Evolution of Steroid Signaling in Metazoans

Gabriel Markov

Numéro d'ordre: 634
Numéro attribué par la bibliothèque: ENSL634

THÈSE

en vue d'obtenir le grade de

Docteur de l'Université de Lyon - Ecole Normale Supérieure de Lyon

spécialité: Sciences de la vie

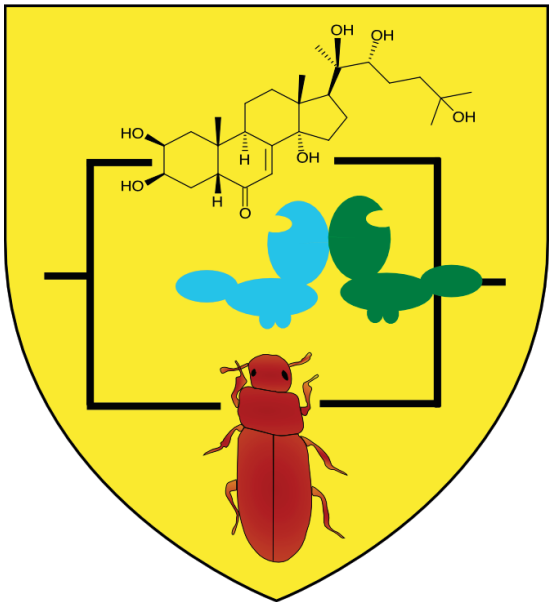
Institut de Génomique Fonctionnelle de Lyon– UMR 5242 IGFL

Ecole Doctorale de Biologie Moléculaire Intégrative et Cellulaire
présentée et soutenue publiquement le 28/6/2011

par Monsieur Gabriel MARKOV

Titre:

Evolution de la signalisation stéroïdienne chez les Métazoaires



Directeur de thèse: Monsieur Vincent LAUDET
Après avis de: M. René LAFONT
Mme Adriana MAGGI

Devant la commission d'examen formée de:

Monsieur Joel DREVET	Membre
Monsieur Thierry GAUDE	Membre
Monsieur René LAFONT	Membre/Rapporteur
Monsieur Vincent LAUDET	Membre/Directeur de thèse
Monsieur Guillaume LECOINTRE	Membre
Madame Adriana MAGGI	Membre/Rapporteur

Summary

Nuclear receptor mediated steroid signaling is involved in many processes in metazoan development, such as puberty in vertebrates, molting in insects and entry into infective stage in some parasitic nematodes. Understanding those phenomena is important regarding public health, agronomical and conservation biology issues. This necessitates to know and to explore the interactions between the evolution of steroid-binding receptors and steroid-synthesizing pathways. My work was articulated around three major parts.

First, using the historical expertise of the laboratory, I updated the relationships between nuclear receptors that are involved in steroid binding, but also from all those that are involved in steroidogenesis regulation, in order to elucidate when and in which context this machinery has arisen.

Second, using a classical comparative genomic approach, I showed that the steroidogenetic enzymes have appeared independently by duplication from xenobiotic-metabolizing enzyme with a wider range of substrate specificity.

Third, I explored the relationships between metabolic pathways using tools from comparative anatomy. This has confirmed and completed the previous results, showing that steroidogenetic pathways have evolved with the pattern of cholesterol degradation pathways.

The synthesis of all these results has led to an evolutionary model where hormonal signaling in bilaterian animals has been inherited from the detoxification of dietary sterols. This model may explain the coupling between nutrient accumulation and sexual maturation, and also the link between metabolic disorders and endocrine disruption due to environmental chemicals or drugs.

Keywords

Evolution, Steroids, Metazoans, Nuclear receptors, CYP450, Hormones

Address of the lab where this work was performed

Molecular Zoology Team
Institut de Génomique Fonctionnelle de Lyon
Ecole Normale Supérieure de Lyon
46 allée d'Italie
FR-69364 Lyon Cedex 07
France.

Titre en français

Evolution de la signalisation stéroïdienne chez les métazoaires

Résumé

La signalisation stéroïdienne médiée par des récepteurs nucléaires est impliquée dans de nombreux processus ayant trait au développement des animaux, tels que la puberté chez les vertébrés, la mue chez les insectes et l'entrée en stade infestant chez certains nématodes parasites. La compréhension de ces phénomènes est importante pour répondre à des questions de santé publique, d'agronomie ou de biologie de la conservation. Ceci nécessite de connaître et de mettre en relation l'évolution des récepteurs qui fixent ces stéroïdes et des voies de synthèse qui produisent les stéroïdes. Mon travail s'est articulé autour de trois grands axes.

Le premier a consisté, en utilisant l'expertise historique du laboratoire d'accueil, à mettre à jour les relations de parenté entre les récepteurs nucléaires impliqués dans la fixation des stéroïdes, mais aussi de ceux qui sont impliqués dans la régulation de la stéroïdogenèse, pour comprendre quand et dans quel contexte cette machinerie est apparue.

Le second axe, utilisant une approche de génomique comparative, a permis de montrer que les enzymes impliquées dans la stéroïdogenèse étaient apparues indépendamment par recrutement d'enzymes à spécificité de substrat plus large impliquées dans la détoxication des xénobiotiques.

Le troisième axe, consistant à explorer les relations de parenté entre des voies métaboliques à l'aide d'outils d'analyse issus de l'anatomie comparée, a complété les conclusions précédentes en montrant que les voies de la stéroïdogenèse avaient évolué suivant des modalités correspondant à une voie de dégradation du cholestérol.

La mise en cohérence de tous ces résultats aboutit à un modèle d'évolution dans lequel la signalisation hormonale des animaux à symétrie bilatérale serait l'héritière de voies de détoxication de molécules stéroïdiennes contenues dans leur alimentation. Ce modèle expliquerait le couplage entre l'accumulation de nutriments et la maturation sexuelle, ainsi que les nombreux dérèglements touchant à la fois le métabolisme et la reproduction dus aux perturbateurs endocriniens ou à certaines molécules thérapeutiques.

Mots-clés

Evolution, Stéroïdes, Métazoaires, Récepteurs Nucléaires, CYP450, Hormones

Blasonnement de la couverture (résumé graphique):

D'or à une ecdysone de sable en chef et à un tribolium castané au naturel en pointe posés entre les branches d'un séparateur aussi de sable à dextre et d'un connecteur du même à senestre et, brochant sur son noeud, un dimère de récepteurs nucléaires d'azur et de sinople.

Acknowledgements

Thanks to the *Ministère de l'Education Nationale, de la Recherche et de la Technologie* and the *Fondation pour la Recherche Médicale* for financial support.
Thanks to the jury members for accepting their task.
Thanks to Raquel TAVARES and Stéphanie BERTRAND for initiation to phylogenetic methods.
Thanks to all members of the Laudet and Demeneix labs for insightful discussions.
Thanks to the administative staff, especially Fabienne ROGOWSKY, Sonia CELARD, Corinne NOVEL-CATIN, and Aïcha BENANNA for their valuable help in many practical aspects.
Thanks to my teaching collegues David BUSTI, Sandrine HEUSSER, Nicolas VIDAL, Amaury DE LUZE, Bruno QUEYRAT, Laurent SACHS, Marie-Stéphanie CLERGET-FROIDEVAUX, Eric GUIBERT, Marie SÉMON, Florent CAMPO-PAYSAA and Alexa STADIER for helping me to widen the perspective around my research interests.
Thanks to Loic PONGER for help in shell programming.
Thanks to Blaise LI for making available the LaTeXtemplate of his PhD thesis.
Thanks to Fabrice GIRARDOT, Laurent COEN, Pierluigi SCERBO and Céline VIVIEN for providing me a basic culture on the stem-cell topic.
Thanks to Mathilde PARIS, Frédéric BRUNET and Michael SCHUBERT for discussions on comparative genomics.
Thanks to Marc ROBINSON-RECHAVI and Philippe DURAND for advice during the annual meetings of my PhD thesis comitee.
Thanks to François BONNETON for numerous discussions on a lot of scientific topics.
Thanks to Michael BAKER for numerous discussions on chordate steroidogenesis.
Thanks to Chantal DAUPHIN-VILLEMANT and Virginie ORGOGOZO for fruitful discussions on arthropod steroidogenesis.
Thanks to Guillaume LECOINTRE for the collaboration on cladistic aspects and for a lot of methodological advice.
Thanks to Barbara DEMENEIX for hosting me in her lab at the MNHN during three years, and thus giving me a solid background on physiological reasoning.
Thanks to Vincent LAUDET for opening my mind to molecular zoology and for coaching me in spite of his multiple responsibilities.
Thanks to my undergraduate teachers, and more specifically to: Solange PIERRAT for basics in history, which can also be useful for an evolutionary biologist, Jean-Philippe ROUX for the physics "olympiades" and first contacts with the scientific method, Philippe LESUR for basics in biology and geology, on for insisting on how the gene notion is complicated - as are many biological concepts indeed... - , Elena SALGUEIRO for two further years on basics in biology and geology, for making me aware of the difference between models and reality, and for helping me to fight against a natural tendency to "japanese-style" overconcision, José BELIN for three years of insightful philosophical teaching, Robert DELORME and Maxence REVAULT D'ALONNES for first lab experiments during TIPE, Jean DEUTSCH for some RT-PCRs and rich discussions about nervous system homology, Ioan NEGRUTIU for the Master 1 module on bioethics, Pierre THOMAS for making me aware of the importance of having a naturalist approach to natural phenomenons, and Michael MANUEL for further dissections of exotical organisms.
Thanks to Serge SIRE for checking the accuracy of the blazon description.
Thanks to my friends, my family and the numerous people from which I learned something, or which helped me to study in very favorable material conditions.
Thanks to my parents for providing some genetical and a lot of environmental imput.

Contents

Part I

Introduction

Chapter 1

Foreword: in defense of molecular zoology

What is the common point between pubertal transformation in teenagers, insect molting and the entry of parasitic nematodes into an infestation stage? All these processes are animal life history transitions that are regulated by steroid signaling through nuclear receptors.

The chemical term ”steroid” refers to a type of organic compound that contains a specific arrangement of four cycloalkane rings that are joined to each other. Nuclear receptors are metazoan transcription factors activated by small lipophilic ligands, such as steroid hormones, thyroid hormones, retinoids or fatty acids. The availability of the ligand controls, in time and space, the transcriptional activity of nuclear receptors

Understanding the molecular basis of such processes by elucidating the relationships between steroids and their receptors has major health and agronomical implications. But, on a more fundamental viewpoint, this is also a beautiful model to link the evolution of molecules to the evolution of organisms. This implies to put the molecular diversification, not only from gene products, but also from lipidic molecules, in a temporal an zoological framework, mapping the new steroids and receptors that were acquired on various nodes of the metazoan tree. This structural description also necessitates a functional recontextualisation, explaining what these steroids and receptor do, with the aim to understand what could have been the functional significance of each new step (Morange, 2011). This does not mean that each acquisition of a new steroid or a new receptor should be necessarily viewed as the acquisition of an adaptative advantage *per se*. But even for an innovation that can be understood as an accidental recruitment of an ancient structure to a new function, it is important to determinate why this has occurred at this precise time in evolutionary history, considering both the physiological consequences of a given innovation in terms of internal body changes, and its consequences on the abilities to cope with the environment and to interact with other organisms living at that time.

We are fully aware that this proposed study scheme can appear as extremely ambitious for a PhD thesis. Even in the very simplified framework that consists of metazoans reduced to metazoan genomic models, there is a lot of steps regarding steroid evolution to discuss. Therefore, we do not aim to analyze in full each of these steps, because for some of them, there would be an obvious lack of precise information. But we think that it is extremely important to put the question of steroid evolution in this general perspective, even if it means, that at some steps, we will be forced to acknowledge that our hypotheses are somewhat speculative or even that we do not have any data.

In this introduction, we will first briefly review some basic concepts on intercellular signaling. Then, we will present the two most important partners at the molecular level

in the steroid signaling pathway, namely steroids and nuclear receptors. After a brief survey of the relationships between the various animal groups that are discussed here, we will pinpoint the anatomical, physiological and molecular innovations that are relevant regarding steroid signaling. After that, we will conclude with a survey of the current knowledge about the evolution of the ligand-binding ability of nuclear receptors, in order to set the stage for the three main sub-questions that have been addressed in this work. These questions are:

- when have nuclear receptor begun to bind steroids as ligands? What was the zoological context at that time?
- when have appeared the enzymes that are responsible for steroid hormone synthesis?
- when have appeared the pathways leading to the current steroids, and how have these pathways evolved?

Chapter 2

Steroid signaling

This chapter aims at presenting the central players in steroid signaling, that are steroids and their receptors. But before that, we will set the stage with a brief reminder of some basic principles in intercellular communication.

2.1 Basic principles in intercellular communication and environmental sensing

2.1.1 The cell is the basic unit of life

All living beings consist of one or many cells (Schwann, 1839), and each cell originates from another cell by division. A cell is characterised by its structure and its metabolism. Both are highly variable, and these variations are modulated by intrinsic and extrinsic parameters. Intrinsic parameters are the metabolic state of the cell, its internal architecture and molecular composition. Extrinsic parameters are the cues from the environment. Even in bacteria, cells are normally not living isolated, but they are closely interacting with other cells from the same clonal population (Rosenberg, 2009) or from a different genetical background. Additionally, all cells can respond to non-cellular stimuli, that can be biophysical parameters, such as light or temperature, or chemical molecules that are present in the environment. Such chemical molecules can be the direct or indirect product of biological activity, or originate from ”purely” geological processes, such as volcanic eruptions or chemical alteration of rocks.

2.1.2 Cells are able to maintain themselves in a changing environment

The intrinsic variability of any living medium thus raises the general problem of homeostasis, which is the conservation of an equilibrium in internal organism parameters (Bernard, 1865), and also the problem of coordination of growth and cell division with the environmental conditions. One of the best examples of this mechanism is the *lac* operon, the first discovered case of gene regulation. Depending on the main available carbon source, the eubacterium *Escherichia coli* is able to grow whether on lactose or on glucose. In a glucose-rich medium, the lactose metabolizing pathway is shut down because a repressor (lacI) blocks the the synthesis of the proteins that permits the uptake of lactose and its breakdown into glucose and galactose. In presence of lactose, its binding to the repressor induces a conformational transition that decreases its affinity

for the operator DNA of the *lac* operon, thus allowing the transcription of the genes that allow the lactose metabolization (Jacob and Monod, 1961).

2.1.3 Metazoan cells have both an internal and external environment

In metazoans, that are multicellular organisms, a new level of interactions is added. Whereas the cells located in the most external layer of epithelia are still in direct contact with the environment, all other cells are in contact through the ”internal milieu”, an internal body fluid whose composition differs from the external world (Bernard, 1865). Except from direct communication between neighboring cells, all other intercellular communications so occur through some small chemical molecules that are released in the internal milieu. These molecules, that are produced by an endocrine cell are transported in the internal milieu up to a receptor cell, which receives and interprets the signal. This process is called signal transduction and it occurs through two main different systems, that are not mutually exclusive.

2.1.4 Molecular mechanisms of signal transduction at the cellular level

Signaling molecules can be bound at the cell surface by membrane receptors, that activate some metabolic processes or activates gene transcription through a series of intermediates (Fig 1., left part). Or, if they are lipophilic, they can go through the cell membrane and bind directly to a nuclear receptor (NR), that will translocate into the nucleus after ligand binding and thus directly trigger gene transcription (Fig. 1, right part).

The signaling molecules are released in response to a stimulus. The stimulus is a detectable change in internal or external environment. Although this term is often associated with neurosensory perception, it is worth to stress that a stimulus can be of all kinds of nature, and is more generally the result of the integration, at the cellular level, of various internal and external parameters. The results of this integration is the emission of a signal, which can be a local variation of membrane potential, in the case of neuronal communication, the synthesis of a membrane protein, in the case of cells that are in direct physical contact one with each other, or the excretion of a signaling molecule, in case of distant intercellular communication. Then occurs a step of transport from the emetting cell to the recepting cell. This transport can be purely passive diffusion, or it can be facilitated by a transporter. The reception implies the binding of the signaling molecule and the transduction of the signal, which can be the activation of the transcription of some target genes or direct changes in cell metabolism or in electron distribution around the membrane. The extinction of the signal implies the degradation of the receptor and the further metabolization of the hormone.

2.2 Steroids, an example of signaling molecules

2.2.1 Structure and biochemical properties

Steroids are one of the groups of signaling molecules between non-adjacent cells in animals. Chemically, steroids are classically defined as organic molecules with four cycloalkane rings that are joined one to each other (Fig. 2). Facultatively, it can also bear a carbon side-chain branched on carbon 17. We insist on this, highlighting the side-chain

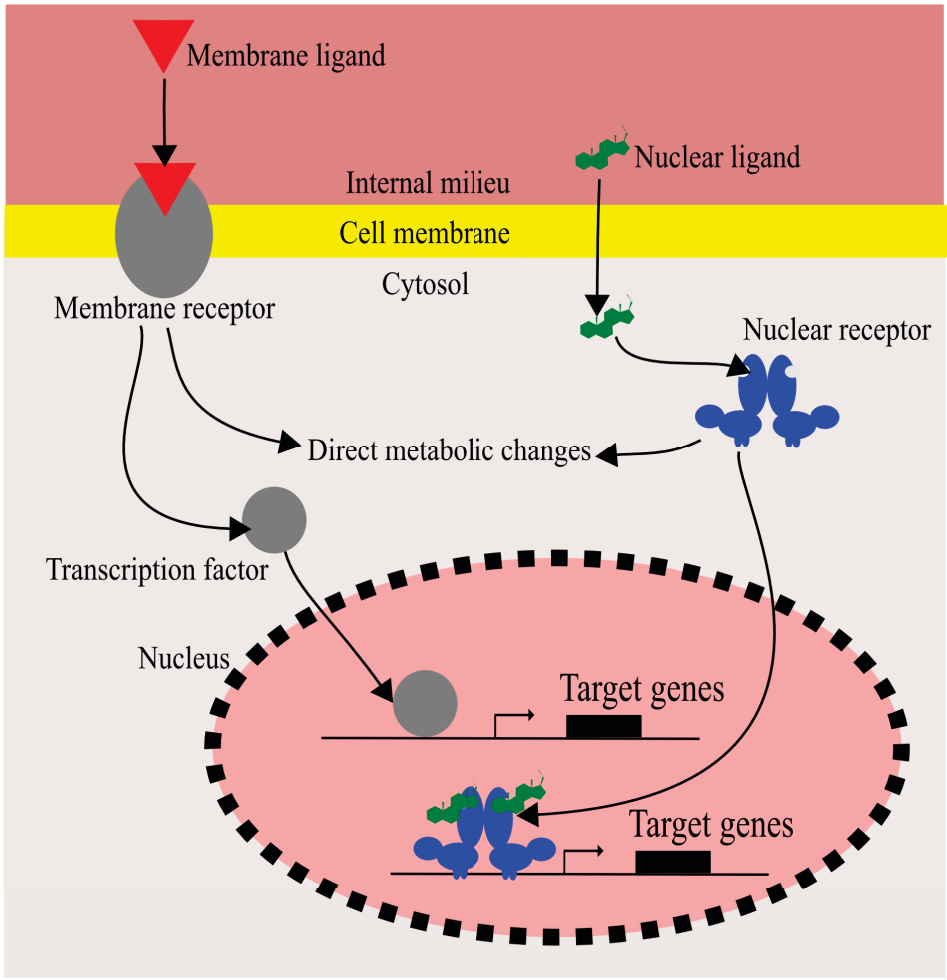


Figure 1: Two main ways of signal transduction in metazoans. On the left, transduction through a membrane receptor. On the right, transduction through a nuclear receptor. Note that for both receptor types, the exact intracellular localisation can vary.

on the figure, because the difference between steroids with and without a side-chain will be an important discussion point along the whole manuscript.

Just as for proteins (Markov et al., 2008a), steroid nomenclature is highly heterogenous, resulting from a complex history, during which these objects were named in different conceptual frameworks. The Fig. 2 gives an example of this diversity.

Cholesterol, the solid component in bile

Cholesterol, the first identified steroid, was firstly discovered in bile and gallstones by François Poulletier de La Salle in 1765, which did not publish his observations (Feltgen, 1993), and then rediscovered in 1815 by Chevreul, who coined the term “cholesterine.”, from the Greek *chole-* (bile) and *stereos* (solid) (Chevreul, 1815). Initially viewed only as a morbid substance or a metabolic waste, it was later acknowledged as a molecule playing important physiological roles, being a modulator of membrane fluidity and as a precursor for bile acids, steroid hormones and vitamin D. Other eucaryotes have also sterols, that are steroids with an hydroxyl group on carbon 3 of ring A. In yeast, this is ergosterol, whereas in plants, there are phytosterols (Summons et al., 2006). In bacteria, sterols are absent, but they role in cell membrane is performed by pentacyclic carbone

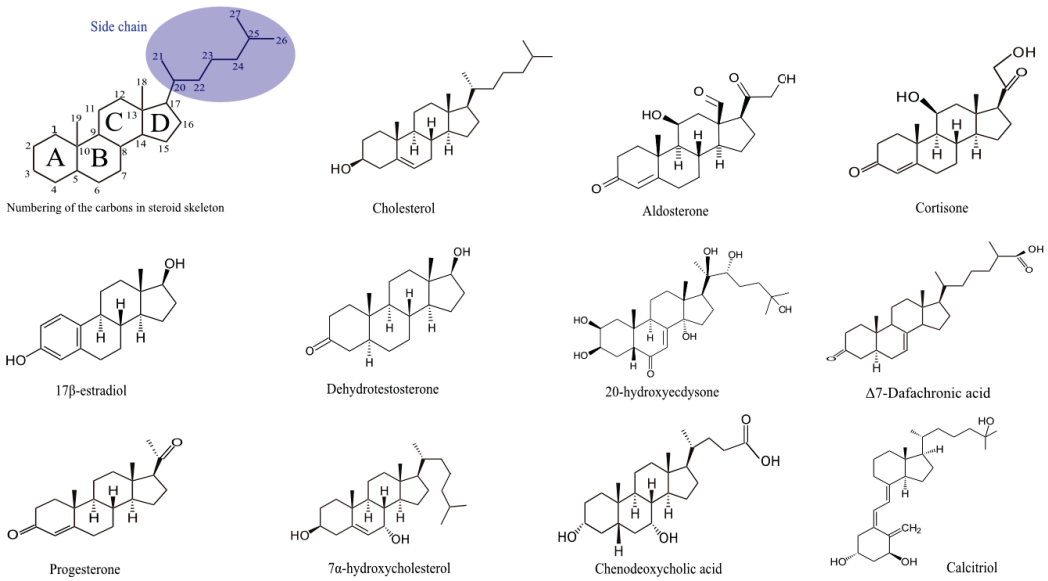


Figure 2: Structure of the sterol ring, of cholesterol, and of the steroids of human, *Drosophila melanogaster* and *Caenorhabditis elegans*. The sterol ring numbering is indicated. Aldosterone, cortisone, estradiol, dihydrotestosterone and calcitriol are the classical human steroid hormones. Oxysterols, such as 7α-hydroxycholesterol, and bile acids, such as chenodeoxycholic acids, are not classically considered as hormones, but they are, as the others, steroid ligands for nuclear receptors. 20-hydroxyecdysone is the main steroid hormone in *Drosophila melanogaster*, whereas Δ7-dafachronic acid is one of the two steroids that are involved in intercellular communication in *Caenorhabditis elegans*.

molecules called hopanoids, that are classified with steroids for some authors (Fahy et al., 2005).

Corticoids, or steroids from the adrenal gland

Cortisone was isolated in 1936 as an important hormonal product secreted by the cortex of the adrenal gland. It was later found to affect glucose metabolism, hence the name glucocorticoid. Another important glucorticoid is cortisol, that is also a sterol from a chemical viewpoint. This is an example of overlap between the various steroid categories.

Aldosterone was initially identified in 1953 under the name ”electrocortin” as a product of the mammalian adrenal gland that was able to affect sodium and potassium transport in the nephrons of the kidney (Simpson et al., 1953). Together with other molecules of similar structure that are able to modulate the ionary equilibrium, it is a member of the mineralocorticoid family. In rodents, amphibians and birds, another molecule, 11-Dehydrocorticosterone, seems to be the main physiological mineralocorticoid (Bury and Sturm, 2007).

Both glucocorticoids and mineralocorticoids are grouped under the general term ”corticoids” or ”adrenal steroids” in reference to their origin in the adrenal gland. In vertebrate lacking an adrenal gland, such as teleosts, they are produced by the interrenal tissue of the head kidney. The distinction between gluco- and mineralocorticoids is not very clear, especially when one considers the diversity of situations in metazoans. For

example, cortisol acts both as a mineralocorticoid and a glucocorticoid in teleosts (Bury and Sturm, 2007).

Estrogens, androgens and progestagens: steroids involved in vertebrate reproduction

17β-estradiol and some other steroids with 17 carbons, called C17-steroids, such as estriol and estrone, are grouped under the term "estrogens", referring to their ability to generate sexual desire in mammals. The classical test to detect estrogenicity of a molecule is the mesure of its ability to stimulate cornification in epidermal cells lining the vagina of castrate female rats, which is a very restrictive definition. But more and more, this test is replaced by *in-cellulo* tests in yeast system with transfected mammalian estrogen receptor (Norris, 2007). Other molecules than steroids that are also able to induce estrogenic effects are also termed"estrogens", such as genistein, a component of soy (Henley and Korach, 2006). So there has been a shift in the "estrogenicity" concept from the description of a physiological effect (induction of female reproductive traits) to a molecular effect (binding to an estrogen receptor). It should also be stressed that estrogens are also present in male and are known to have effects that are distinct from feminization.

Dihydrotestosterone and other C18-steroids are androgens, i. e. compounds that stimulate development of male characteristics. Here again there is cohabitation of two concepts, one referring to the function (androgen) and one to the organ where the steroid is produced (testosterone is the steroid from the testis). Even in females, testosterone is synthesized, even if not in testis.

Progesterone and other C21-steroids are grouped under the name "progestogen", referring to their ability to maintain pregnancy in mammals. However, progesterone also exists in other vertebrates, that are not necessary life-bearers, with various and only partially investigated functions in modulation of vitellogenesis and gamete proliferation (Norris, 2007) . It should be noted that the "C21-steroid" category is overlapping, because corticoids, that are not sex steroids, are also C21-steroids (Fahy et al., 2005).

Progestagens, estrogens and androgens are sometimes grouped together under the name "sex steroids", referring to what is inferred to be their main biological role in vertebrates. There have been many claims for their presence in other animal groups too (Lafont and Mathieu, 2007).

Oxysterols: a purely biochemical name for widespread molecules

7α-hydroxycholesterol is involved in cholesterol homeostasy. Here the name refers only to the chemical structure: a cholesterol bearing an hydroxyl group. Oxysterols are also widespread among animals (Lafont and Mathieu, 2007).

Bile acids: steroids that are present in the bile

Chenodeoxycholic acid was first isolated in goose - hence the prefix "cheno" in its name, and is one of the two most important bile acids in human, being involved in cholesterol homeostasis as well as in facilitation of lipid digestion (Russell, 2009).

Secosteroids: steroids that lack the canonical steroid structure

Calcitriol is involved in calcium homeostasis. A lack of calcitriol leads to rickets, a softening of the bones in children suffering from malnutrition, and osteomalacia and

osteroporosis in adults (Holick, 2003). The importance of calcitriol in diet has led to its naming as vitamin D3. It is worth to mention that calcitriol lacks the canonical four-ring structure, as a consequence of the opening up of the B-ring. Hence the term "secosteroid" was coined as a synonymous to vitamin D to group all steroids that underwent this opening. Here a temporal dimension is added in the definition of steroids, because secosteroids have no more the four polyalkane rings but are formed from a molecule that has it. Secosteroids can also be produced from ergosterol in fungi, and this leds to vitamin D2.

Ecdysteroids: steroids that trigger arthropod ecdysis

20-hydroxyecdysone was isolated from a crayfish (Hampshire and Horn, 1966) and was rapidly established as the main moulting hormone in all arthropods (Lafont et al., 2005). 20-hydroxyecdysone and similar compounds are grouped under the general name of ecdysteroids. Since they have been also identified in plants, a further distinction is made between "zooecdysteroids" and "phytoecdysteroids" (Lafont et al., 2005). A number of steroids have also been isolated in sponges (Lafont and Mathieu, 2007).

Dafachronic acids: steroids that block dauer entry in nematodes

Δ7-dafachronic acid was recently identified as a ligand for the nuclear receptor DAF-12 in the model nematode *Caenorhabditis elegans*. The name "dafachronic" was coined referring to its ability to block dauer formation (a nematode-specific diapause stage) and to modulate other so-called "heterochronic" developmental pathways (Motola et al., 2006). A second molecule, Δ4-dafachronic acid, was identified at the same time and was shown to have similar effects.

Additionally, there are many steroids with functions that are not related to intercellular communication within one organism. They also play a role as intra- or interspecific communication substances, being pheromones, alarm substances, feeding deterrent or toxins (Lafont and Mathieu, 2007).

It must also be noted that outside animals, steroids are also implicated in cell signaling through NR-independant mechanisms. In plants, brassinosteroïds are growth-promoting hormones that act through receptor kinases at the cell membrane (Kim and Wang, 2010).

May an evolutionary approach help in giving structure to such heterogenous classifications?

Seeing this diversity of steroid names, we can now try to list the main types of steroid categories:

- steroid names referring to a chemical structure: sterols, secosteroids, oxysterols, C17, C18 and C21 steroids. Concerning secosteroids, it is worth to mention that the names refers not only to a structure, but also to a process, because "seco" means that the steroid has undergone a cleavage of its B-ring.
- steroid names referring to a localisation in the body: bile acids, cholesterol, testosterone.
- steroid names referring to an organism: ergosterol, phytosterols, zoosterols, chenodeoxycholic acid.

- steroid names referring to a function: androgens, estrogens, progestagens, mineralo- and glucocorticoids, ecdysteroids and dafachronic acids.

One important unifiying factor between these various categories is the temporal dimension. This is obvious for secosteroids, where the name refers to a synthesis process with successive steps. But this is also true concerning names related to a localisation, that is mainly historical. Cholesterol is indeed present in bile, but it is also present in all cells as a membrane component. In names referring to an organism, the temporal dimension comes because these organisms have diverged during evolution.

Concerning functional names, we can observe that they refer to processes occurring at different time scales. Some are unambiguously rooted in a developmental timeframe. Androgens and estrogens mediate puberty, the transition from a juvenile to a sexually mature adult. Ecdysteroids triggers the molts that punctuate arthropod growth. Dafachronic acids repress the entry into dauer stage. In some other cases, such as mineralo- and glucocorticoids, the situation is more complicated. Corticoids are well known modulators of amphibian development (Denver, 2009), and more generally, unexpected variations of corticoids are well known as disruptors of reproduction in captive vertebrates (Norris, 2007). Bile acids are regulators of the digestive cycle, which is not considered as a developmental process, because of its daily periodicity, but which is nevertheless a periodic process.

From this we draw two major conclusions:

- these observations raise a question about the biological basis for the implication of steroids in rhythmic processes. Is it purely fortuitous or is it the result of they special properties? We will discuss that point a little in the Part II (nuclear receptors in a zoological context) and more precisely in the discussion.

- the way steroids are made is important to classify them: we will esplore this more precisely in the following subsection.

2.2.2 Steroid synthesis

Most of animal steroids are synthesized from cholesterol, which is itself synthesized from acetate that is produced via glycolysis or fatty acid oxidation in the liver (Norris, 2007). Hence, the term steroidogenesis refers to both the synthesis of cholesterol from acetate and to the synthesis of other steroids from cholesterol. Here we will briefly present what is known about the synthesis of the main steroids.

Synthesis of human sex and adrenal steroids

Vertebrate sex and adrenal steroids are synthesized by various organs, the main being the gonads (for estrogens, androgens and progesterone), the placenta (for estrogens and progesterone in eutherian mammals), the adrenal cortex (for corticoids and androgens) and the brain (Norris, 2007).

The synthesis of steroid hormones is performed by enzymes belonging to four different families (Payne and Hales, 2004), such as the cytochrome P450 (CYP), the short-chain dehydrogenases-reductases (SDR), the 3β-hydroxysteroid dehydrogenases (HSD3B) and the 5α-reductases (SRD5A).

In the Fig. 3, we present the classical pathway in human and rodents. There are some variations from one vertebrate to another (Bury and Sturm, 2007) in the terminal products. For examples, teleosts can also synthesize 11ketotestosterone from testosterone, using the CYP11B and HSD11B enzymes (Lokman et al., 2002).

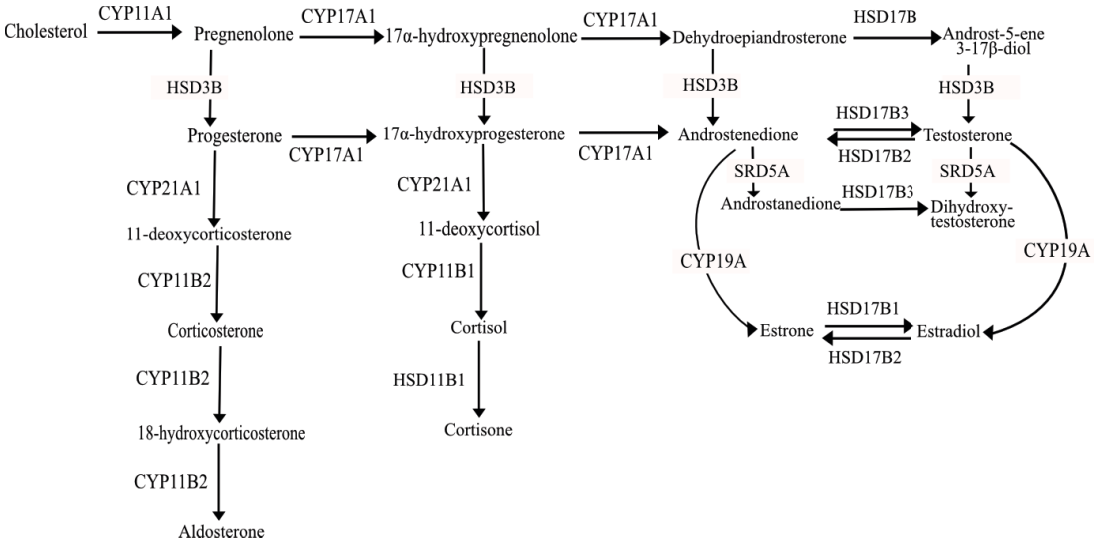


Figure 3: Synthesis of sex and adrenal steroids, showing all intermediates and enzymes.

Synthesis of human 1,25-dihydroxyvitamin D3

In presence of ultraviolet light in the skin, 7-Dehydrocholesterol is modified to vitamin D3 (Holick, 2003). Then it is released into the blood from which it is removed by the liver, and then converted to 25-hydroxyvitamin D3, and returned to the blood. The final step is carried out in the kidney and leads to 1,25-dihydroxyvitamin D3 that is the active molecule (Fig. 4). So, synthesis of 1,25-dihydroxyvitamin D3 is the only animal steroidogenic process that does not start from cholesterol and that includes an non-enzymatic step, the cleavage of the B-ring by ultraviolet light. The other reactions are carried out by members of the CYP family, but concerning the vitamin D3 25-hydroxylation, almost two different paralogs, CYP2R1 or CYP27A1, may perform that reaction *in vivo* (Gupta et al., 2007).

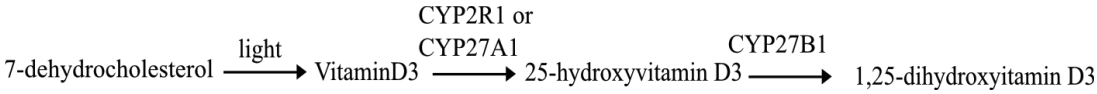


Figure 4: Synthesis of vitamin D3

Synthesis of human bile acids

Synthesis of bile acids occurs in the liver (Russell, 2009). There are two major pathways for bile acid synthesis. In the classical or neutral pathway (Fig. 5), the first reaction is 7α-hydroxylation, whereas in the alternative or acidic pathway, that was discovered later, the first reaction is a 26-hydroxylation, leading to 3β,7α-dihydroxycholestenoic acid (Fig. 5). Chenodeoxycholic acid can also be synthesized starting from 25-hydroxycholesterol or 24-hydroxycholesterol.

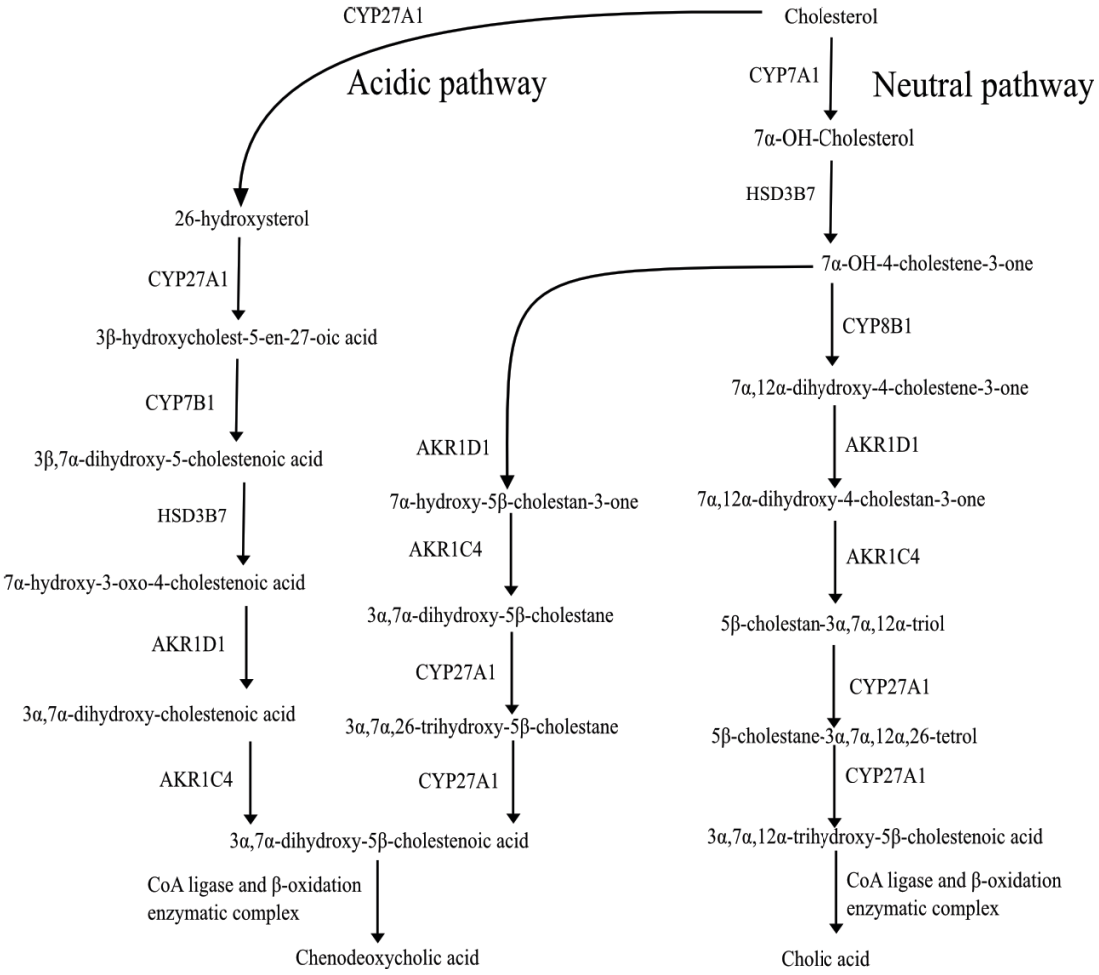


Figure 5: Synthesis of human bile acids, showing all intermediates and enzymes.

All the bile acids that are endogenously synthesized are called primary bile acids. They can be further metabolized by gut bacteria, leading to secondary bile acids, some of which, such as lithocholic acid, can be carcinogenic (Russell, 2009).

Synthesis of arthropod ecdysteroids

Ecdysone is synthesized from dietary sterols in the prothoracic gland of insects (Huang et al., 2008) or in the Y organs of crustaceans (Lafont and Mathieu, 2007). The last four steps of the synthesis are performed by enzymes from the CYP family (Fig. 6) and the first one is performed by the rieske-domain oxygenase Neverland, but some reactions in the middle, called the ”black box reactions”, are still misterious (Huang et al., 2008).

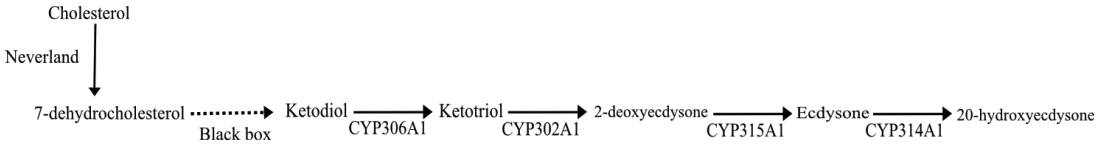


Figure 6: Synthesis of arthropod ecdysteroids. The names of the known enzymes performing the reactions are indicated on the arrows.

Synthesis of nematode dafachronic acids

Synthesis of nematode dafachronic acids involves the CYP enzyme CYP22 (Motola et al., 2006), the rieske-like oxygenase DAF-36 (Rottiers et al., 2006) and the HSDB3 HSD-1 (Patel et al., 2008). But the enzymes that transform 7-dehydrocholesterol into lathosterone are not yet kown (Fig. 7).

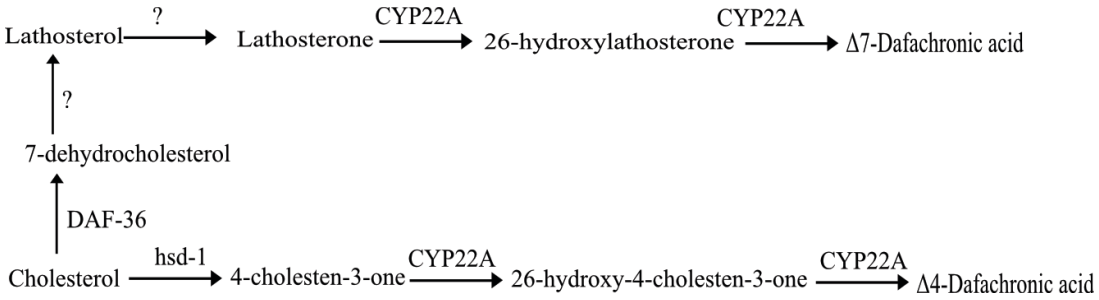


Figure 7: Synthesis of nematode dafachronic acids. The names of the known enzymes performing the reactions are indicated on the arrows. Question marks indicate unknown enzymes.

Regulation of steroid synthesis

Regulation of steroid synthesis occurs by different ways. The availability of steroidogenic enzymes is controlled by various transcription factors, among which are some nuclear receptors, that do not necessarily bind steroids themselves ((He et al., 2010); (Parvy et al., 2005); (Horner et al., 2009))(Fig. 8).

2.2.3 Physiological roles of steroids, other than through NR-binding

Steroids bind to nuclear receptors, as we will detail in further section, but they are also able to bind membrane receptors. For example, in mammals (Revankar et al., 2005) and in some teleosts (Thomas et al., 2010), estradiol binds the receptor GPER (formerly GPR30) that is located in the membrane of the endoplasmic reticulum. In mammals, some bile acids also bind the transmembrane receptor TGR5 (Thomas et al., 2008). Similary, in the butterfly *Manduca sexta*, ecdysone regulates the proliferation of neural precursor cells through a nongenomic mechanism in the optic lobes (Champlin and Truman, 2000).

The ability of steroids to act independently of nuclear receptors has two important consequences. From a methodological viewpoint, this means that the identification of a given steroid in an animal is not sufficient to draw conclusions on its physiological

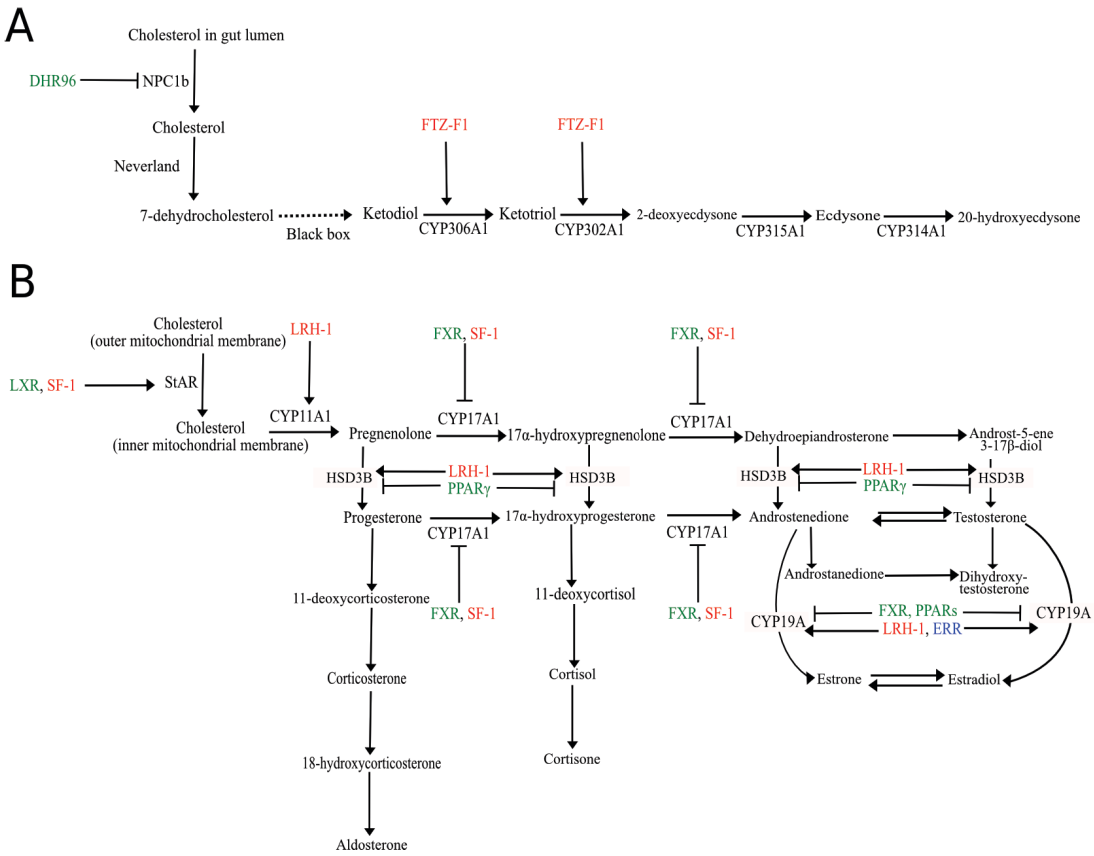


Figure 8: Control and regulation of steroid synthesis by steroidogenic enzymes and transcription factors. A. Ecdysteroid synthesis in *Drosophila melanogaster*. B. Sex and adrenal steroid synthesis in human. The colors of the nuclear receptors pictured here refer to their phylogenetic relationships, as they are described in section 2.3.

role, and on the molecular pathway in which it could be involved (Lafont and Mathieu, 2007). From a fundamental viewpoint, this means that one important open question regarding the evolution of steroid signaling is when and why they became ligands for nuclear receptors. But before to discuss this, it is important to explain what nuclear receptors are.

2.3 Nuclear receptors

Nuclear receptors are metazoan transcription factors activated by small lipophilic ligands, such as steroid hormones, thyroid hormones, retinoids or fatty acids. The availability of the ligand controls, in time and space, the transcriptional activity of nuclear receptors. However, this general and traditional definition has to be put into perspective, since natural ligands are still missing for several receptors in mammals and for most of them in insects. Furthermore, it appears that some nuclear receptors are true orphans. However, most members of this family are expected to be regulated by one or several ligands. Testing this hypothesis has proven to be an arduous challenge, especially when endocrinological knowledge is non-existent for a given receptor.

2.3.1 Modular proteins with a conserved structure

Nuclear receptors are modular proteins with two well structured domains, the DNA-binding domain (DBD) and the ligand-binding domain (LBD), that are separated by an hinge region. Before the DBD in the N-term, there is a regulatory domain that is highly variable in sequence. The C-term also ends with a variable region. The DBD and LBD are highly conserved and their structures have been determined for several receptors (Huang et al., 2010) (Fig. 9). Nuclear receptors bind to the regulatory regions of target genes as homodimers (when two receptors of the same type bind together) or heterodimers (when two receptors of different types bind together), more rarely as monomers.

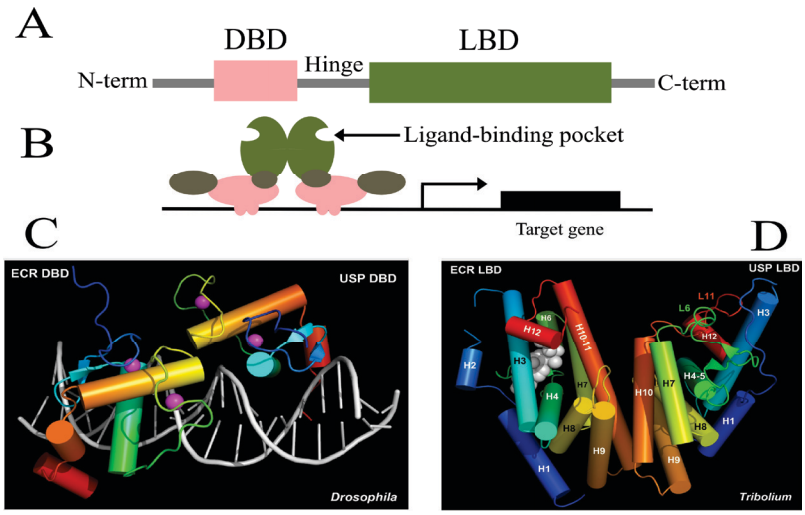


Figure 9: Structure of a nuclear receptor. (A) Primary structure, with the less conserved domains in grey, the DBD in pink and the LBD in green. (B) Nuclear receptor dimer on the promoter of a target gene, with the ligand-binding pocket shown in the LBD. (C) Tridimensional structure of two DBDs on a DNA helix. The pink balls are the zinc atomes. (D) Tridimensional structure of a dimer of two LBDs. In the EcR LBD, the white balls represent the ecdysone ligand. The surface around it determines the ligand-binding pocket. The names H1 to H12 refer to the twelve α-helixes that make the LBD. Concerning the USP LBD, two of them are transformed into loops and are therefore named L6 and L11 (Bonneton and Laudet, 2011)

2.3.2 A phylogenetic classification of nuclear receptors

The high degree of sequence conservation in the DBD and LBD of nuclear receptors has made possible a phylogenetic classification of the family (Laudet et al., 1992). This classification (Fig. 10) has provided a basis for the official nomenclature of the family (Nuclear Receptors Nomenclature Committee, 1999), that divides classically into six subfamilies (NR1, NR2, NR3, NR4, NR5 and NR6). Each family itself divides into groups (NR1A, NR1B, NR1C...).

Among them, steroid-binding nuclear receptors are located both in the NR3 subfamily (ER, GR, MR, PR, AR) and in the NR1 subfamily (EcR, LXR, FXR, VDR, DAF-12...). We will discuss their relative binding- abilities more extensively in a following section about evolution of ligand-binding ability.

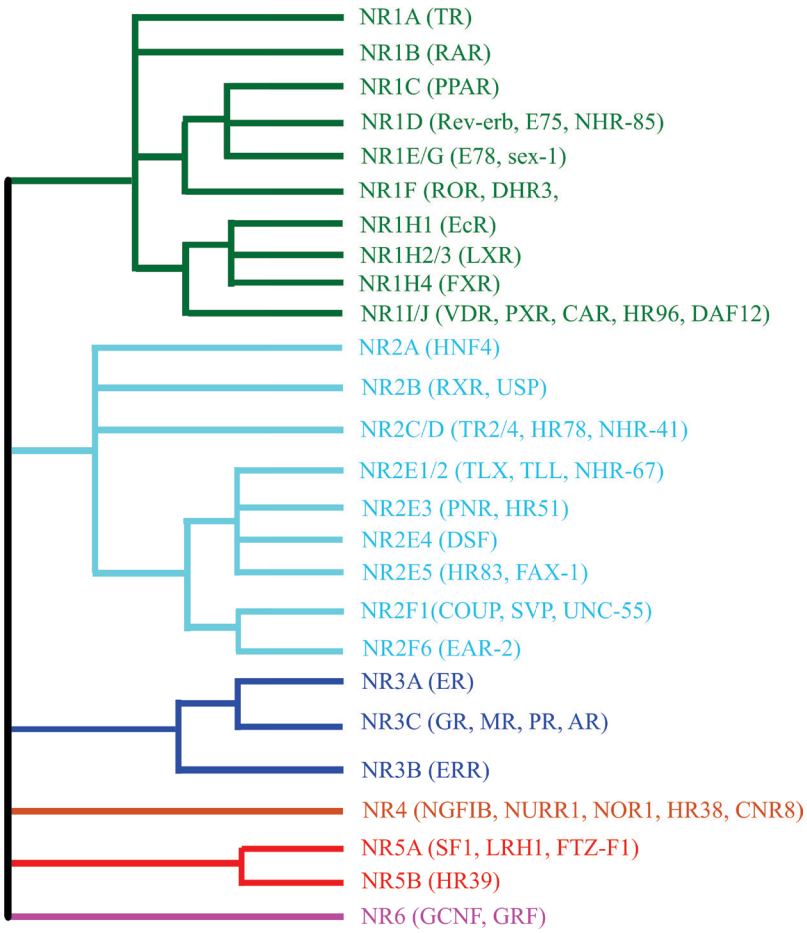


Figure 10: A classification of nuclear receptors based on their evolutionary relationships. The main trivial names for nuclear receptors from human, *Drosophila melanogaster* and *Caenorhabditis elegans* are indicated between brackets.

2.3.3 The DNA binding domain

The DNA binding domain (DBD) is made up of two non-equivalent zinc-finger structures (C4-zinc fingers) with each zinc atom being necessary to retain stable domain structure and function (Khorasanizadeh and Rastinejad, 2001). Recent genomic studies have shown that the number of nuclear receptor binding sites range from several hundreds (Gauhar et al., 2009) to several thousands (Cheung and Kraus, 2010). The canonical hormone response element (HRE) has the core sequence RGGTCA (Umesono and Evans, 1989). Mutations, extensions, duplications and distinct relative orientations of repeats of this motif generate response elements that are selective for a given class of receptor (Fig.11). Some receptors can bind to DNA as monomers through a single core sequence. In this case, an A/T-rich region 5' to the core element governs the binding specificity (Laudet and Adelmant, 1995). All members of subfamily 2 are able to form homodimers on direct repeat sequence. Steroid receptors (NR3) bind as homodimers to HREs containing two core motifs separated by 1-3 nucleotides and organized as palindromes. Most of the receptors of subfamilies 1 and 4 heterodimerize with RXR (NR2B, homologous to the insect USP) and bind to direct repeats (DRs) of the core motif (Brelivet et al., 2004). The spacing between the two halves of the DR dictates the type of heterodimer that will bind. For example, a DR separated by five nucleotides (DR5) will

usually be recognized by RXR:RAR and a DR4 by RXR:TR.

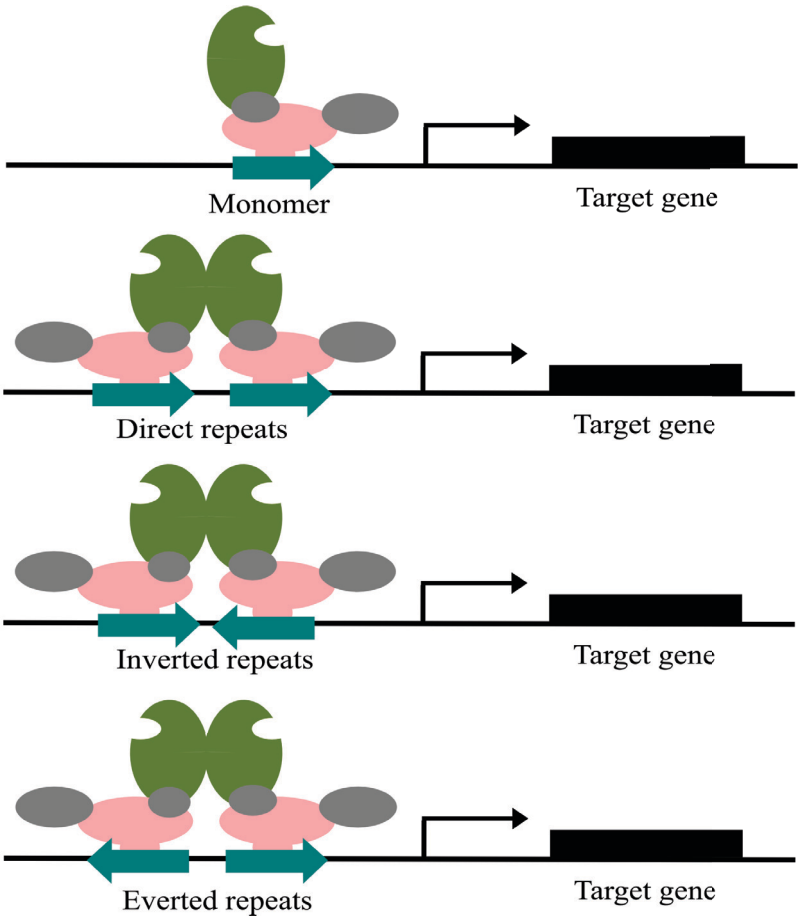


Figure 11: Various ways of DNA binding in nuclear receptors. The various receptor domains are colored according to the Fig. 9B.

2.3.4 The ligand binding domain

The ligand binding domain (LBD) is composed of 10-12 α-helixes that form three antiparallel helical layers that combine to make an α-helical sandwich (Fig. 12). The LBD fulfils three main functions: ligand binding, dimerisation and recruitment of coregulators. The ligand-binding pocket (LBP) of the receptor is located in the interior of the structure and is formed by a subset of the surrounding helices. The core of the dimerisation interface is mainly constituted by helices H9 and H10 (more than 75% of the total surface), together with other residues from helices H7 and H11 as well as from loops L8-9 and L9-10 (Huang et al., 2010). The heterodimeric arrangement closely resembles that of a homodimer, except that the heterodimer interfaces are asymmetric (Folkertsma et al., 2005). Many unliganded (apo) nuclear receptors are transcriptional silencers as a result of interaction with corepressors. The LBD domain undergoes a conformational change upon ligand binding (holo), allowing the interaction with coactivators and the transactivation of target genes.

Different types of receptors are distinguished according to their various ligand-binding abilities (Benoit et al., 2004).

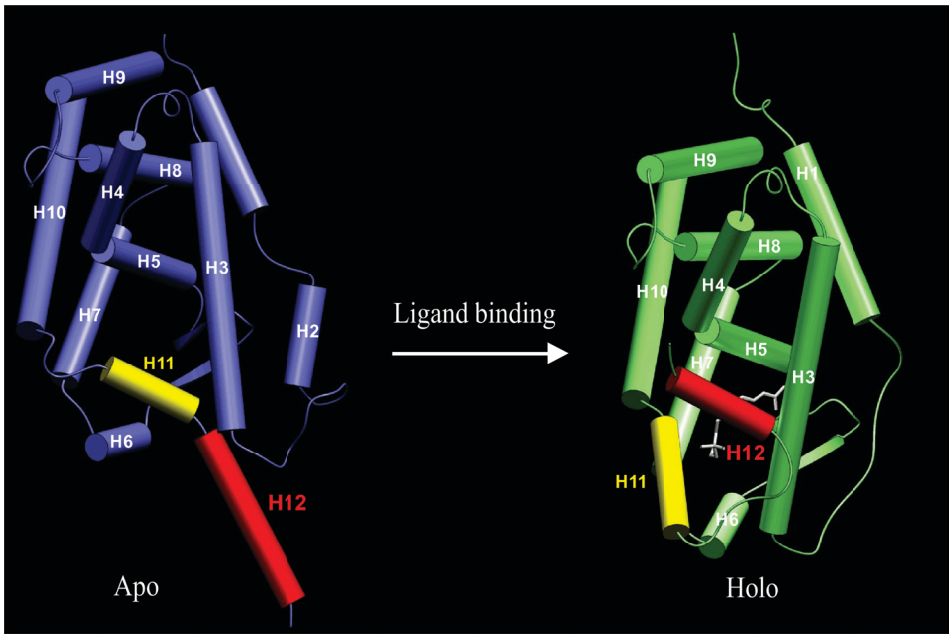


Figure 12: Allosteric transition in response to ligand activation in the human RXRα. The fixation of oleic acid (in white) triggers a conformational change, where the most visible effect is the move of the H12 helix (in red) that locks the ligand into the binding pocket. (Bourguet et al., 2000).

Hormone receptors

Classical hormone receptors, such as the vertebrate estrogen receptor (NR3A) or the insect ecdysone receptor (NR1H1) bind only to a few very specific molecules with high specificity and high affinity, in the nanomolar range. Their ligands are generally synthesized endogenously.

Nutritional sensors

Another more recently defined category, that comprises the oxysterol receptors (LXR/NR1H2-3) and the bile acid receptors (FXR/NR1H4-5) are able to bound more loosely, with affinity in the micromolar range, to a variety of molecules, that are often food derivatives. Hence was coined the term ”nutritional sensor”, implying that these receptors are able to sense the metabolic state of the cell and to regulate its balance (Benoit et al., 2004).

Constitutive activators

Constitutive activators are receptors that can activate transcription in transfected cells even in the absence of an exogenously added ligand. Classical examples from this group are the ERR (NR3B). It is possible that such receptors are really activated by a yet unknown endogenous ligand (Benoit et al., 2004).

Receptors with ligands as structural cofactors

Receptors with structural cofactors are receptors that are constitutively active in absence of exogenous ligand, but that require an endogenous lipid molecule to take their conformation. This is for example the case for RORα (NR1F), which is stabilized by

the binding of cholesterol, and such a situation also appears with the USP (NR2B) in dipterans (Clayton et al., 2001) and lepidopterans (Billas et al., 2001).

Receptors without a ligand-binding pocket

Receptors without a ligand-binding pocket are receptors where the space that is occupied by the ligand in liganded receptors is filled by the amino-acids of the surrounding helixes. The most salient examples are members from the NR4 family (Wang et al., 2003). They are nevertheless regulated during signaling cascades, but this regulation goes through post-translational modifications, such as phosphorylation (Rochette-Egly, 2003), as it is the case for mainly unliganded transcription factors.

2.3.5 Variations in nuclear receptors mechanism of action

The canonical mechanism of action that is described here is not the only way nuclear receptors play a role in cell signaling. They also can be cofactors modulating the activity of another transcription factor. For example, the glucocorticoid receptor (GR) in vertebrates is known to repress through direct interaction the activity of the transcription factor NF-κB, a central transcription factor in the inflammatory response (Necela and Cidlowski, 2004). More surprisingly, they are also involved in signal transduction in the cytoplasm through non-genomic pathways, triggering kinase cascades (Ordóñez-Morán and Muñoz, 2009).

2.3.6 Function at the organismal level

Nuclear receptors are involved in a considerable number of developmental and physiological processes (Huang et al., 2010). In insects, their role has been well characterised in embryo segmentation, development of the nervous system, moulting and metamorphosis (King-Jones and Thummel, 2005). In vertebrates, they are also involved in so many various processes that it is difficult to propose a satisfactory classification of their functional roles.

Classically, nuclear hormone receptors, such as steroid receptors or thyroid hormone receptors, are associated with the functioning of the regulatory axes. Here we present the case of the gonadal steroid axis (Fig. 13).

In such an axis, external or internal signals are integrated through the brain, in which some neurons from the hypothalamus secrete a peptidic hormone that stimulates the hypophysis. After that, hypophysal neurones secrete a second neurohormone that goes into the general circulation and triggers the synthesis of steroids in the gonad. The gonadal steroids that are produced by that way are also released into the blood, where they trigger a variety of biological responses on various cell types, including some brain cells. The retroaction of gonadal steroids on the brain is currently viewed as a feedback retroaction (Norris, 2007).

However, this scheme cannot be universally valid for all nuclear receptors, especially for those who are not ligand-activated. Some researchers have yet tried to propose a functional classification of all the mammalian nuclear receptors, proposing that each of the mammalian receptors fits in one of the following categories: reproduction, development, central and basal metabolic functions, dietary-lipid metabolism, and energy homeostasis (Bookout et al., 2006). This classification was based on the results of high-throughput analysis of the target genes of all mouse nuclear receptors. However, such a classification is somewhat arbitrary. For example, TRβ, that is involved in gut remodeling during

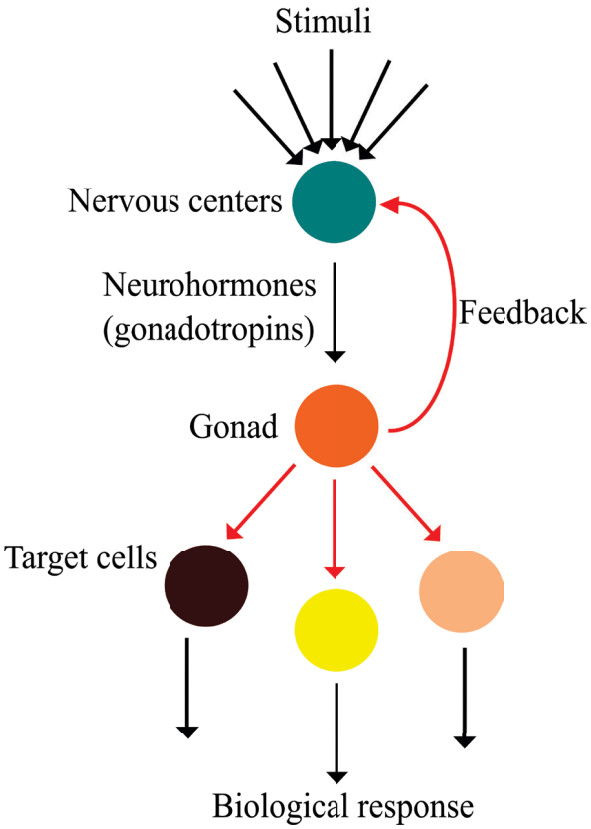


Figure 13: A simplified version of the gonadal steroid signaling axis. Red arrows indicate in which steps of the signaling cascade steroid hormone signaling through nuclear receptors is involved.

weaning of mouse pups, is classified in the category ”dietary-lipid metabolism, and energy homeostasis”, but could also well fit in the ”development” category. Here as for many other questions, an evolutionary perspective could be instrumental to design a good functional categorization. Again, this implies that we have to understand in which order these various physiological processes appeared and how they interacted during evolution.

Chapter 3

Diversity of metazoans regarding intercellular communication

The physiological regulations that we have previously presented are occurring in animals with a complicated internal organisation, that is not universal. To put steroid signaling evolution in an appropriate framework, it is necessary to have a brief overview of metazoan morpho-anatomical diversity and phylogeny.

Here we will provide a quick tour on metazoan diversity, focussing on structural features that can play a role in intercellular communication.

3.1 Metazoan phylogeny

The necessity to class living beings according to their relative degrees of kinship was already expressed in the first edition of *The Origin of species* (Darwin, 1859), but the conceptual tools necessary to make this efficiently have been lacking during almost an entire century.

Until the 1970's, the classification of animals was an highly confuse and controversial working area, with no consensus about the real relationships between the animal groups (Jenner, 2000). This has spectacularly changed during the last decades due to two main factors.

The first was the cladistic revolution, triggered by the publication of an english translation of Willi Hennig's work about phylogenetic systematics ((Hennig, 1950), (Hennig, 1966)). Before that, animals were classified mainly on the basis of global similarity, and the characters taken into account to define this similarity were highly arbitrary. Hennig introduced a distinction between shared ancestral characters (symplesiomorphies) and shared derived characters (synapomorphies), compared to an external reference, the outgroup, that indicates the common ancestral state (Fig. 14).

Of course, characters are not always shared due to common descent, but can also arise independently due to similiar functional constraints or selective pression. This process in named convergence, or recruitement, at the molecular level. Homology, the hypothesis about common ancestry of similar characters in two taxa, is always an hypothesis, that has to be tested by phylogenetic reconstruction. When the test fails, characters that were acquired convergently in different lineages are named homoplasies (Lankester, 1870).

The second major factor that drastically modified metazoan phylogeny was the increasing abundance of molecular data, that made possible large-scale comparisons of taxa that are very different morphologically, and that provided a very abundant source of characters (Adoutte et al., 2000). This made possible for the first time to acquire a

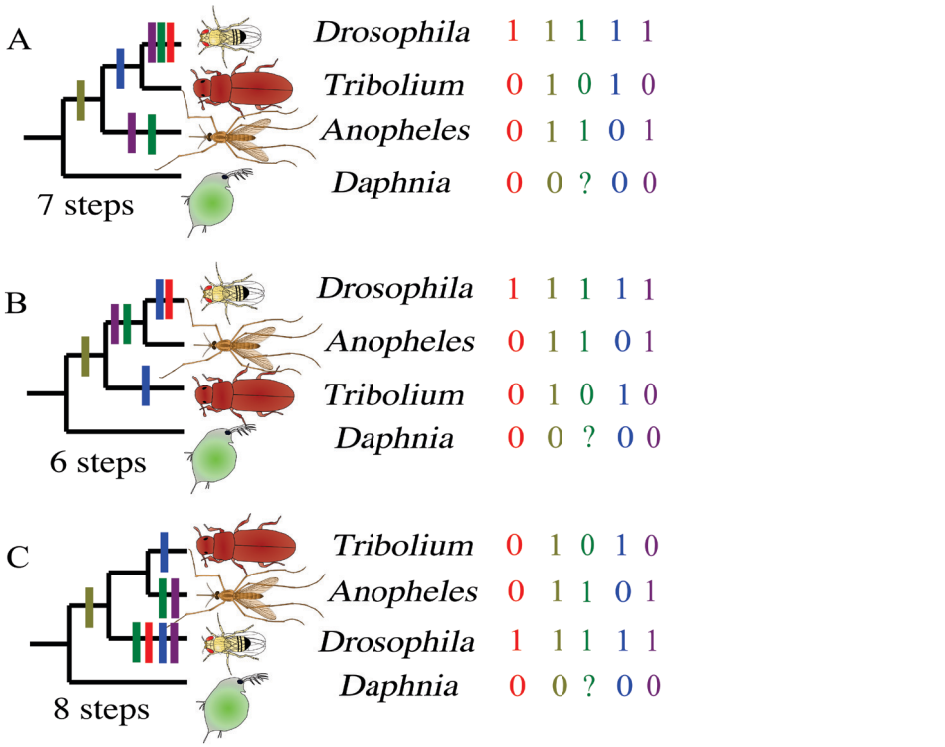


Figure 14: An example of phylogenetic classification established using cladistics. The three possibilities regarding the relationships between *Drosophila*, *Tribolium* and *Anopheles*, compared to the outgroup *Daphnia* are represented by three possible trees (A, B and C). In each case, the number of steps (changes in character state) necessary to explain the character distribution pattern is indicated on the branches. With only six steps in comparison to seven steps for trees A and eight steps for tree C, tree B is the most simple way to explain the given character distribution. The character distribution on tree B implies that the presence of imaginal discs in *Drosophila* and *Tribolium* is an homoplasy, and that this structure appeared independently in both animals (Svácha, 1992).

consensus view about the relationships between metazoans.

The Fig. 15 gives the currently accepted consensus view about the relationships between the discussed animal groups. Not all groups are presented, because we concentrated on those where there is consistent genomic knowledge, with at least one fully sequenced genome. Some points, such as the monophyly of sponges or the exact position of placozoans, are still debated, but we will not discuss this here.

3.2 Choanoflagellates, the sister group of metazoans

To understand what is a metazoan, it is necessary to know what is its closely-related group, in order to identify which characters are really specific to metazoans. The sister group of metazoans is choanoflagellates. These are small mainly unicellular organisms, and about 120 species are currently known. A choanoflagellate cell has a collar made from microvillosities, and in its center a flagellum beats, thus creating a water current that

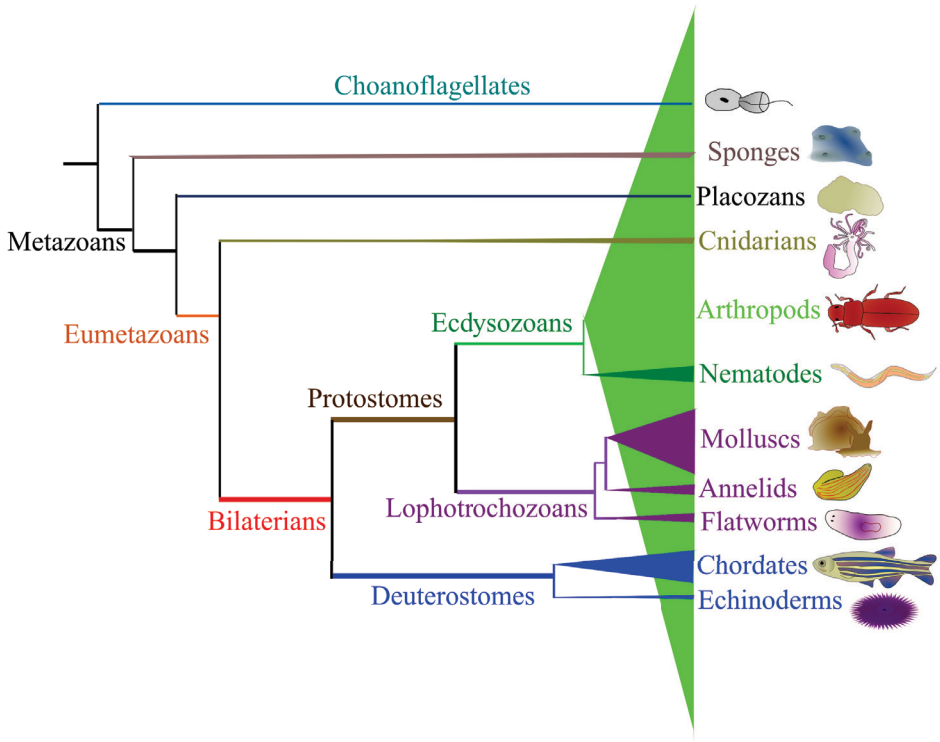


Figure 15: Current consensus about metazoan phylogeny. The size of the clades is proportional to the number of described species.

attracts bacterias eaten by the cells (Fig. 16A). There can be benthic or planctonic, and they live mainly in the sea, but also in freshwater environments, where there are able to form cysts in response to overcrowding (Leadbeater and Karpov, 2000). Within the group, the ability to synthesize an extracellular matrix (the lorica) and the presence of a colonial stage during the life cycle are quite widespread, and may even be ancestral characters, that were present in the last common ancestor of choanoflagellates and metazoans (Carr et al., 2008). Observation of the transition from unicellular state to multicellularity by cell division was observed in *Salpingoeca rosetta*, and the asynchronous division pattern suggests that cell division is not coordinated between sister cells in a colony (Fairclough et al., 2010). The sequencing of the genome of *Monosiga brevicollis* revealed that they differ from animals by the absence of some protein families such has T-box and ETS transcription factors and nuclear receptors (King et al., 2008). Preliminary studies on its sterol profile reveal the ability to synthesize sterols with a structure that is intermediate between this of fungi and this of metazoans (Kodner et al., 2008).

3.3 Sponges

Sponges are sessile metazoans that have water intake (the pores) and outake (the osculum) openings connected by chambers lined with choanocytes, cells with whip-like flagella, whose organisation is strikingly similar to choanoflagellates (Fig. 16B and C).

All known living sponges can remold their bodies, as most types of their cells can move within their bodies and a few can change from one type to another. The existence of complex intercellular coordination of cell movements and differenciation is well established, in particular during asexual budding (Hammel et al., 2009), and for the

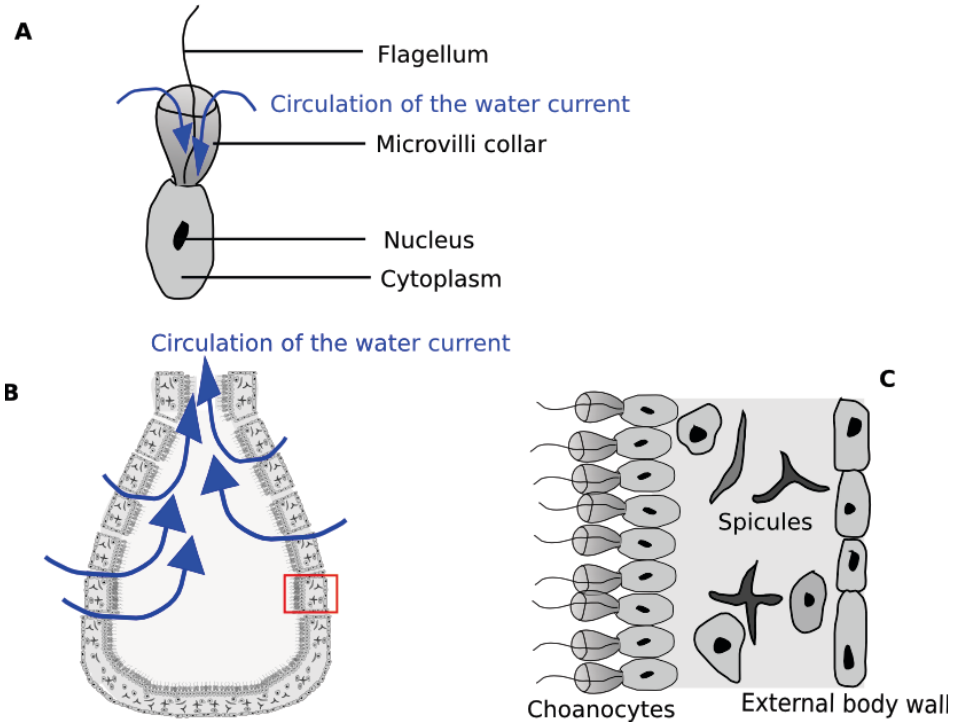


Figure 16: Similarity in cellular organisation of choanoflagellates and sponge choanocytes. (A) An isolated choanoflagellate cell, without lorica. (B) Transversal section of a sponge, showing the circulatory current. (C) Focus on a portion of the sponge body wall, showing the rows of choanocytes that are the feeding cells.

control of the opening of the osculum through contraction of myocytes. They have no clearly delimited tissues and no internal body fluid with a strictly controlled composition. Trophic exchanges between cells occur mainly through the amebocytes, which are mobile and the most plastic cell type. Some species have huge sizes and complex structures, so probably most of the intercellular communication occurs not through an ”internal milieu”, but through releasing of signaling molecules in the cell neighborhood (exocrine communication), and through intracytoplasmic communication.

Demosponges are able to produce a peculiar sterol 24-isopropylcholesterol, which seems to be specific for them (Kodner et al., 2008), and they produce a huge variety of oxysterols with anticancer properties. Proliferation can be modulated in a reversible manner by retinoic acid, and the expression level of a nuclear receptor is known to be activated by that (Wiens et al., 2003), which does not necessary imply that retinoic acid is the physiological ligand of that receptor, or even that the activation is direct. No link between steroid signaling and nuclear receptor was reported for the moment.

In *Amphimedon queenslandica*, two nuclear receptors were identified (Bridgham et al., 2010). AqNR1 is expressed specifically in the ciliar cells of the external layer, which transdifferenciate into choanocytes after metamorphosis, and also in cells associated with the pigment ring, a ring of ciliated cell at the posterior part of the embryon, that is involved in its locomotion (Larroux et al., 2006). AqNR1 is able to activate gene transcription after binding of a wide range of bacterial free fatty acids (Bridgham et al., 2010). AqNR2 is also able to bind fatty acids but not to activate the transcription after ligand binding, and is expressed quite ubiquitously during larval development

(Bridgham et al., 2010).

Currently, only one sponge genome is sequenced, this of the demosponge *Amphimedon queenslandica* (Srivastava et al., 2010), and it is not clear yet, whether sponges are truly monophyletic or if other metazoans are nested in a peculiar sponge subgroup.

3.4 The enigmatic placozoans

The group "placozoans" was coined for only one animal species, *Trichoplax adhaerens*. Trichoplaxes are very flat animals around a millimeter in diameter, lacking any organs or internal structures. They have three cellular layers: the top epitheloid layer is made of ciliated "cover cells" flattened toward the outside of the organism, and the bottom layer is made up of cylinder cells which possess cilia used in locomotion and gland cells which lack cilia. Between these layers is the fiber syncytium, a liquid-filled cavity strutted open by star-like fibers (Schierwater, 2005). The digestion is made in a temporary cavity formed by contraction of the ventral side around the feeding particle. The absence of a basal lamina makes possible the direct entry of the particules into the body cavity.

Its life cycle and sexual reproduction are not known, but its genome is fully sequenced (Srivastava et al., 2008), showing the presence of four different nuclear receptors (Baker, 2008), but there is currently no functional data about them. Nothing is known about steroid metabolism. However, the presence of lipid-accumulating cells on the dorsal side of the animal and the observation that placozoan may repel predators or even kill them if ingested (Pearse and Voigt, 2007) suggest that may be able to synthesize a variety of lipid molecules that can disrupt some metabolic pathways of potential predators.

3.5 Eumetazoans

Eumetazoans are animals whose body is divided into germ layers. They have true epithelia delimited by a basement membrane. This has important consequences on nutrition, because food particles can not go through the digestive epithelium, which so determinates the formation of a gut, with extracellular digestion. A corollary is a more strict division of labour between cell types, with much less plasticity in terms of cellular differenciation ability, and with the appearance of true muscular and neuronal cells. With the neuronal cells appears the possibility of long-distance intercellular communication through nervous and neuroendocrine pathways.

The two major groups of eumetazoans are cnidarians and bilaterians.

3.6 Cnidarians

Cnidarians are animals that have peculiar urticant cells, the cnidocytes, that allow them to catch their preys. They have no internal circulatory system, but they are highly differentiated animals, with complex life cycles, and sometimes colonial structures. In the colonial species, such as corals, trophic connections occur through the gut.

Cnidarians have quite important regeneration abilities, and are able to produce steroids that are used as anticancer compounds. The application of exogenous steroids is known to disrupt the reproductive physiology at least of anthozoans (Twan et al., 2003), and it is known that cnidarians have some nuclear receptors, but no functional link is known between those partners.

3.7 Bilaterians

Bilaterians are animals where the body has a bilateral symmetry axis. They have an internal body fluid, which composition is regulated by an excretory system, but with no systematic division between an internal circulating fluid (the blood) and the intercellular fluid (the lymph). They all have differentiated glands.

The two major groups of bilaterians are deuterostomes and protostomes.

3.8 Deuterostomes

Deuterostomes are animals where the blastopore forms the anus during gastrulation. A debated synapomorphy is the presence of pharyngeal pouches (Fig. 17), from which many glands derive, and from the endostyle or thyroid gland, that is a major partner in crosstalks with steroid signaling at least in chordates. In tetrapods, the parathyroid glands also derive from pharyngeal pouches. The parathyroid glands produce the parathyroid hormones, that stimulate the last step of 1,25-dihydroxyvitamin D3 synthesis (Norris, 2007).

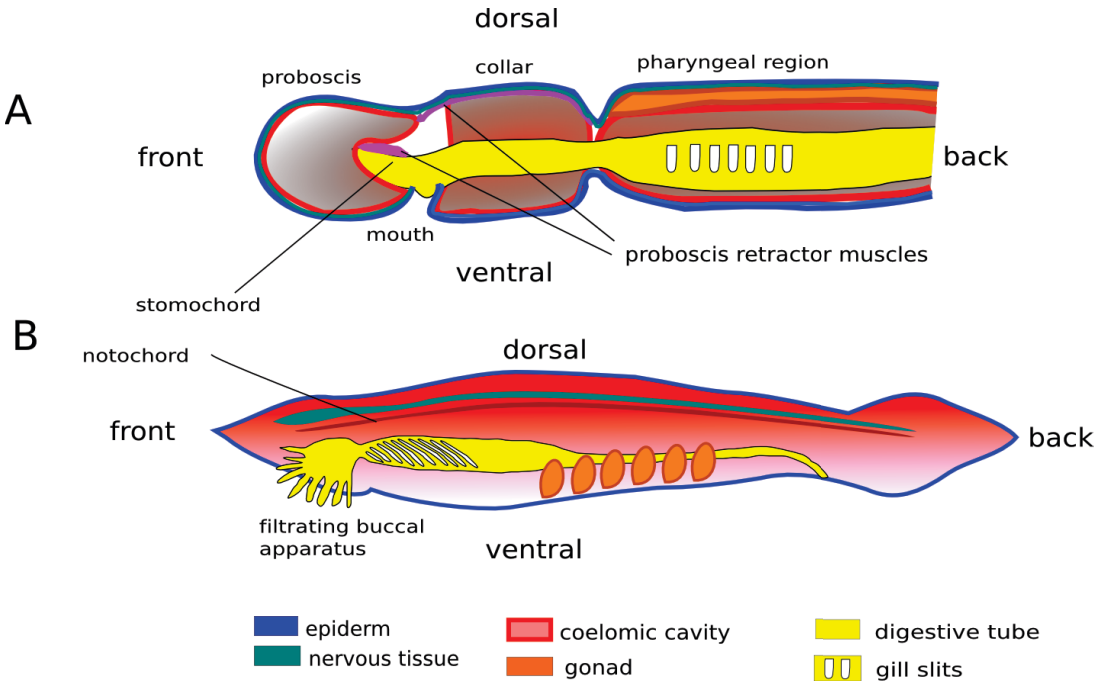


Figure 17: Comparison of the body structure of two deuterostomes, an enteropneust (A) and an amphioxus (B). In both cases, the anterior part of the gut wall bears holes that are called gill slits. In vertebrates, the thyroid gland develops from the pharynx wall between the first and second pouches.

There are two major groups of deuterostomes: ambulacralians and chordates.

Ambulacralians comprise hemichordates and echinoderms. Hemichordates, which comprise the enteropneust portrayed on Fig. 17A, are poorly studied, and almost nothing is known about their physiology. We mention them here only because they are the only living ambulacralians that have gill slits and an endostyle, whereas echinoderms are supposed to have secondarily lost these characters (Ruppert, 2005).

Echinoderms are highly derived animals with a pentaradial symmetry. They have an open circulatory system, but its role in body fluid circulation seems to be greatly diminished by the presence of a water vascular system, that is implicated in locomotion, food and waste transportation, and respiration. This water system is connected to the sea water (except for sea cucumbers), and so cannot be viewed as a true internal body fluid, but could be an important vector of long-distance signaling molecules between distant parts of the echinoderm body. They are various reports of vertebrate-type steroids in these animals but no proof of endogenous synthesis (Lafont and Mathieu, 2007).

Chordates have a closed circulatory system, dividing the body fluid into blood in the vessels and lymph in the rest of the body. Among them, vertebrates are the best studied group, comprising many endocrinological models, where steroid binding by nuclear receptors is well established, as we detailed previously. In urochordates and cephalochordates (comprising the amphioxus portrayed above), the data are more incomplete. A receptor that was able to bind estrogens was identified in the floridan amphioxus (Bridgham et al., 2008), and there have been claims about a complete estrogen-synthesizing pathway in this animal (Mizuta et al., 2008). We will discuss this in full at various places of this manuscript, especially in results part III and IV.

3.9 Protostomes

Protostomes are animals where the blastopore forms the mouth during gastrulation. The two major groups of protostomes are ecdysozoans and lophotrochozoans.

3.10 Ecdysozoans

Ecdysozoans are animals that have a discontinuous growing cycle, punctuated by molts (Aguinaldo et al., 1997). The two ecdysozoan groups that are well known from a genomic viewpoint are arthropods and nematodes.

Arthropods have an open circulatory system, where there is no blood compartment with specific composition. The hemolymph that baths the organs is moved by the contraction of the heart. In arthropods, the nuclear-receptor mediated steroid signaling plays a central role, because molting is triggered by an ecdysone peak, and the ecdysone binds to the nuclear receptor EcR (Koelle et al., 1991).

Nematodes have no circulatory system at all. Moves in the body fluid are further limited by the absence of circular muscles in the body wall, that makes dorso-ventral contractions not possible. In nematodes, the ortholog of the insect ecdysone receptor is not universally conserved and molting seems to rely on a partially different regulatory mechanism (Thummel, 2001). But nuclear receptor steroid signaling through dafachronic acid is known to be at the core of regulation of dauer entry (Antebi et al., 1998) in various clades and mouth polyphenism in the diplogastrid *Pristionchus pacificus* (Bento et al., 2010). In some parasitic nematodes, the entry into the infestation L3 stage, that is believed to be homologous to the dauer stage, is also regulated by dafachronic acid signaling (Wang et al., 2009).

3.11 Lophotrochozoans

Lophotrochozoans are an highly heterogenous grouping of animals that was defined on a molecular basis (Halanych et al., 1995). Currently it is difficult to have a unified view on

the common specificities of all members of this group because many of them are poorly known at both the anatomical and molecular level. The three mainly studied groups are annelids, mollusks and flatworms.

Annelids have a closed circulatory system. Recently, a candidate ortholog of vertebrate estrogen receptor was cloned and caracterised in two model annelids. It can be activate by vertebrate estrogens (Keay and Thornton, 2009), but the physiological meaning of such a finding is not yet clear. Reports on annelid estrogens are numerous but evidence for de-novo synthesis of such molecules is still lacking (Lafont and Mathieu, 2007).

Mollusks have an open circulatory system, except from cephalopods where it is closed. Many candidate orthologs of the vertebrate estrogen receptor were caracterised, but none of them binds estrogens. As for annelids, the data on vertebrate-type steroids in molluscs are not reinforced by data on their biosynthesis (Markov et al., 2008b).

Platyhelminths, or flatworms, have no circulatory system. Their digestive system is very specific, with only one opening, and many secondary branchings that bring directly the nutrients to various cell types. The body cavity is totally filled with mesenchymal cells with no space for a circulating body fluid.

Chapter 4

Evolution of ligand-binding ability for nuclear receptors

4.1 Nuclear receptor phylogeny and the origin of the ancestral orphan hypothesis

The way of regarding the relationships between NRs and their ligands has been historically biased by the fact that the first identified ligands were mammalian steroid and thyroid hormones, hence the name often given to the family: the steroid/thyroid hormone receptor family (Evans, 1988). When EcR (NR1H1), the receptor for the insect hormone ecdysone was identified in *Drosophila*, it was clear to everyone that NRs were high affinity receptors (at the nanomolar range) for hormones in all animals (Koelle et al., 1991). The question then arose of the origin of the ligand-binding ability. Based on the observation that two kinds of major NR ligands, steroids and retinoids, are products of terpenoid metabolism, the first hypothesis was that the ancestral receptor would have been liganded by a terpenoid molecule (Moore, 1990). At that time, the only known non-terpenoid NR ligands were thyroid hormones, and there were many candidate ligands within terpenoid molecules with signaling roles in various eukaryotes, such as juvenile hormone in insects or abscissic acid and gibberellins in plants. In this context, the first orphan receptors that were cloned were mainly considered as receptors waiting for a yet unknown high-affinity ligand (Giguère et al., 1988). When the possibility was raised that orphans could be constitutively active, the hypothesis that ligands could be derived from intracellular metabolism, was preferred (O'Malley, 1989).

With the first phylogenies of the family, it became clear that all NRs share a common ancestor, and it became possible to speculate on the ancestral state of the first nuclear receptor ((Amero et al., 1992), (Laudet et al., 1992)). This led to the proposal that the evolution of the ligand binding specificity of NRs involved several independent gains and losses of ligand-binding ability from an ancestral orphan (Fig. 18 , (Escriva et al., 1997), reviewed in (Escriva et al., 2000), and (Baker, 2003)). This view was supported by three types of arguments.

First, orthologs of classically liganded vertebrate receptors, such as TR (NR1A), RAR (NR1B) and steroid receptors from the NR3 family (ER, GR, MR, PR, AR) were not identified outside vertebrates, suggesting a late appearance of liganded receptors during animal evolution. It was later shown that homologs of these receptors are present in some mollusks and platyhelminths, but to date it remains true that they are not present outside bilaterians. This means that the early steps of NR diversification, leading to the common set of 25 bilaterian NRs (Bertrand et al., 2004) may have taken place in a

context where classical hormone receptors did not exist, as animals living at that time did not have an internal circulatory system linking differentiated organs. However, this does not necessary imply that the first receptors were true orphans. Among the NRs that exist in cnidarians, there are HNF4 (NR2A), which has a ≪structural ligand≫ in mammals, that may mirror the ancestral situation ((Sladek, 2002); (Benoit et al., 2004); (Yuan et al., 2009)), and also RXR (NR2B), which can be viewed as a sensor (discussed below) due to the broad diversity of its ligands ((Mic et al., 2003); (Calléja et al., 2006)).

Second, the hypothesis that orphan receptors evolved early also raises the questions of the mechanism that would allow them to be activated, and what structural constraints allow for the ligand binding domain (LBD) to be conserved in orphan receptors. In 1997 came the first results of NR regulation through conformational changes in a ligand-independant way, for example due to phosphorylation (Rochette-Egly, 2003), or other types of post-translational modifications. This provided an explanatory mechanism for the regulation of orphan receptors by something other than ligand binding, which could explain the structural conservation of the LBD in absence of ligand binding. This fact has gained increasing support: it is clear that the ligand is just one possible trigger for the conformational change or more appropriately termed the allosteric transition (Faus and Haendler, 2006).

The third argument was that there seemed to be nothing in common between the synthesis pathways of diverse ligands such as thyroid hormones, retinoic acids and steroids. This should be now reassessed in light of our more complete understanding of the wide diversity of NR ligands (Fig. 18, and Sladek, 2011). For example, even if some of these data still need to be confirmed, retinoids can apparently bind mammalian receptors other than RAR (NR1B) and RXR (NR2B), with comparable affinity for PPARβ/δ (NR1C2) and lower affinities for the two other mammalian PPARs (NR1C), RORβ (NR1F2), or COUP-TFII (NR2F2) (Theodosiou et al., 2010). NR2B (RXR/USP) is sensitive to a wide range of different molecules, not only retinoids but also fatty acids and phospholipids, and, in fact, retinoid binding to RXR may have no relevance in vivo ((Mic et al., 2003); (Calléja et al., 2006)). PPARs bind fatty acids and their eicoisanoid derivatives. Furthermore, there is evidence for crosstalk between retinoic acid and fatty acid signalling based on alternate activation of RAR or PPARβ/δ depending on retinoic acid concentration (Schug et al., 2007). It is worth noting that retinoids are transported to cells as esters and that the dissociation of the ester produces not only retinoic acid, but also releases a fatty acid. This raises the possibility that other unknown cross-talks due to the sharing of one of many ligands may connect the different NR-signaling pathways. Rev-erbs (NR1D1/2) and their ortholog in insects, E75 (NR1D3), were shown to bind hemes, a very big molecule when compared to other ligands in mammals, as in *Drosophila* (Burris, 2008). Recent data indicate that other NRs, such as RXRα (NR2B1) in mammals (Gotoh et al., 2008) or HR51 (NR2E3) bind heme with micromolar affinity, whereas HNF4 (NR2A) and HR83 (NR2E5) bind heme with lower affinity in *Drosophila* (de Rosny et al., 2008). This suggests that heme could be a more widespread ligand than previously expected. Even historical ≪hormone-receptors≫, can be activated by a variety of ligands. For example, the Vitamin D Receptor (NR1I1) is also activated by a bile acid (Makishima et al., 2002), and 5α-androstane-3β,17β-diol appears to be a natural agonist of the estrogen receptor ERβ (NR3A2) (Weihua et al. 2002). On the other hand, many steroid hormones activate the ≪xenobiotic receptor≫ PXR (NR1I2) at a micromolar range (Ekins et al., 2008). Therefore, generally there is no exclusive pairing between ligands and receptors: one ligand can activate many receptors and one receptor can be activated by many ligands. Furthermore, a receptor can be both a sensor or a liganded receptor, depending on the context (Fig. 18).

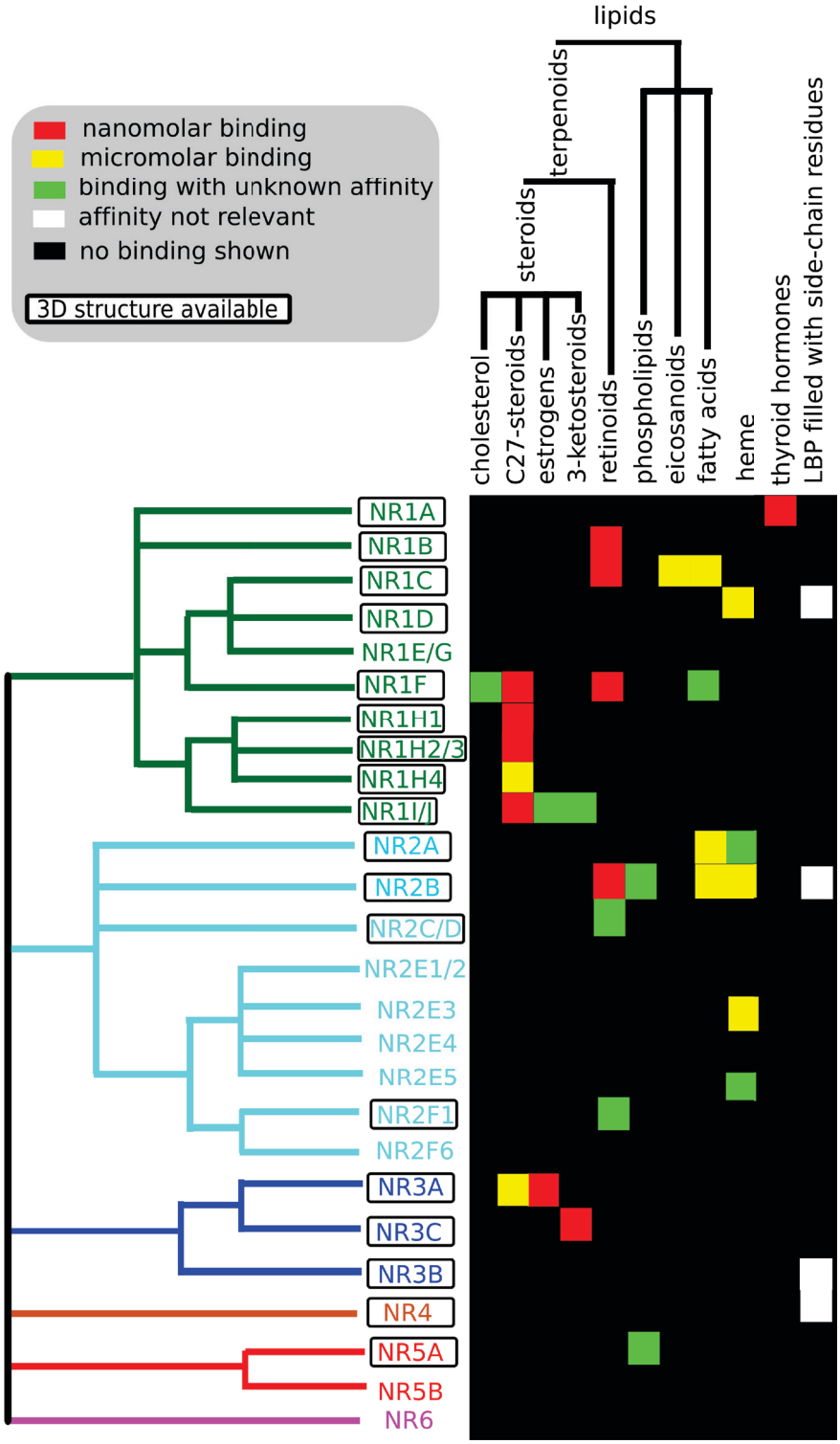


Figure 18: Phylogenetic distribution of NR ligands. A phylogeny of putative bilaterian NRs, adapted from Bertrand et al., 2004. The presence of a ligand in front on the receptor does not mean that all the orthologs bind it, but only that the binding was shown for some of them. Boxed receptors are those for which at least one crystal 3D structure is available. The binding ability for each type of ligand is indicated by the colour code. Red: nanomolar affinity; yellow: micromolar affinity; green: binding without data on the affinity, or affinity criterion not relevant (for pockets filled with amino-acid side-chains).

Even if the various NR ligands are members of different chemical families from a nomenclature viewpoint, they share common physical properties, such as hydrophobicity and a volume between 250 and 550 $Å^3$. Indeed, there are other proteins families that interact with the NR ligands (Fig. 19), where different paralogs bind different ligands. Retinoic acid, retinol, fatty acids, eicosanoids, heme and bile acids are all bound by proteins of the FABP family (Zimmerman and Veerkamp, 2002), that are involved in their intracellular transport, whereas the extracellular albumins bind all kinds of NR ligands (Baker, 2002). Short-chain dehydrogenase reductases (SDR) are also known to metabolize both retinoids and steroids (Baker, 2001). Proteins of the CYP family are involved in the metabolism of retinoids, steroids, fatty acids, eicosanoids and xenobiotics. Important substrate shifts between closely related paralogs are also well known (Brown et al., 2008). For example, CYP8A1 and CYP8B1, which are the result of a vertebrate-specific duplication, metabolize respectively, prostacyclin (an eicosanoid) and a bile-acid precursor (Thomas, 2007). These shifts in substrates may indicate that these ligands share some common properties allowing a rapid switch from one ligand to another on an evolutionary timescale.

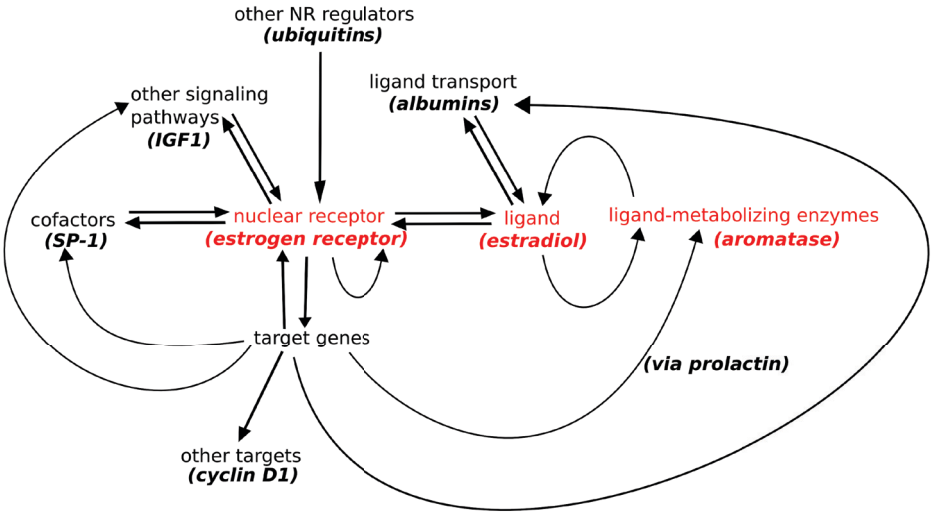


Figure 19: The interaction network between NRs and their ligands. The basic interaction network in which NRs are involved comprises NR target genes, for which transcription is activated or inhibited by NRs, cofactors, that bind to NRs during activation or repression, other signalling pathways, that led to post-transcriptional modifications of NRs, and ligands, whose presence is controlled by ligand-metabolising enzymes, and ligand transport proteins, intracellular such as FABP or extracellular such as albumins. NRs can also act on other signalling pathways through non-genomic mechanisms, that do not involve their binding to DNA. Members of the estrogen receptor network are indicated as an example.

Finally, the orphan hypothesis raises the question on how an unliganded transcription factor shifts to a ligand binding receptor. This can only be addressed by focusing on some precise case studies.

4.2 Origins and evolution of the steroid receptors and their implications on the ancestral NR

There are many well studied cases of variations on binding ability between similar ligands ((Bridgham et al., 2006); (Escriva et al., 2006); (Paris et al., 2008); (Reschly et al., 2008a)), and some cases of transition from a liganded receptor to an orphan ((Krylova et al., 2005); (Iwema et al., 2007)) or transition from an orphan to a receptor with a structural ligand (Iwema et al., 2007), but these studies do not deal with the transition from an orphan to a liganded receptor. This is partly due to the fact that, except for insects and nematodes, functional data on non-vertebrate animals are still scarce, and this is further complicated by the lack of data regarding the physiological significance of some putative ligands, and by uncertainties regarding the topology of the NR trees (see the numerous polytomies on Fig. 18). The only families where there is sufficient genomic sampling and understanding of the physiological significance of the ligands are the NR3 subfamily (Fig. 21) and the NR1H/I/J group (Fig. 22), that contain the steroid receptors, making them the best proxies to address the question of ligand binding acquisition.

4.2.1 Hypotheses on the binding ability of the ancestral steroid receptor in the NR3 subfamily

The NR3 family contains receptors for vertebrate sex and adrenal steroids, but also some mollusk receptors that do not bind sex and adrenal steroids. However, a resurrected ancestral estrogen receptor was found to be activated by estrogens, implying that estrogen binding would have been secondarily lost in some mollusks (Thornton et al., 2003). This was followed by the further characterization of mollusk constitutive activators in this family (Keay et al., 2006), whereas secondary losses of ligand binding-ability was further documented in the rodent LRH-1, a receptor from the NR5A2 group (Krylova et al., 2005). Taken together, these data were interpreted as evidence against the ancestral orphan receptor theory, and it was proposed that, on the contrary, constitutive activation has evolved several times in parallel from a ligand-dependant nuclear receptor ancestor (Keay et al., 2006).

Hypotheses about steroid binding in the NR3 subfamily are strongly dependent on the understanding of the relationships between the various receptors (Fig. 20). The NR3 subfamily diversified specifically in bilaterians (Baker, 2008). The first duplication produced the common ancestor of bilaterian ERR (NR3B) and the common ancestor of the bilaterian steroid receptor, AncSR1 ((Thornton, 2001); (Thornton et al., 2003)). There is also clear evidence that vertebrate ER has an ortholog in amphioxus (Paris et al., 2008), and that another amphioxus receptor, named SR (Bridgham et al., 2008), is orthologous to the vertebrate ancestor gene that gave rise to the current GR (NR3C1), MR (NR3C2), PR (NR3C3) and AR (NR3C4).

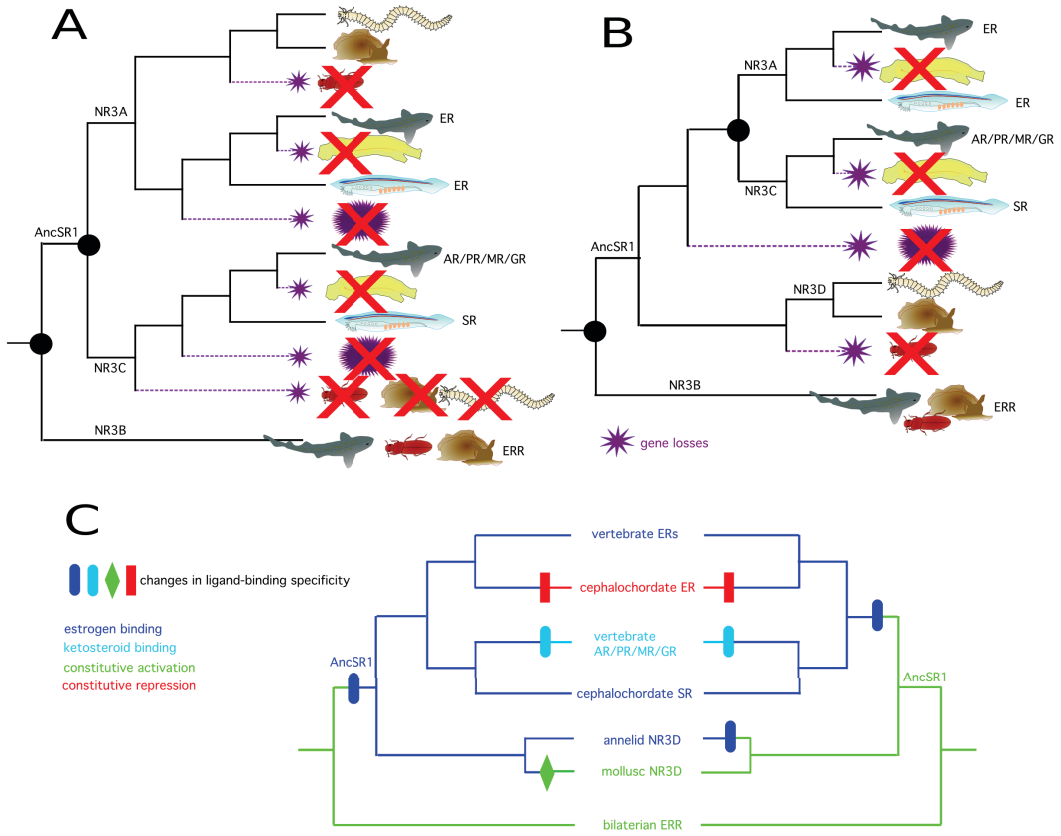


Figure 20: Uncertainities in parsimony-based scenarios about the ligand-binding abilities of the ancestral steroid receptor (AncSR1). (A) If there was a gene duplication of AncSR1 leading to NR3A and NR3C in the common ancestor of all bilaterian animals, six independent gene losses are necessary to explain the current gene distribution. These losses are indicated by the purple stars at the end of dotted branches. (B) If NR3A and NR3C are the products of a chordate-specific gene duplication, four gene losses would be sufficient to explain the observed gene distribution. Thus, this scenario is more parsimonious than (A), but would require the renaming of lophotrochozoan ≪ER≫ as NR3D, to stress that they are not more related to vertebrate NR3A than to vertebrate NR3C. (C) Under the tree topology presented in (B) and taking into account the fact that ERR is a constitutive activator, acquisition of steroid binding in all bilaterians (on left side) or convergent acquisition in chordates and annelids (on right side) are both equiparsimonious scenarios. In both cases, constitutive repression in cephalochordate ER (red box) and a shift in ligand specificity in vertebrate AR/PR/MR/GR (light blue oval) would be acquired from a liganded receptor. On the contrary, the early acquisition of estrogen-binding in AncSR1 (dark blue oval on left) would necessitate a reversion to constitutive activity in mollusk NR3D.

Things become more complicated when dealing with the recently cloned NR3 in mollusks and annelids. The first analyses of mollusk sequences provided support for their grouping with the chordate ER (NR3A), and they were also named NR3A in some publications ((Thornton et al., 2003); (Paris et al., 2008). This topology is shown in Fig. 9A. But the addition of two cloned annelid receptors decreased the support for this grouping (Keay and Thornton, 2009). In this paper, the authors acknowledged that they ≪cannot rule out the possibility that the protostome ERs could be equally orthologous to the entire SR family ≫(topology shown in Fig. 9B).

Two additional lines of evidence may be worth taking into account to discriminate between both scenarios. First, when one considers the minimal number of secondary gene losses, it appears that there would have been at least six independent losses if lophotrochozan ≪ERs≫ are orthologous to NR3A (Fig. 20A), whereas if it were orthologous to the chordate gene that gives rise to NR3A and NR3C after duplication, there would have been only four losses (Fig. 20B). This second scenario is therefore more parsimonious. The second line of evidence is that vertebrate NR3C has undergone an acceleration of evolutionary rate, which has led to reconstruction artifacts at the base of NR3 phylogeny. At that time, when only mammal sequences where available, NR3C branched basally to a group containing NR3A and NR3B (Laudet, 1997). In fact, similar artifacts are also present for other NR subfamilies, even when methods that diminish the effect of long-branch attraction, such as maximum-likelihood, are used. For example, a recent paper on the phylogeny of bilaterian RXR/USP (NR2B) showed nematode, platyhelminth and mecopteridan insect sequences branching erroneously at the basis of bilateria, and this was interpreted using the same reasoning that favours the topology in Fig. 20B (Tzertzinis et al., 2010).

Whatever the true topology, this debate raises an important nomenclature issue. In such ambiguous cases, it would be preferable to give the controversial sequence a name that does not favour one or the other hypothesis. This is why we suggest that mollusk and annelid sequences, that where up to now unofficially designated as ≪ER≫ or ≪NR3A≫ should preferably be named ≪NR3D≫, as we do in Fig. 20B and 20C, and as it has already been done for other ambiguous cases, such as vertebrate VDR/PXR/CAR (NR1I) and insect HR96 (NR1J) (NRNC, 1999; see also Fig. 21). This formal problem should not be underestimated, given the fact that non-neutral names can significantly bias further experimental research (Markov et al., 2008a). In spite of the mentioned uncertainties, and maybe due to the nomenclature bias, the evolution of the steroid binding ability in the NR3 family was, to date, only discussed based on the topology presented in Fig. 20A. It was proposed that the early acquisition of estrogen-binding in bilaterians was more parsimonious than the late convergent acquisition from a constitutive activator occurring three times independently in vertebrates, cephalochordates and annelids ((Keay and Thornton, 2009), (Eick and Thornton, 2011)). However, if we take into account the topology proposed on Fig. 20B and the fact that ERR is a constitutive activator in mammals and mollusks ((Giguère et al., 1988); (Bannister et al., 2007)), we find that the early acquisition of estrogen binding in bilaterians (Fig. 20C, left) and the late acquisition of estrogen binding in chordates (Fig. 20C, right) are equally parsimonious hypotheses. In both cases, four evolutionary steps are required. Thus, additional data are required to discriminate between the two possibilities.

4.2.2 Acquisition of hormonal binding from a steroid sensing background in the NR1H/I/J group

Steroid-binding nuclear receptors are not restricted to the NR3 subfamily. In the NR1 subfamily (Fig. 21), there is a group of steroid-binding receptors containing the arthropod ecdysone receptor EcR (NR1H1), the vertebrate oxysterol-binding LXR (NR1H2 and NR1H3), the vertebrate bile acid receptor FXR (NR1H4) and the vertebrate bile alcohol receptor FXRβ (NR1H5). The orthologs of LXR and FXR in the urochordate *Ciona intestinalis* were recently shown to bind oxysterols and sulfated steroids ((Reschly et al., 2008b), (Reschly et al., 2008a)). The NR1I/J group also contains the vertebrate PXR (NR1I2) and VDR (NR1I1), which bind vitamin D, bile acids and other cholesterol derivatives, the nematode DAF-12, which binds dafachronic acids (also a kind of steroids) and the insect HR96 (NR1J1), which binds cholesterol (Horner et al., 2009). This group also contains many not yet characterized receptors from amphioxus (Schubert et al., 2008), sea urchin (Howard-Ashby et al., 2006) and various nematodes (Abad et al., 2008), each with lineage-specific duplications. Thus, because most of the characterized receptors from this group, either in chordates or in ecdysozoans, are able to bind cholesterol or steroid derivatives, it is highly likely that the common ancestor of this group was also able to do so.

Functional data allow us to be more precise at least for the NR1I/J group. Members of this group regulate the xenobiotic response in vertebrates, *Drosophila* (HR96; (King-Jones et al., 2006)) and nematodes (NHR-8; (Lindblom et al., 2001)), thus indicating that the common ancestor of bilaterian NR1I/J may have had this ancestral function. If this hypothesis is true, it would be necessary to explain how the nematode DAF-12 and vertebrate VDR shifted from xenobiotic sensing, which implies the binding of many different ligands at the micromolar range, to hormone binding, that implies a binding of one specific molecule with high affinity. For nematode DAF-12, data are currently insufficient to draw an evolutionary scenario. But concerning VDR, it is possible to compare its properties to that of its paralogs PXR and CAR and also to compare the properties of VDR in various vertebrates. The three receptors share binding to some targets implicated in xenobiotic responses, such as the CYP3A genes. CYP3A are involved in hydroxylation of various xenobiotics and endobiotics, such as lithocholic acid, a cytotoxic secondary bile acid that is produced by mammalian intestinal bacteria. Additionally, the sea lamprey VDR is able to activate the transcription of a reporter gene bearing a response element of the mammalian CYP3A4 xenobiotic-metabolising enzyme, but not a that of rat osteocalcin, which is a mammalian VDR target gene involved in bone physiology (Whitfield et al., 2003). This is consistent with the lack in lamprey of a calcified skeleton and plasma levels of calcitriol that are 7 to 8 times higher than those of other vertebrates, which correlates with the lower affinity of lamprey VDR for this ligand. Thus, even if calcitriol is present in lamprey and able to bind VDR, it may not have an hormonal function in that animal. The physiologically relevant lamprey VDR ligand may be a bile acid or another steroid-like molecule. We suggest that results from NR3 family steroid receptors should be interpreted in a similar way. Physiologically relevant estrogen-binding would be a chordate-specific feature. Estrogen binding by annelid NR3D may be viewed as a purely pharmacological property, as observed for the calcitriol-binding VDR in lamprey. Or estrogens may be one of the numerous ligands that can activate annelid NR3D during a xenobiotic response, as it is the case for estrogen binding by the vertebrate PXR or CAR.

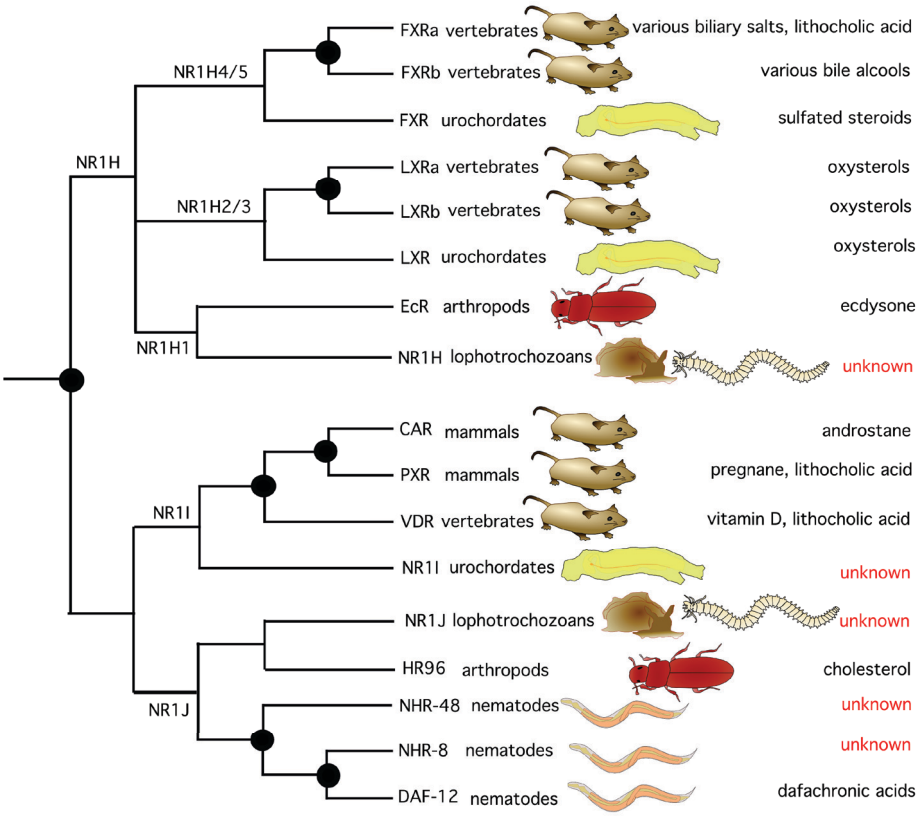


Figure 21: Current data about steroid binding in the NR1H/I/J family. Trivial names of various NR1H/I/J members are plotted on the tree, together with indication on the species where there are present, and with information about the ligand-binding ability of the receptor. Gene duplication events are indicated with black disks. In the NR1H family, all chordate receptors bind steroids, and the NR1H of insects also binds a steroid, so there is high probability that lophotrochozoan NR1H also bind an oxydated cholesterol derivative. Similarly, since there are steroid-binding receptors in vertebrate NR1I and nematode NR1J, whereas the insect ortholog NR1J may bind cholesterol, the currently uncharacterized NR1J from lophotrochozoan would be an obvious candidate for steroid binding.

Chapter 5

Organization of the manuscript

To understand the origin and evolution of NR-mediated steroid signaling, it is necessary to understand when and in which context nuclear receptors became able to bind steroids, but also when metazoan became able to synthesize steroids. As we detailed in part I, the nuclear receptor side was much more studied than the ligand side. In order to balance this, we made some updates on the nuclear receptor side, but focussed mainly our research effort on the ligand side.

- In part II, we update the distribution of nuclear receptors in metazoans, using comparative genomics. We pinpoint the unexpected diversity in this already well known family, naming three new subfamilies. We show that NRs have undergone two major diversification waves: one after the sponge-eumetazoan split, and a second in bilaterians, that correlates with major steps in the diversification of metazoan intercellular communication systems. This article in preparation will be submitted to *Molecular Endocrinology*. In the rest of the manuscript, we follow the below organization:

- In part III, we investigate the relationships between the enzymes that are involved in steroidogenesis using comparative genomics, and we conclude that they were independently recruited in various metazoan groups from a xenobiotic-metabolising background. This article was published in *PNAS* in july 2009.

- In part IV, we explore the relationships between metabolic pathways using tools from comparative anatomy. This has confirmed and completed the previous results, showing that steroidogenetic pathways have evolved with the pattern of cholesterol degradation pathways. This article was just rejected after review in *PLoS Biology* and will be resubmitted after corrections.

- In part V, we discuss all these observations in a more precise zoological context, exploring the functional implications of our results and showing some remaining problems that will need further investigation.

- In part VI, we briefly conclude on the proposed evolutionary model and its implications on our view about endocrinology.

Part II

Nuclear receptor diversification from a metazoan viewpoint

Nuclear receptor diversification from a metazoan viewpoint

Gabriel Markov†, François Bonneton* and Vincent Laudet**

*Molecular Zoology Team; Institut de Génomique Fonctionnelle de Lyon; Université de Lyon; Université Lyon 1; CNRS; INRA; Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France.
†UMR 7221 - Evolution des Régulations Endocriniennes. *Muséum National d'Histoire Naturelle, Paris, France*

Abstract

Nuclear receptors are metazoan-specific transcription factors that are involved in intercellular communication. Here, taking into account genomic data from various recently sequenced metazoan genomes, we reveal the unexpected diversity in this already well known family, naming three new subfamilies. The first new subfamily is the eumetazoan subfamily NR7, that disappeared in vertebrates and ecdysozoans, and may be an important player in lophotrochozoan and cnidarian steroid signaling. The second is the cnidarian-specific NR8 subfamily, that branches at the basis of the bilaterian-specific clade containing NR1 and NR4, and the third is the cnidarian-specific NR9 subfamily, at the basis of the bilaterian-specific clade containing NR5 and NR6. We also elucidate the origin of some of the receptors with no DBD or LBD, and we show that NRs have undergone two major diversification waves: one after the sponge-eumetazoan split, and a second in bilaterians, that correlates with major steps in the diversification of metazoan intercellular communication systems.

Introduction

Nuclear receptors (NRs) are a major component of metazoan intercellular signaling systems, being involved in signal transduction mediated by small hydrophobic molecules such as steroids, fatty acids, thyroid hormones, or retinoic acid (Gronemeyer et al., 2004). Recently, there have been attempts to elucidate what are the general characteristics of NRs in metazoans and how they can explain the metazoan-specific features of their physiology, studying globally their spaciotemporal expression pattern and the set of their target genes (Bookout et al., 2006; Palanker et al., 2006). However, our understanding of metazoan intercellular communication is strongly biased by the fact that largely dominant models in experimental endocrinology are mainly mammals and insects, that are not representative of metazoan diversity. There are three main clades of bilaterian animals, deuterostomes, ecdysozoans and lophotrochozans. Mammals belong to deuterostomes whereas

insects belong to ecdysozoans, but there is no representative of lophotrochozans as a complete endocrinological model up to the genetical level. Moreover, there are other animal groups, such as sponges or cnidarians, where the concept of endocrinology is generally not used but where intercellular signaling occurs. There have been early attempts to grasp a more complete picture on the distribution of nuclear receptors in metazoans (Escriva et al., 1997), but such attemps have been limited because during a decade, only bilaterian genomes were available for metazoans (Bertrand et al., 2004). Now, this lack of data has been partially filled with the publication of four complete genome of non-bilaterian metazoans : the sea anemone *Nematostella vectensis* (Putnam et al., 2009), the placozoan *Trichoplax adhaerens* (Srivastava et al., 2008), the freshwater hydra *Hydra magnipapillata* (Chapman et al., 2010) and the demosponge *Amphimedon queenslandica* (Srivastava et al., 2010). Furthermore, even within bilaterians, the long-lasting lack of genomic data in lophotrochozoans has strongly biased our understanding of the basal NR set of metazoans. The genomes of the blood fluke *Schistosoma mansoni* has been recently published (Berriman et al., 2009) and complete genome data are now available in one mollusk, the limpet *Lottia gigantea*, and two annelids, the polychaete *Capitella teleta* and the leech *Helobdella robusta*. Therefore, it is now possible to have a somehow representative glance at NR diversity in this third major clade of bilaterian animals.

NRs are unique in metazoan physiology in that there are the main family of ligand-activated transcription factors, providing a very direct way to trigger a specific genomic response to environmental changes (Huang et al., 2010). Even if other transcription factors such as the bHLH-PAS family are also able to be activated by the binding of a small lipophilic molecule (Furness et al., 2007), for the moment it is not clear if this specificity is really due to specific structural features of the NRs or if this is the result of an observation bias, due to the important role of NRs in human physiology. Nuclear receptors interact with many different partners, such as the DNA that they bind, the cofactors with which they modulate the activity of the transcription machinery, the ligand that they bind and the proteins that modify their structure through post-translational modifications. Among those interacting partners, the most studied from an evolutionary viewpoint are the ligands. The perception of what a NR ligand is has significantly evolved during the last 20 years (Sladek, 2011), so did the models on the evolution of ligand-binding ability (Markov and Laudet, 2011). Starting from sharply opposing alternatives (orphan receptor versus hormone liganded receptors), a consensus is growing to a vision where the ancestral receptor may have been a sensor of various environmental molecules of exogenous origin, such as dietary lipids or xenobiotics (Eick and Thornton, 2011 ; Markov and Laudet, 2011). Recently, an important point has been clarified about the relationships between the various NRs, showing that almost one receptor, HNF4 (NR2A), is present in all metazoans, and that a second one in sponges is orthologous to the common ancestor of

all the remaining members of the family (Bridgham et al., 2010). Both sponge receptors are able to bind fatty acids, but only one was shown to transactivate gene expression in response to ligand-binding. However, this general study has not gone in depth in the precise NR content of the various metazoans and has not discussed the implications of such a distribution. This is nevertheless important to understand the precise physiological context in which NR signaling through ligand-binding arose. During recent years, a number of genome papers have reported the NR content of various animals. Most of the annotated NRs are orthologous to classically known receptors but in almost each genome there is also a few number of « orphan sequences » whose relationships with other family members are unclear. These separate studies and the use of peculiar nomenclature systems for almost each newly annotated species have created some confusion. This was also amplified with the discovery of especially odd receptors with two DNA-binding domains in some protosomes (Wu et al., 2007). Indeed, the majority of nuclear receptors are made of five to six domains, the most conserved being the DNA-binding domain (DBD) and the ligand-binding domain (LBD). But the family also contains recepors with no DBD in chordates (Zanaria et al., 1994; Seo et al., 1996; Schubert et al., 2008) or no LBD in protostomes (Nauber et al., 1988; Sengupta et al., 1994). The orthology of these odd NRs to canonical members of the family has not been checked since a long time (Laudet, 1997). This increasing accumulation of pending questions makes an extensive reanalysis of all genomic data available urgent. Here, we provide a more extensive update on the NR distribution in metazoans, in order to clarify the increasing nomenclature confusion, and in order to put more precisely this distribution in a zoological context.

Results

Species sampling as a tool to resolve some difficult nodes in the phylogeny.

Using all metazoan sequences available, we completed a global NR tree (Fig. 1*A*), in which the newly identified families are provided with an official nomenclature name, for the sake of clarity. The accession numbers of the newly described sequences and an updated nomenclature are given in Table 1. The general backbone of our phylogeny is consistent with the tree recently published by Bridgham et al. (2010), which was itself consistent with the first analyses concerning the relationships between the six subfamilies (NRNC, 1999), but provided additionally a rooting for the tree and a better statistical support for the nodes between the subfamilies. Here, our purpose is to go more deeper in the analysis of the NR distribution along the metazoan tree to pinpoint some important events regarding the evolution of ligand-binding ability and more generally regarding the contribution of NRs to diversification of intercellular signaling in metazoans. We confirm that HNF4 (NR2A) is the only receptor that is present in all metazoans including sponges, and, additional to the previousely reported sequences, we found in EST-databases two DBD portions of

HNF4 in the demosponge *Ephydatia muelleri* and the hexactinellid *Heterochone calyx*, as well as a LBD-like sequence from the demosponge *Carteriospongia foliascens*. However, the EST sampling available to date is very limited and represents only a tiny fraction of sponge genomic diversity. For example, no sequences of calcareous or homoscleromorph sponge are available, to cite two groups that have been sometimes proposed as sister-groups for the eumetazoans. For the sponge sequence at the basis of all the other families, with the exception of NR2A, we propose the name NR2I. With this system there is a slight inconsistency: the NR2 family is paraphyletic, and this contradicts our own recommendation that official protein names should refer only to their phylogenetic relationships (Markov et al., 2008). However, we consider that it would be very unpractical and prematurate to rename all proteins of the NR2 family for the moment, because such a thing cannot be proposed with the assentiment of the whole NR community, and it would render difficult the following of the literature. So, as it was proposed by zoological taxonomists to deal with the polyphyletic genus *Drosophila* (O'Grady and Markow, 2009), we suggest that, until we have a really complete picture of the NR phylogeny, we keep this framework with a paraphyletic NR2 family. Additionally, the increased species sampling allows to solve some ambiguous nodes using the parsimony principle. For example, the cnidarian sequences here referred as NR8A1 and NR8A2 branch with a weak support at the basis of the eumetazoan NR7 family (Fig. S1). But as the NR1 and NR4 families are bilaterian-specific, this topology implies that NR8 was an eumetazoan family that was lost in bilaterians whereas the common ancestor of the NR1-NR4 family was lost in cnidarians. A more parsimonious explanation is that the cnidarian NR8 are orthologous to the common ancestor of the bilaterian NR1-NR4 family, and that the real topology is perturbed by the lineage-specific expansion in the NR1-NR4 family that makes the sequences very divergent to their cnidarian counterparts. Using the same reasoning, we propose to name NR9 the cnidarian sequence which weakly branches between the bilaterian NR5 and NR6. We also propose to name NR2J the *Trichoplax* sequence at the basis of the eumetazoan NR2E and NR2F. In all these cases, our interpretations are fully consistent with those of Bridgham et al. (2010), and they are also in agreement with a reasoning that was already used to interpret the topology of the NR2A family (Robinson-Rechavi et al., 2005) or of the NR2B family (Bonneton et al., 2003). But there are also two other important nodes that can be analysed with such reasoning, which were not explicitely addressed in the last Bridgham paper. The first is in the NR3 family. As we already pointed out and discussed previousely (Markov et al., 2011), the most parsimonious way to interpret the topology of the subtree grouping chordate NR3A, chordate NR3C and lophotrochozoan « ER », that we propose to call NR3D, is that NR3A and NR3C are chordate-specific duplicates of an ancestral gene that was orthologous to the lophotrochozoan NR3D. This means that, from an evolutionary viewpoint, the lophotrochozoan NR3D in no more closely related to vertebrate ER (NR3A) than to vertebrate

MR/GR/PR/AR (NR3C). Additionally, the *Trichoplax* sequence at the basis of the NR3 subfamily should be called NR3E, to stress that it is not more related to NR3B (ERR) than to other members of the NR3 subfamily. The last important node where phylogenetic sampling can help to solve an ambiguous branching is the NR1H group. This group contains the protostome NR1H1, the vertebrate NR1H2 and NR1H3 (LXR) and their orthologs in other deuterostomes (NR1H8), as well as the vertebrate NR1H4 and NR1H5 (FXR) and their deuterostome counterparts (NR1H7 and maybe NR1H6). Here again, we hypothesize that the basal position of NR1H4/5 and other deuterostome sequences reflects the high divergence in some sequences that have undergone lineage-specific extension, with eight paralogs of NR1H4/5 in amphioxus (NR1H7) and three paralogs in sea urchin (NR1H6). The position of the sea urchin sequences (NR1H6) are more ambiguous, because their grouping with other NR1H is not well supported (57% bootstrap). Since there is no sea urchin sequence in the NR1I/J family, it could also be that the so-called NR1H6 in sea urchin are very divergent orthologs of the chordate NR1I.
To sum up, the extended species sampling makes possible to put many sequences in a nomenclature framework that reflects their relationships (Table S1) and to propose a detailed scheme about the NR distribution in the main groups of metazoan genomic models (Fig. S2).

Unexpected diversity of the NR family and spectacular parallel losses in model bilaterians. The importance of taking into account sequences from lophotrochozoans and cnidarians is illustrated here by the identification of a new eumetazoan subfamily, that we call NR7, with secondary losses in the major model animals that are ecdyzosoans and vertebrates (Fig. 1*B*), and two cnidarian-specific subfamilies (NR8 and NR9). Due to its position as the sister group of a clade containing the bilaterian NR4 and NR1 as well as the cnidarian NR8, the functional characterisation of proteins from the NR7 subfamily will be of primordial importance to understand the deep history of the NR family. Indeed, it is located at an intermediate node between the two groups containing the major hormone receptors that are NR3 and NR1. So, the characterisation of the ligand-binding abilities of NR7 would be important to discriminate if there existed a clade of receptors primarily involved in high affinity ligand binding, or if such an ability has evolved independently in some members of the NR1 and NR3 families. Similarly, within the NR2E clade, we confirm that the proteins that were unofficially named NR2E6 in the genome papers from bee (Velarde et al., 2006), red flour beetle (Bonneton et al., 2008) and sea urchin (Howard-Ashby et al., 2006) are orthologous members of an eumetazoan clade, that underwent a duplication event in cnidarians, and that was lost in vertebrates, in some nematodes and in the *Drosophila* genus (Fig. 1*C*). This protein is still clearly present in another dipteran, *Anopheles gambiae*, and the LBD is disappearing in *Aedes aegypti*, for which a LBD is non significantly predicted in Pfam with an e-value of 0.19. This

represents a kind of direct evidence for the supposed mechanism of formation for NR0, that are proteins lacking either a DBD or a LBD, as we discuss further in more detail. Also noteworthy is the presence of many unclearly placed sequences, at the basis of the NR2E and NR2F clade in *Trichoplax* and *Nematostella*, and in the clade grouping NR1C, D and E, where maybe up to five new bilaterian families exist, due to the presence of sea urchin and lophotrochozan sequences. These sequences are mainly predicted, but on EST basis, and some cDNAs that where cloned in various mollusks are also parts of this new clade, indicating that there are probably not a mere prediction artifact. Since many of the « 2DBD NRs » that were identified in protostomes (Wu et al., 2007) are in this family, it seems that careful further studies will be needed to fully elucidate the evolutionary events that took place at that node. We also confirm that there are many cnidarian-specific receptors in the NR2F family, as it was already supposed (Gauchat et al., 2004), that may indicate bilaterian-specific gene losses.

Diversification by domain losses

It was hypothesized for long that some odd nuclear receptors (with a DBD or a LBD only) have elvolved through domain losses (Laudet, 1997). Here, our abundant sampling enables to follow some of these processes quite precisely. The spectacular case of NR2E5 (Fig. 2A) shows that in different animals, a same receptor can even loss either its LBD or its DBD. In chordates, the DBD completely disappears. In sea urchin, the LBD is very well recognised (3.1e-30 prediction score for domain recognition in Pfam), even with a better score than the DBD (4.9e-25). On the contrary, this score becames twice lower (on a log scale) in the case of *Capitella* (2.8e-14), with a shortening of the terminal part of the LBD. It further decreases at the basis of ecdysozoans (2.3e-06 in *Apis*, 8.9e-08 in *Trichinella*), and the LBD becomes totally absent twice independently in *Drosophila* and in *Caenorhabditis*. Additionally, the *Caenorhabditis* mentioned here is not FAX-1, which is classically considered as a NR2E5 (Bertrand et al., 2004) but is another paralog called NHR-239. According to our tree and consistently with previous propositions (DeMeo, 2009), we argue that FAX-1 is a member of the NR2E3 subclass (Fig. 1). Since there is only one NR2E3 in other nematodes, such as *Trichinella spiralis*, *Loa loa*, *Brugia malayi* and *Ascaris suum*, FAX-1 seems to be the result of a duplication of NR2E3 in the *Caenorhabditis* lineage, leading to NHR-111, that kept a LBD, and FAX-1, that lost its one. The previous confusion is probably due to the blurring of phylogenetic signal as a result of the loss of a recognisable LBD in *Caenorhabditis* and in *Meloidogyne incognita*, a tylenchine nematod. A third classical nematode nuclear receptor, NHR-67, that belongs to the NR2E2 family, has lost its LBD very early in nematode evolution, because it is recognisable only in *Trichinella spiralis*, and absent in *Meloidogyne incognita, Brugia malayi* and *Caenorhabditis elegans* (Fig. S2). LBD losses have occurred also in the NR2E2 family, where in

nematodes, the only NHR-67 with an identifiable LBD belongs to the basal *Trichinella spiralis*. LBD loss also occurred in nematode NHR-85/NR1D, and in nematode NHR-48/NR1J. These abundant losses in crown nematodes are maybe facilitated by the lineage-specific expansion of HNF4, which seem to have occurred after the divergence between *Trichinella* and the other nematodes. But there is also a major counter-example (Fig. 2B). The LBD of the *Caenorhabditis elegans* DAF12 (NR1J) is not recognised by Pfam whereas there is experimental evidence showing that it binds dafachronic acid (Motola et al., 2006). This activity seems to be at least partially conserved in various nematodes (Ogawa et al., 2009; Wang et al., 2009), and the study of the 3D structure of *Strongyloides stercoralis* (Wang et al., 2009) shows that the canonical LBD structure is conserved at least in this species, in spite of some small variations, such as the appearance of very small L3' and L7' helixes. This strikingly shows that the tertiary structure of a protein can be more conserved that the primary sequence, and that not only prediction scores, but also the total lenght of the protein and additional bibliographical information should be taken into account to draw an accurate picture of domain evolution events. Moreother, DAF-12 has two paralogs in the NR1J class: NHR-8, which has a canonical NR structure, and NHR-48, for which no LBD is recognised by Pfam. The NR1J class also contains the insect receptors NR01, NR02 and NR03 that have also lost their LBD. A single unambiguous ortholog of these receptors exists in other arthropods, such as the myriapod *Strigamia maritima*, the chelicerate *Ixodes scapularis* and the crustacean *Caligus rogercresseyi.* It also exists in two lophotrochozans, *Capitella teleta* and *Schistosoma mansoni*. But it is not yet possible to conclude that all these receptors are the product of a unique domain loss in a protostome ancestor, because in the same unresolved group there are also a lot of receptors with one DBD and one LBD, and even a receptor with one DBD and two LBDs in *Daphnia pulex*. The situation is further complicated by the presence of two duplicates of the canonical « NR1J » in *Capitella teleta* and in *Ixodes scapularis* (Fig. S1). The apparent domain loss in DAF-12 illustrates the difficulties to resolve this node, where the choice of the sequences to use for an alignement becames highly complex.

All these discussions on domain losses and duplications illustrate how difficult can be the accurrate phylogenetic reconstruction in a family where the protein structure is very constant on average and where the sampling and functional data are of exceptionally good quality. Even in such a favourable case, careful manual curation is still needed to draw a precise picture of the family history.

Discussion

NR diversification correlates with the diversification of intercellular communication systems

The diversification of the NR family strikingly correlates with the diversification of animal intercellular communication systems (Fig. 3). Despite of the uncertainities about the root of the NR

tree, it seems very likely that the canonic six subfamilies, and the newly defined NR7 subfamily have arisen after the sponge/eumetazoan split, and that the NR set present in sponges was reduced to only two nuclear receptors. Even if the exact timing of early NR diversification remains to be adressed in a more robust manner with extended sampling among sponges, the presence of NRs in sponges, and the fact that the basalmost NRs are sensors that bind various fatty acids should be discussed on a functional and physiological perspective, in order to increase our comprehension of the context in which the ancestral NR arose. Intercellular communication and a certain degree of coordination in cell development at the colony level exists even in bacteria (Rosenberg, 2009). Data on the volvocales, the clade which contains the unicellular flagellate algae *Chlamydomonas reinhardtii*, the multicellular *Volvox carterii* and many organisms showing intermediate character distribution suggests that transition between a colony of unicellular cells and a meta-organism is mainly a question of soma-germen division of labour (Kirk, 2005). In a colony of unicellular cells, each cell can pass through the germ-cell phase, whereas in a meta-organism this role is limited to a subset of cells. This is consistent with genomic studies showing that many protein modules that were traditionally thought to be metazoan-specific are already present in the last common ancestor of metazoans and choanoflagellates (King et al., 2008), or even before. Re-evaluating traditional models about metazoan origin in that context has led to the proposal that metazoans arose through the integration of transient cell-types that were previously temporally successive into a single spatial organism (Mikhailov et al., 2009). NRs are known to be important developmental timers, that is proteins controlling the temporal identity of developmental steps (Thummel, 2001). Moreover, the unique functionally characterised sponge NR, that of *Suberites domuncula*, is implicated in inducing gemmulation in response to retinoid acid (Wiens et al., 2003). Even if retinoic acid is not necessarily the unique physiologically relevant ligand, it can here be viewed as a proxy for xenobiotic or nutritional signal, and it seems reasonable to hypothesize that the ancestral NRs was an important actor in that transition, maybe due to its ability to target a rapid switch in the cellular gene expression integrating information about the cell chemical environment due to his ligand-binding pocket.

The presence of orthologs of some classical NRs even in *Trichoplax adhaerens*, with secondary losses of at least a NR7/4/1/8 and a NR5/6/9 supports the hypothesis that this aenigmatic animal is secondarily simplified, being composed of only four cell types and lacking a nervous system. Strikingly, while *Trichoplax* has orthologs of NR2A and NR2B, and has secondarily lost the homolog of NR2C/D/H, it has a unique NR2J at the basis of the NR2E and NRF families, that have greatly diversified in cnidarians and bilaterians, with probably about 6 to 8 receptors at the split between both lineages (Fig. 3). In today species, many of these receptors, such as NR2E1/TLL, NR2E3/PNR, NR2E6 and NR2F/COUP-TF are known to be implicated, among other functions, in

the developpement of nervous system in chordates (Langlois et al., 2000), ecdysozoans (Velarde et al., 2006) and cnidarians (Gauchat et al., 2004). So their diversification specifically in animals with a nervous system may indicate that their contributed from the beginning in the edification of such a system.

The increased sampling on cnidarian sequences indicates that almost all members of the NR2 family have orthologs in cnidarians, whereas subfamilies NR1/NR4, NR3 and NR5/6 may have diversified after the cnidarian/bilaterian split (Fig. 3). Within the bilaterian-specific NR1 and NR3 families are many notorious « hormone receptors », whereas the NR5 are implicated in regulation of steroidogenesis in both mammals and insects. Additionally, the vertebrate NR2E5-NR0B have a very peculiar structure, with no DBD, that makes us able to play the role of dominant-negative regulators of steroidogenesis, that are physiologically crucial as indicate the human diseases linked to mutations in those receptors. This function is probably quite different from the function of the ancestral NR2E5. In insects, this dominant-negative role is played by a specific isoform of NR1D (E75B) that lacks half of its DBD and is thus not able to bind DNA (Thummel, 1997). This again argues in favour of convergent acquisition of the systems involved in steroid synthesis regulation in arthropods and vertebrates.

As previously stressed (Tarrant, 2005), cnidarians lack an internal circulatory system that would be necessary to have a vertebrate or insect-like endocrine system. Of course this does not mean that there are no long-distance intercellular communications through other means than nervous and neurendocrine system. For example, other molecular signals may be transported around the body through the gut, especially in colonial species. But this, correlated with the previously reported lineage-specific diversification of CYP450 enzymes, that may be implicated in the synthesis of NR ligands (Markov et al., 2009), is consistent with the idea that if really existing, the non-nervous long-distance communication system of cnidarians would be probably not homologous to the vertebrate one, contradictory with many claims based on indirect non-genomic evidences (Twan et al., 2003). What may have existed in a common cnidarian/bilaterian ancestor is a kind of molecular dialog between neuronal and germinal cell, inherited from the common germen-soma communication network in which the ancestral NR may have been implicated. Starting from this common toolkit, some players may have further developped more specific interactions independently in cnidarians and bilaterians, in response to similar functional constraints.

Unexpected genomic diversity of lophotrochozoan NRs favours independent acquisition of endocrine signaling pathways.

The presence of the NR7 subfamily in lophotrochozoans, as well as many new receptors in the NR1 family is of major relevance concerning the discussion on lophotrochozoan endocrinology. In spite

of repeated claims for vertebrate-type steroid signalisation in molluscs and annelids, there is an increasing amount of data favouring independant diversification of such signaling system. Many orthologs of key enzymes, such as aromatase or side-chain cleavage enzyme, seem to be vertebrate or chordate-specific (Markov et al., 2009). Orthologs of steroid receptors are constitutive activators in all molluscs studied to date, and even if some annelid NR3D were shown to bind vertebrosteroids in heterologous system (Keay et al., 2009), it should be noted that they behave very differently than vertebrate receptors in presence of various specific agonists or antagonists. It should also be noted that claims for the presence of vertebrosteroids occur both in animals without NR3s, such as ecdyzosoans, echinoderms or cnidarians, and in animals with NR3s being either constituive activators or « hypothetical receptors », which favours the hypothesis of an NR3-independent steroid signaling pathway. Indeed, another important family involved in vertebrate and ecdysozoan steroid signaling is the NR1H/I/J clade (Fig. 1A), containing the vertebrate oxysterol, vitamin D and bile acid receptors, as well as arthropods ecdysone receptor and nematode dafachronic acid receptor. Given these similarities in ligand specificity, that are reinforced by regulatory similarities, such as the involvement in NR5 in these signaling pathways, such receptors should be first-rate candidates when adressing studies about steroid signaling in lophotrochozans. Additionally, given the possiblity that NR1F may bind to cholesterol in vertebrates, members of the NR1C/D/E/F clade should also be checked. The evolutionary history of this NR clade is far from being solved, and is further complicated by some domain losses or recombination events. But what is already clear is that attempts to understand lophotrochozoan endocrinology should pay attention to them. Accordingly with our previous suggestion that the last common ancestor of all steroid receptors may have been a sensor, which binding ability was not necessarily restricted to steroids, but may have encompassed also various fatty acids and terpenoid derivatives (Markov and Laudet, 2011), we suggest that the acquisition of a highly specific steroidal ligand-receptor couple in lophotrochozoans may have occurred in any of the descendants of this ancestral sensor, and that experimental effort should aim at equally characterising all these receptors if we really want to have a clear view on this question.

Such a discussion could also be valid to some extent for some non-vertebrate deuterostomes, for which data are still scarce. The NR7s from sea urchin and amphioxus, as well as the many specific receptors in the NR1 clade may greatly help to shed light on the endocrine physiology of these understudied animals.

Material and methods

NR sequences were retrieved from various databases (Supplementary Table 1) using a blasting

query set from all six subfamilies. Protein sequences were aligned with Muscle (Edgard, 2004), and alignments were checked by eye and edited with Seaview (Galtier and Gouy, 1996). Phylogenetic trees were made using PHYML (Guindon et al.), a fast and accurate maximum likelihood heuristic method, under the JTT substitution model (Jones et al.), with 100 bootstrap replicates. Predictions for DBDs and LBDs were made with the online domain recognition software Pfam, version 24.0 (Sonnhammer et al., 1997). The threshold for significant domain recognition is an e-value under 1e-5.

Acknowledgements

We thank Gérard Benoit, Mathilde Paris and Marie Sémon for fruitful discussions. This work was supported by FRM, MENRT, the Cascade Network of Excellence (FOOD-CT-2003-506319), ENS Lyon and CNRS.

References

Berriman M, *et al.* (2009) The genome of the blood fluke *Schistosoma mansoni*. *Nature* 460:352-358.

Bertrand S, *et al.* (2004) Evolutionary Genomics of Nuclear Receptors: From Twenty-Five Ancestral Genes to Derived Endocrine Systems. *Mol Biol Evol* 21:1923-1937.

Bonneton F, Zelus D, Iwema T, Robinson-Rechavi M, Laudet V (2003) Rapid divergence of the ecdysone receptor in Diptera and Lepidoptera suggests coevolution between ECR and USP-RXR. *Mol Biol Evol* 20:541-553.

Bonneton F, Chaumot A, Laudet V (2008) Annotation of *Tribolium* nuclear receptors reveals an increase in evolutionary rate of a network controlling the ecdysone cascade. *Insect Biochem Mol Biol* 38, 416-429.

Bookout AL, *et al.* (2006) Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. *Cell* 126:789-799.

Bridgham JT, *et al.* (2010) Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol* 8.

Chapman JA, *et al.* (2010) The dynamic genome of *Hydra*. *Nature* 464:592-596.

DeMeo SD, *et al.* (2008) Specificity of DNA-binding by the FAX-1 and NHR-67 nuclear receptors of *Caenorhabditis elegans* is partially mediated via a subclass-specific P-box residue. *BMC Mol Biol.* 9, 2.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC*

Bioinformatics 5:113.

Escriva H, *et al.* (1997) Ligand binding was acquired during evolution of nuclear receptors. *Proc Natl Acad Sci U S A* 94:6803-6808.

Furness SGB, Lees MJ, Whitelaw ML (2007) The dioxin (aryl hydrocarbon) receptor as a model for adaptive responses of bHLH/PAS transcription factors. *FEBS Lett* 581:3616-3625.

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12:543-548.

Gauchat D, *et al.* (2004) The orphan COUP-TF nuclear receptors are markers for neurogenesis from cnidarians to vertebrates. *Dev Biol* 275, 104-123.

Gronemeyer H, Gustafsson J, Laudet V (2004) Principles for modulation of the nuclear receptor superfamily. *Nat Rev Drug Discov* 3:950-964.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696-704.

Howard-Ashby M, *et al.* (2006) Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. *Dev Biol* 300, 90-107.

Huang P, Chandra V, Rastinejad F (2010) Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. *Annu Rev Physiol* 72, 247-272.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275-282.

Kirk DL (2005) A twelve-step program for evolving multicellularity and a division of labor. *Bioessays* 27, 299-310.

King N, *et al.* (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783-788.

Laudet V, Hanni C, Coll J, Catzeflis F, Stehelin D (1992). Evolution of the nuclear receptor gene superfamily. *EMBO J* 11:1003-13.

Langlois MC, *et al.* (2000) Amphicoup-TF, a nuclear orphan receptor of the lancelet *Branchiostoma floridae*, is implicated in retinoic acid signalling pathways. *Dev Genes Evol* 210, 471-482.

Markov G, Lecointre G, Demeneix B, Laudet V (2008) The "street light syndrome", or how protein taxonomy can bias experimental manipulations. *Bioessays* 30:349-357.

Markov GV, *et al.* (2009) Independent elaboration of steroid hormone signaling pathways in metazoans. *Proc Natl Acad Sci U S A* 106:11913-8.

Markov GV, Laudet V (2011) Origin and evolution of the ligand-binding ability of nuclear receptors. *Mol Cell Endocrinol* 334, 21-30.

Mikhailov KV, *et al.* (2009) The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* 31, 758-768.

Motola DL, *et al.* (2006) Identification of ligands for DAF-12 that govern dauer formation and reproduction in *C. elegans*. *Cell* 124:1209-1223.

Nauber U, et al. (1988) Abdominal segmentation of the *Drosophila* embryo requires a hormone receptor-like protein encoded by the gap gene *knirps*. *Nature* 336:489-492.

Ogawa A, Streit A, Antebi A, Sommer RJ (2009) A conserved endocrine mechanism controls the formation of dauer and infective larvae in nematodes. *Curr Biol* 19:67-71.

O'Grady PM, Markow TA (2009) Phylogenetic taxonomy in *Drosophila*. *Fly (Austin)* 3:0-14.

Palanker L, *et al.* (2006) Dynamic regulation of *Drosophila* nuclear receptor activity in vivo. *Development* 133:3549-3562.

Putnam NH, *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317:86-94.

Robinson-Rechavi M, Maina CV, Gissendanner CR, Laudet V, Sluder A (2005) Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J Mol Evol* 60:577-586.

Rosenberg SM (2009) Life, death, differentiation, and the multicellularity of bacteria. *PLoS Genet* 5, e1000418.

Sengupta P, Colbert HA, Bargmann CI (1994) The *C. elegans* gene *odr-7* encodes an olfactory-specific member of the nuclear receptor superfamily. *Cell* 79:971-980.

Seol W, Choi HS, Moore DD (1996) An orphan nuclear hormone receptor that lacks a DNA binding domain and heterodimerizes with other receptors. *Science* 272:1336-1339.

Sladek FM (2011) What are nuclear receptor ligands? *Mol Cell Endocrinol* 334:3-13.

Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 3:405-20.

Srivastava M, *et al.* (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466:720-726.

Tarrant AM (2005) Endocrine-like Signaling in Cnidarians: Current Understanding and Implications for Ecophysiology. *Integr Comp Biol* 45:201–214.

Thummel CS (2001) Molecular mechanisms of developmental timing in *C. elegans* and *Drosophila*. *Dev Cell* 1:453-465.

Twan WH, Hwang JS, Chang CF (2003) Sex steroids in scleractinian coral, *Euphyllia ancora*: implication in mass spawning. *Biol Reprod* 68:2255-60.

Velarde RA, Robinson GE, Fahrbach SE (2006) Nuclear receptors of the honey bee: annotation and expression in the adult brain. *Insect Mol Biol* 15:583-595.

Wang Z, *et al.* (2009) Identification of the nuclear receptor DAF-12 as a therapeutic target in parasitic nematodes. *Proc Natl Acad Sci U S A* 106:9138-914.

Wiens M, Batel R, Korzhev M, Müller WEG (2003) Retinoid X receptor and retinoic acid response in the marine sponge *Suberites domuncula. J Exp Biol* 206, 3261-3271.

Wu W, Niles EG, Hirai H, LoVerde PT (2007) Evolution of a novel subfamily of nuclear receptors with members that each contain two DNA binding domains. *BMC Evol Biol* 7, 27.

Zanaria E, *et al.* (1994) An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita. *Nature* 372:635-641.

Figure 1. New members in the NR family

A. Simplified version of the complete NR phylogeny. The full maximum-likelihood tree that was used to produce this consensus is shown in Fig. S1. B. Focus on the new NR7 clade. C. Focus on the NR2E6 clade. In B and C lineage-specific losses in current laboratory models are indicated.

Figure 2. Variability of LBD recognition patterns

A. Parallel LBD losses in *Caenorhabditis, Drosophila* and *Tribolium*, and loss of chordate DBD for NR2E5. The e-values refer to Pfam domain prediction scores. The lower the score, the better the prediction. Prediction scores below the 1e-5 threshold are considered unsignificant. Note that not only the prediction score, but also the total length of the protein varies. The red disk indicates a vertebrate-specific duplication of the “NR0B”.

B. Apparent LBD loss in a functionally conserved receptor, the nematode DAF12/NR1J. The conservation of the total length of the protein as well as additional functional data indicate that in spite of an unrecognisable LBD at the primary sequence level, it is still present.

Figure 3. Correlations between major steps of endocrine system evolution and NR diversification

On the left side are plotted the apparition of intercellular communication within a single organism (paracrine signaling), the long-distance intercellular communication through the nervous system and neuropeptides (neuroendocrine signaling), and the circulatory system making able long-distance transport of molecular signals through body fluids (endocrine signaling). On the right side are plotted the NR2E/F diversification, that concerns receptors mainly involved in nervous system development, and the NR1 and NR3 diversification, which concerns the classical hormone receptors.

Table 1. An updated nomenclature of the NR family

Figure S1. A complete metazoan NR phylogenetic tree.

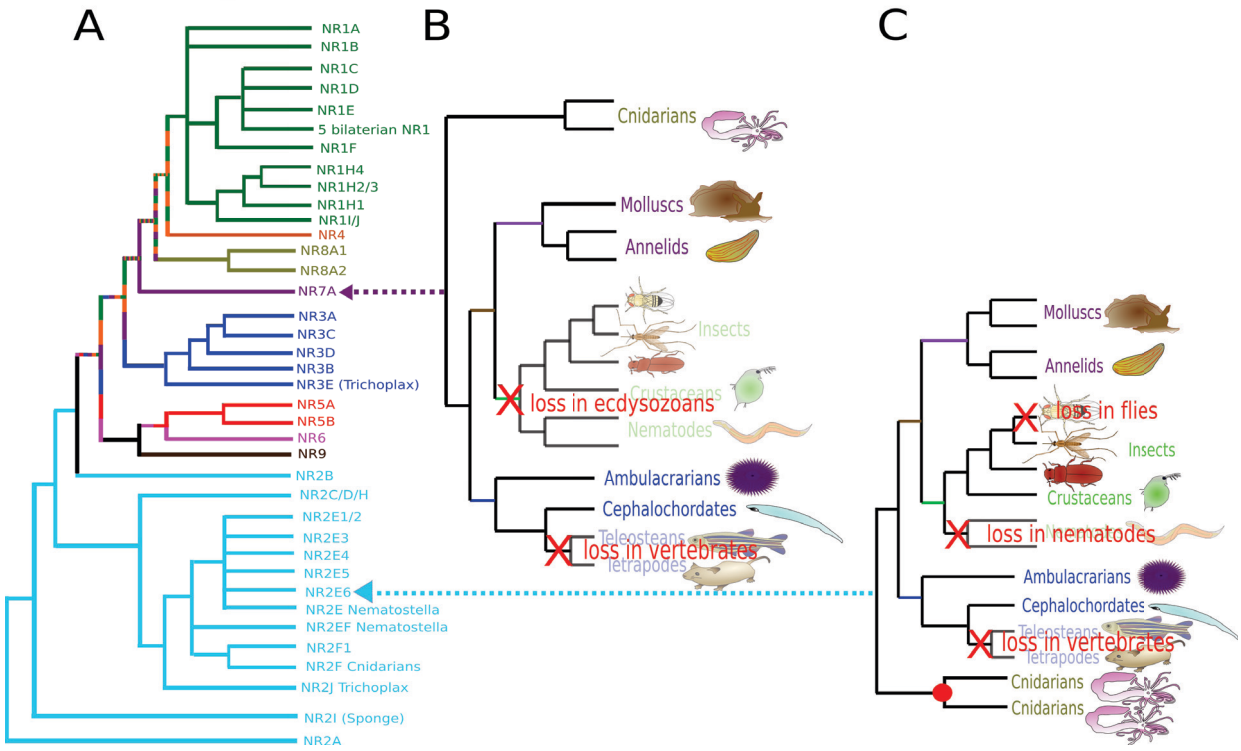
A maximum-likelihood tree of the metazoan nuclear receptors. The families are indicated with color boxes using the same colour code as in the Fig. 1. Accession numbers of the used sequences are provided in Table 1. Bootstrap values (100 repetitions) are indicated for nodes that are important for the discussion.

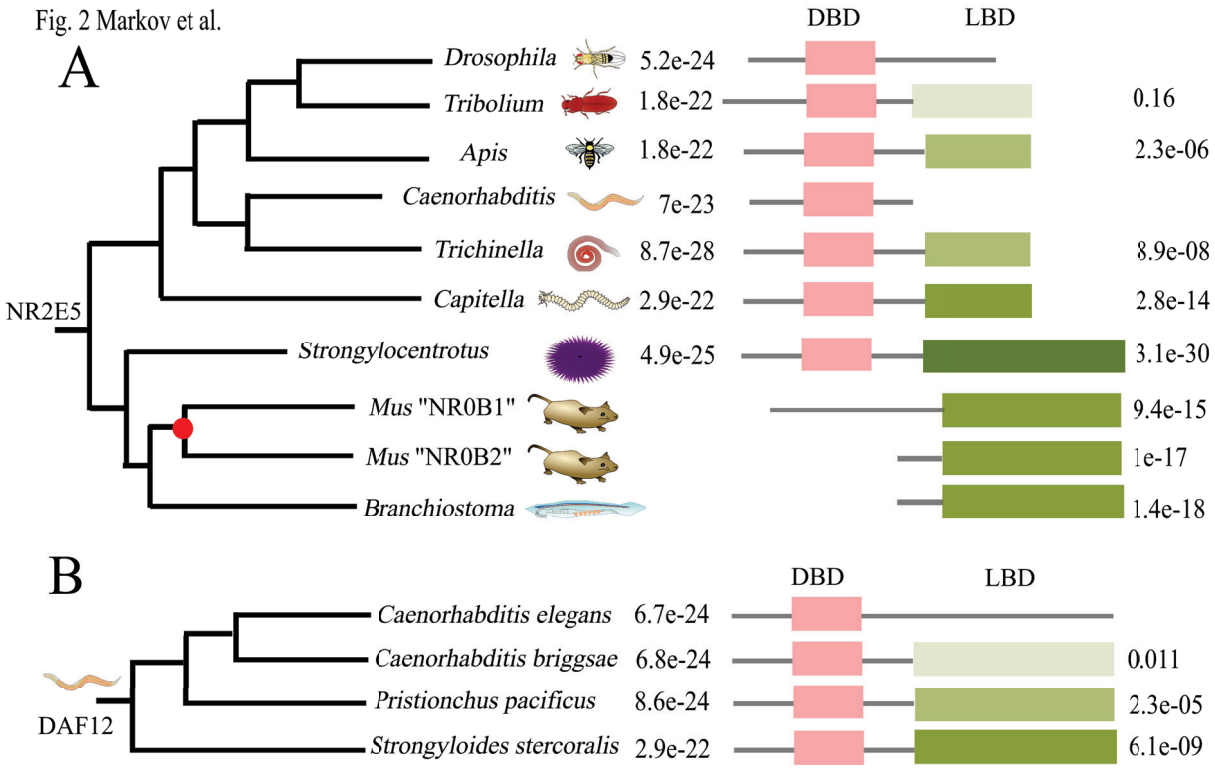
Figure S2. A synthetic update about the NR distribution among animals.

The NR contain of the main metazoan genomic models is plotted against a phylogeny of nuclear receptors. For *Amphimedon, Trichoplax* and cnidarians, where sometimes one receptor is homologous to many bilaterian receptors, the box symbolizing the receptor is elongated to match the whole set of bilaterian receptor to which it is homologous.

Lost receptors are indicated by crosses, ambiguous cases or lack of information is indicated by a question mark, and lineage-specific duplications are indicated by “xN”, where N is the number of paralogs. 2R and 3R refer to the rounds of whole-genome duplication that occurred in vertebrates, and concerned also NRs.

Fig. 1 Markov et al.





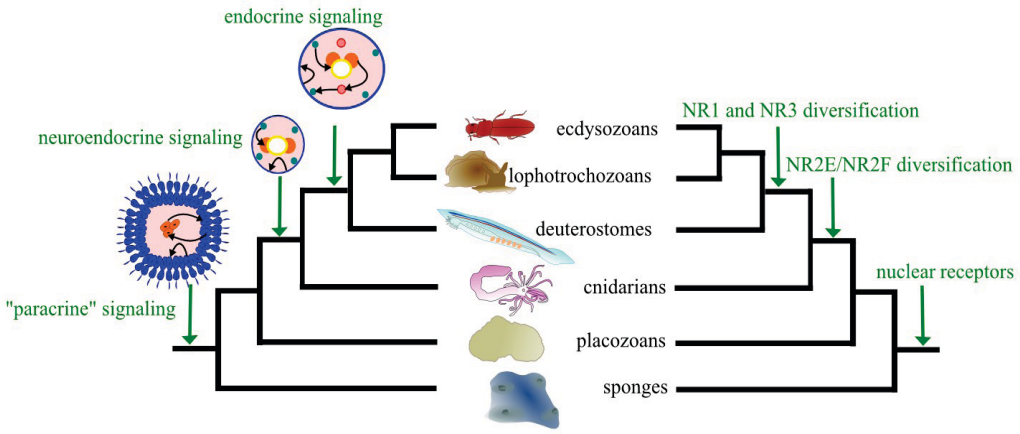
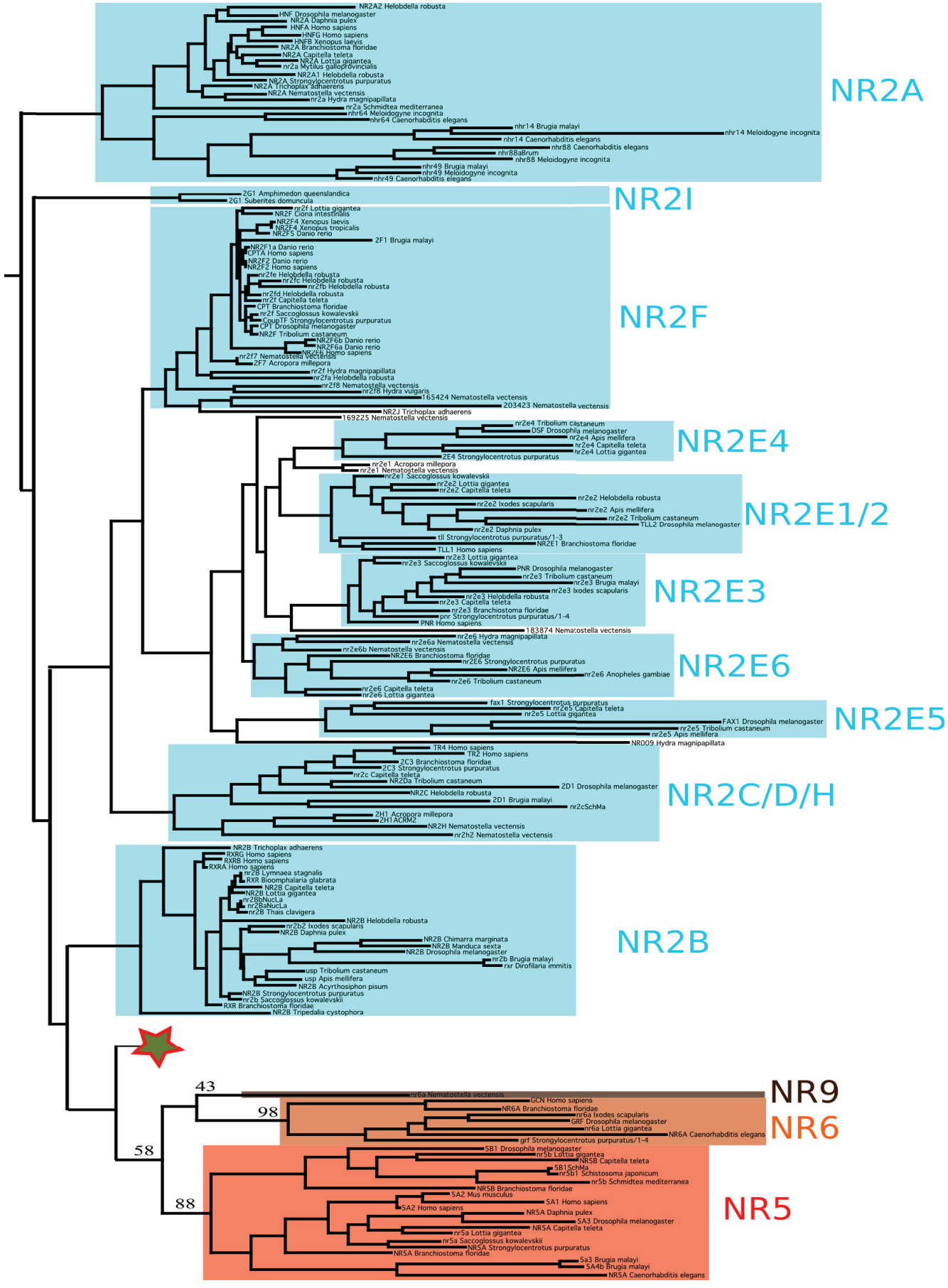


Fig. 3 Markov et al.


NR2A
NR2I
NR2F
NR2E4
NR2E1/2
NR2E3
NR2E6
NR2E5
NR2C/D/H
NR2B
NR9
NR6
NR5
43
98
58
88

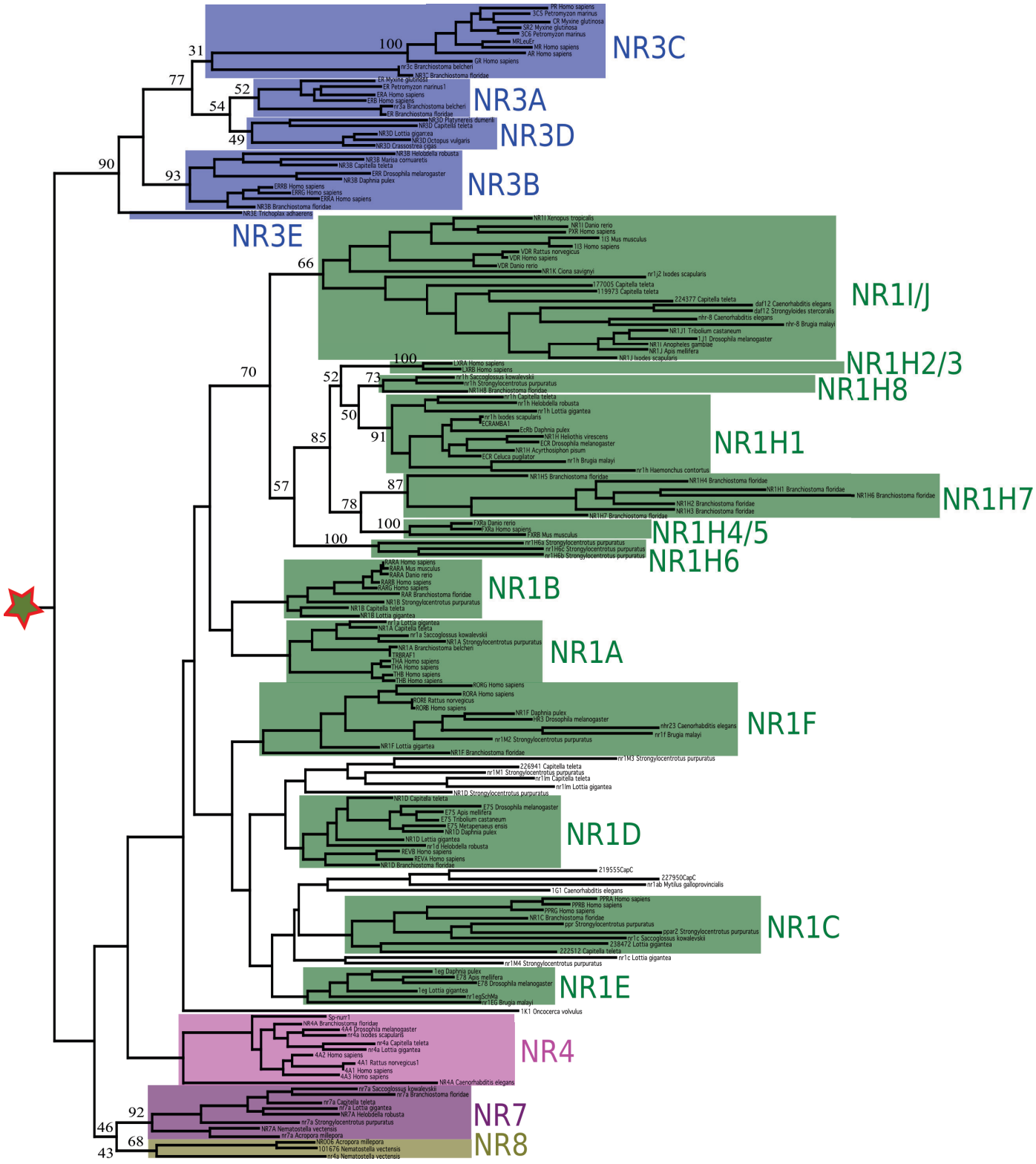


Fig. S1 (in two parts)

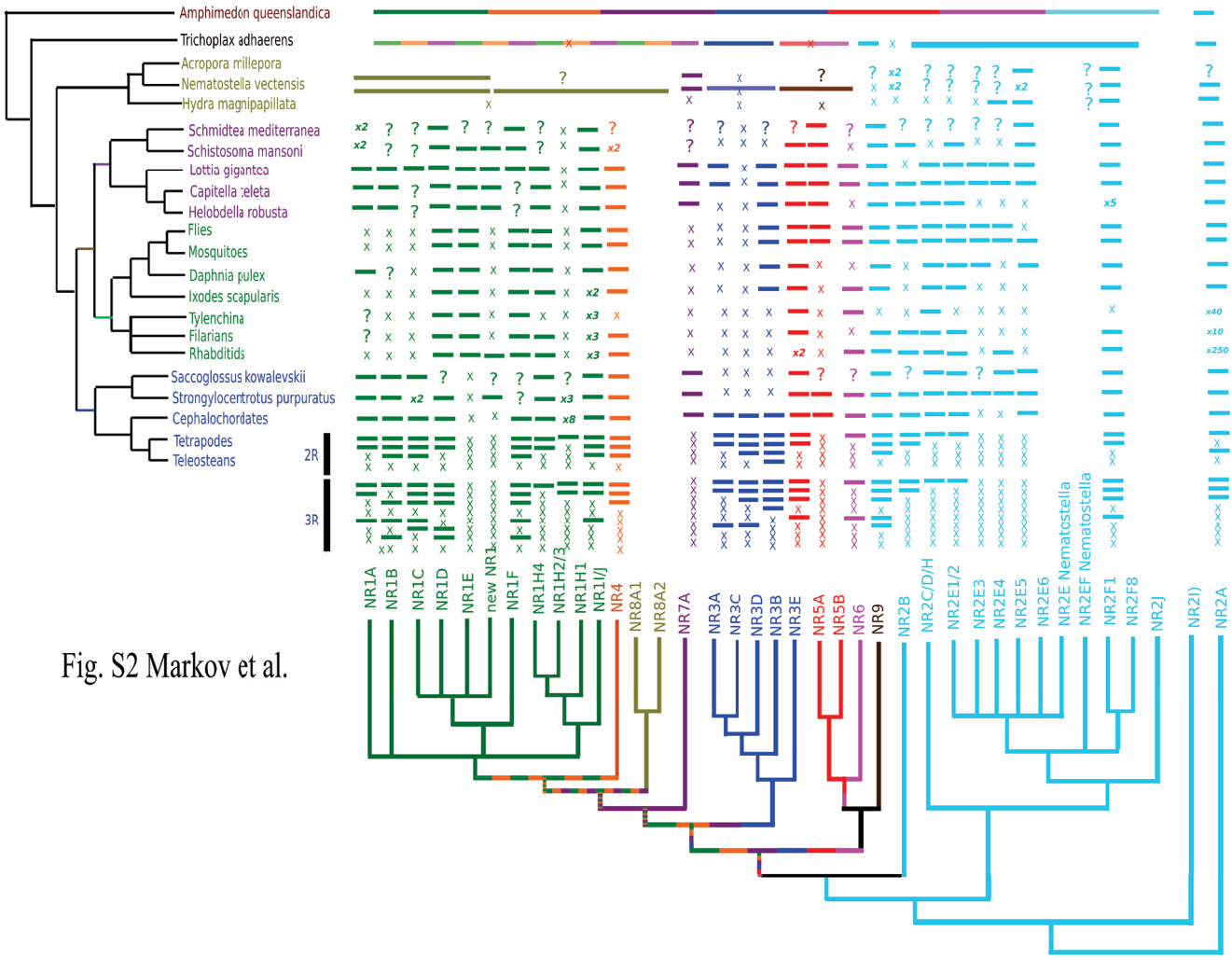


Fig. S2 Markov et al.

Sheet1

Table 1. An updated nomenclature of the NR family

Species	Accession number	Family	Trivial Names
Capitella teleta	167148	NR1A	
Lottia gigantea	171663	NR1A	
Saccoglossus kowalevskii	NP_001161669.1	NR1A	
Strongylocentrotus purpuratus	XP_001185977.1	NR1A	Sp-thr
Branchiostoma belcheri	ACR15148.1	NR1A	
Branchiostoma floridae	ABS11249.1	NR1A	
Homo sapiens	M24748	NR1A1	TRα
Danio rerio	Q98867.1	NR1A	TRαA
Homo sapiens	X04707	NR1A2	TRβ
Danio rerio	NP_571415.1	NR1A2	TRβ
Homo sapiens	X06538	NR1B1	RARα
Mus musculus	P11416	NR1B1	RARα
Danio rerio	Q90271	NR1B	RARαA
Homo sapiens	P10826	NR1B2	RARβ
Homo sapiens	M57707	NR1B3	RARγ
Branchiostoma floridae	XP_002598475.1	NR1B	
Strongylocentrotus purpuratus	XP_779976.2	NR1B	
Capitella teleta	168520	NR1B	
Lottia gigantea	142734	NR1B	
Homo sapiens	L02932	NR1C1	PPARα
Homo sapiens	L07592	NR1C2	PPARβ/δ
Homo sapiens	L40904	NR1C3	PPARγ
Branchiostoma floridae	XP_002598634.1	NR1C	
Strongylocentrotus purpuratus	XP_784429.2	NR1C4	Sp-ppar2
Strongylocentrotus purpuratus	XP_781750.1	NR1C5	Sp-ppar1
Saccoglossus kowalevskii	NP_001164713.1	NR1C	
Lottia gigantea	174409	NR1C	
Lottia gigantea	238472	New NR1 group	
Capitella teleta	222512	New NR1 group	
Strongylocentrotus purpuratus	XP_781794.2	New NR1 group	NR1M4
Capitella teleta	219555	New NR1 group	
Capitella teleta	227950	New NR1 group	
Mytilus galloprovincialis	ABU89802.1	New NR1 group	NR1ABC
Caenorhabditis elegans	NP_001024662.1	NR1G1	CNR14
Homo sapiens	M24898	NR1D1	REVERBα
Homo sapiens	L31785	NR1D2	REVERBβ
Branchiostoma floridae	XP_002598635.1	NR1D	
Capitella teleta	62897	NR1D	
Lottia gigantea	136477	NR1D	
Helobdella robusta	67463	NR1D	
Drosophila melanogaster	X51548	NR1D3	E75
Apis mellifera	NP_001073579	NR1D3	
Tribolium castaneum	XP_971362	NR1D3	
Metapenaeus ensis	AAC71770	NR1D3	
Daphnia pulex	ADB79814.1	NR1D3	
Strongylocentrotus purpuratus	XP_785820.1	NR1D	Sp-reverb
Strongylocentrotus purpuratus	XP_797496.1	New NR1 group	NR1M3
Capitella teleta	226941	New NR1 group	
Strongylocentrotus purpuratus	NP_001123279.1	New NR1 group	NR1M1
Capitella teleta	227484	New NR1 group	

Sheet1

Lottia gigantea	168854	New NR1 group	
Daphnia pulex	50511	NR1E1	
Apis mellifera	XP_396527	NR1E1	
Drosophila melanogaster	U01087	NR1E1	E78
Lottia gigantea	233381	NR1E1	
Schistosoma mansoni	AAR30507.2	NR1E1	
Brugia malayi	EDP32378.1	NR1E1	
Homo sapiens	U04897	NR1F1	RORα
Homo sapiens	Y08639	NR1F2	RORβ
Rattus norvegicus	P45446.3	NR1F2	RORβ
Homo sapiens	U16997	NR1F3	RORγ
Branchiostoma floridae	174225	NR1F	
Strongylocentrotus purpuratus	XP_783869.1	New NR1 group	NR1M2
Lottia gigantea	167096	NR1F	
Daphnia pulex	EFX87520.1	NR1F4	HR3
Drosophila melanogaster	M90806	NR1F4	HR3
Caenorhabditis elegans	U13075	NR1F4	Nhr23
Brugia malayi	EDP30145.1	NR1F4	
Drosophila melanogaster	M74078	NR1H1	ECR
Heliothis virescens	O18473.1	NR1H1	
Acyrthosiphon pisum	NP_001152831.1	NR1H1	
Celuca pugilator	AAC33432.2	NR1H1	
Daphnia pulex	EFX68327.1	NR1H1	
Ixodes scapularis	XP_002405625.1	NR1H1	
Amblyomma americanum	AAB94565.1	NR1H1	EcRA3
Brugia malayi	ABQ28713.1	NR1H1	
Haemonchus contortus	ADD49663.1	NR1H1	
Helobdella robusta	108893	NR1H1	
Capitella teleta	125155	NR1H1	
Lottia gigantea	170342	NR1H1	
Homo sapiens	U07132	NR1H2	LXRβ
Homo sapiens	U22662	NR1H3	LXRα
Homo sapiens	Q96RI1	NR1H4	FXRα
Danio rerio	NP_001002574	NR1H4	FXRα
Mus musculus	NP_941060.2	NR1H5	FXRβ
Strongylocentrotus purpuratus	XP_786370.2	NR1H6a	
Strongylocentrotus purpuratus	XP_782721.1	NR1H6b	
Strongylocentrotus purpuratus	XP_783089.2	NR1H6c	
Branchiostoma floridae	XP_002595803.1	NR1H7	NR1H-5
Branchiostoma floridae	XP_002603811.1	NR1H7	NR1H-4
Branchiostoma floridae	XP_002603810.1	NR1H7	NR1H-1
Branchiostoma floridae	XP_002588118.1	NR1H7	NR1H-6
Branchiostoma floridae	124680	NR1H7	NR1H-2
Branchiostoma floridae	124679	NR1H7	NR1H-3
Branchiostoma floridae	156544	NR1H7	NR1H-7
Branchiostoma floridae	222287	NR1H8	NR1H-8
Strongylocentrotus purpuratus	NP_001123279.1	NR1H8	Sp-fxr
Homo sapiens	J03258	NR1I1	VDR
Rattus norvegicus	P13053.1	NR1I1	VDR
Danio rerio	Q9PTN2.2	NR1I	VDRα
Homo sapiens	O75469	NR1I	
Danio rerio	NP_001092087	NR1I	
Xenopus tropicalis	ENSXETT00000039111	NR1I2	PXR

Sheet1

Homo sapiens	Z30425	NR1I3	CAR1
Mus musculus	NP_033933.2	NR1I4	CAR2
Ciona savignyi	ENSCSAVT00000003276	NR1K	
Drosophila melanogaster	NP_524493.1	NR1J1	HR96
Anopheles gambiae	XP_313130	NR1J1	HR96
Tribolium castaneum	XP_968487.1	NR1J1	HR96
Apis mellifera	XP_624213	NR1J1	HR96
Ixodes scapularis	XP_002404556.1	NR1J1	
Caenorhabditis elegans	AAD34462.1	NR1J	DAF12
Strongyloides stercoralis	AAD37372.1	NR1J	DAF12
Caenorhabditis elegans	AAD03684.1	NR1J	nhr-8
Brugia malayi	XP_001896592.1	NR1J	nhr-8
Capitella teleta	224377	NR1I/J/K	
Capitella teleta	177005	NR1I/J/K	
Capitella teleta	119973	NR1I/J/K	
Ixodes scapularis	XP_002402961.1	NR1J2	
Onchocerca volvulus	AAA87173.1	NR1K	
Helobdella robusta	67417	NR2A2	
Drosophila melanogaster	U70874	NR2A4	
Daphnia pulex	59378	NR2A	
Homo sapiens	X76930	NR2A1	HNF4α
Homo sapiens	Z49826	NR2A2	HNF4γ
Xenopus tropicalis	ENSXETT00000035778	NR2A3	HNF4β
Branchiostoma floridae	XP_002612502.1	NR2A	
Capitella teleta	172322	NR2A	
Lottia gigantea	108689	NR2A	
Mytilus galloprovincialis	CAJ53825.2	NR2A	
Helobdella robusta	68918	NR2A1	
Strongylocentrotus purpuratus	XP_780389.1	NR2A	
Trichoplax adhaerens	50786	NR2A	
Nematostella vectensis	89471	NR2A	NvNR5
Hydra magnipapillata	XP_002159483.1	NR2A	
Schmidtea mediterranea	Contig2310.1	NR2A	
Meloidogyne incognita	Minc02318	NR2A	nhr-64
Caenorhabditis elegans	O44960.2	NR2A	nhr-64
Brugia malayi	EDP36393.1	NR2A	nhr-14
Meloidogyne incognita	Minc11307	NR2A	nhr-14
Caenorhabditis elegans	O02151.3	NR2A	nhr-14
Caenorhabditis elegans	AAG15179.1	NR2A	nhr-88
Brugia malayi	EDP36547.1	NR2A	nhr-88
Meloidogyne incognita	Minc1542	NR2A	nhr-88
Brugia malayi	EDP35600.1	NR2A	nhr-49
Meloidogyne incognita	Minc02316	NR2A	nhr-49
Caenorhabditis elegans	O45666.2	NR2A	nhr-49
Trichoplax adhaerens	49897	NR2B	
Tripedalia cystophora	AAC80008.1	NR2B	
Homo sapiens	X52773	NR2B1	RXRα
Homo sapiens	M84820	NR2B2	RXRβ
Homo sapiens	X66225	NR2B3	RXRγ
Lymnaea stagnalis	AAW34268.1	NR2B4	
Biomphalaria glabrata	Q8T5C6.1	NR2B4	
Capitella teleta	164614	NR2B4	
Lottia gigantea	162352	NR2B4	

Sheet1

Nucella lapilius	ABS70715.1	NR2B	NR2Ba
Nucella lapilius	ABS70716.1	NR2B	NR2Bb
Thais clavigera	AAU12572.1	NR2B4	
Helobdella robusta	62045	NR2B4	
Ixodes scapularis	XP_002435070.1	NR2B4	
Daphnia pulex	219609	NR2B4	
Chimarra marginata	AAZ38141.1	NR2B4	USP
Manduca sexta	P54779.1	NR2B4	USP
Drosophila melanogaster	X52591	NR2B4	USP
Tribolium castaneum	CAL25729.1	NR2B4	USP
Apis mellifera	AAF73057.1	NR2B4	USP
Acyrtosiphon pisum	NP_001155140.1	NR2B4	USP
Brugia malayi	ABQ28715.1	NR2B4	
Dirofilaria immitis	AAM08269.1	NR2B4	
Strongylocentrotus purpuratus	XP_001201896.1	NR2B	
Saccoglossus kowalevskii	ADB22634.1	NR2B	
Branchiostoma floridae	AAM46151.1	NR2B	
Homo sapiens	M29960	NR2C1	TR2
Homo sapiens	L27586	NR2C2	TR4
Branchiostoma floridae	129718	NR2C3	
Strongylocentrotus purpuratus	NP_001116968.1	NR2C3	
Capitella teleta	224222	NR2C	
Helobdella robusta	92366	NR2C	
Schistosoma mansoni	XP_002581285.1	NR2C	
Drosophila melanogaster	U36791	NR2D1	HR78
Tribolium castaneum	EFA11562.1	NR2D1	HR78
Nematostella vectensis	167880	NR2C/D/H	NvNR15
Nematostella vectensis	209681	NR2H	NvNR17
Acropora millepora	AAL29201.1	NR2H1b	NR8Am, 2H1ACRM2
Acropora millepora	AAL29197.1	NR2H1a	NR4Am, 2H1ACRM1
Homo sapiens	S72373	NR2E1	TLX
Branchiostoma floridae	155937	NR2E1	
Strongylocentrotus purpuratus	XP_794533.2	NR2E1	Sp-tll
Saccoglossus kowalevskii	NP_001158362.1	NR2E1	
Lottia gigantea	130367	NR2E2	
Capitella teleta	226190	NR2E2	
Helobdella robusta	71774	NR2E2	
Ixodes scapularis	XP_002403220.1	NR2E2	
Daphnia pulex	299738	NR2E2	
Apis mellifera	XP_001121187.2	NR2E2	
Tribolium castaneum	EEZ99198.1	NR2E2	
Drosophila melanogaster	M34639	NR2E2	TLL
Homo sapiens	AAD28301.1	NR2E3	PNR
Branchiostoma floridae	225454	NR2E3	
Strongylocentrotus purpuratus	XP_780706.2	NR2E3	Sp-pnr
Saccoglossus kowalevskii	NP_001158447.1	NR2E3	
Lottia gigantea	137595	NR2E3	
Capitella teleta	171557	NR2E3	
Helobdella robusta	139475	NR2E3	
Brugia malayi	EDP32855.1	NR2E3	
Ixodes scapularis	XP_002409993.1	NR2E3	
Tribolium castaneum	EFA07486.1	NR2E3	HR51
Drosophila melanogaster	NP_611032.2	NR2E3	HR51

Sheet1

Nematostella vectensis	183874	NR2E	NvNR6
Nematostella vectensis	114090	NR2E1	NvNR5
Acropora millepora	AAL29193.1	NR2E1	
Tribolium castaneum	EEZ99270.1	NR2E4	
Drosophila melanogaster	AAD05225.1	NR2E4	DSF
Apis mellifera	XP_624265.2*	NR2E4	
Capitella teleta	224811	NR2E4	
Lottia gigantea	135497	NR2E4	
Strongylocentrotus purpuratus	XP_789465.1	NR2E4	
Nematostella vectensis	169225	NR2E	NvNR7
Drosophila melanogaster	AAF54133.1	NR2E5	
Tribolium castaneum	EFA04538.1	NR2E5	
Apis mellifera	XP_001121181.1*	NR2E5	
Capitella teleta	177303	NR2E5	
Lottia gigantea	132866	NR2E5	
Strongylocentrotus purpuratus	XP_795547.1	NR2E5	
Hydra magnipapillata	XP_002156561.1	NR2E	NR009
Hydra magnipapillata	XP_002154441.1	NR2E6	
Nematostella vectensis	132075	NR2E6b	
Nematostella vectensis	99425	NR2E6a	NvNR2
Lottia gigantea	120392	NR2E6	
Capitella teleta	53417	NR2E6	
Tribolium castaneum	EFA04575.1	NR2E6	
Apis mellifera	XP_624042.2	NR2E6	
Anopheles gambiae	AGAP001348-RA.1	NR2E6	
Branchiostoma floridae	236186	NR2E6	
Strongylocentrotus purpuratus	SPU_017375*	NR2E6	
Lottia gigantea	125514	NR2F	
Ciona intestinalis	Q4H3S1	NR2F	
Xenopus laevis	CAA44806.1	NR2F4	COUP-TFIII
Xenopus tropicalis	ENSXETT00000024153	NR2F4	
Danio rerio	Q06726	NR2F5	SVP46
Brugia malayi	EDP32461.1	NR2F	
Danio rerio	Q06725.1	NR2F1A	
Homo sapiens	X12795	NR2F1	COUP-TFI
Homo sapiens	P24468	NR2F2	COUP-TFII
Danio rerio	NP_571258	NR2F2	
Helobdella robusta	193553	NR2Fe	
Helobdella robusta	77679	NR2Fc	
Helobdella robusta	76907	NR2Fb	
Helobdella robusta	106233	NR2Fd	
Capitella teleta	171549	NR2F	
Branchiostoma floridae	AAO61416.1	NR2F	
Saccoglossus kowalevskii	NP_001158369.1	NR2F	
Strongylocentrotus purpuratus	XP_782295.2	NR2F	Sp-coupTF
Drosophila melanogaster	M28863	NR2F3	SVP
Tribolium castaneum	EFA11548.1	NR2F3	
Danio rerio	NP_991120.1	NR2F6a	
Danio rerio	NP_998404.1	NR2F6b	
Homo sapiens	X12794	NR2F6	EAR2
Acropora millepora	AAL29200.1	NR2F7	AmNR7
Nematostella vectensis	189134	NR2F7	NvNR10
Hydra magnipapillata	XP_002159396.1	NR2F	

Sheet1

Helobdella robusta	191405	NR2Fa	
Nematostella vectensis	242271	NR2F8	NvNR11
Hydra vulgaris	AAU11312.1	NR2F8	
Nematostella vectensis	165424	NR2	NvNR12
Nematostella vectensis	203423	NR2	NvNR13
Trichoplax adhaerens	21656	NR2J	
Suberites domuncula	Q8I748	NR2I	
Amphimedon queenslandica	ACA04755.1	NR2I	AqNR1
Homo sapiens	X03635	NR3A1	ERα
Homo sapiens	U57439	NR3A2	ERβ
Petromyzon marinus	AAK20929.1	NR3A	
Myxine glutinosa	ACC85903.1	NR3A	
Branchiostoma floridae	XP_002613220.1	NR3A	
Branchiostoma belcheri	BAI59767.1	NR3A	
Homo sapiens	X51416	NR3B1	ERRα
Homo sapiens	X51417	NR3B2	ERRβ
Homo sapiens	NP_001429.2	NR3B3	ERRγ
Branchiostoma floridae	AAU88062.1	NR3B	
Helobdella robusta	106750	NR3B	
Capitella teleta	108381	NR3B	
Marisa cornuaretis	ABI97120.1	NR3B	
Drosophila melanogaster	NP_729340.1	NR3B	
Daphnia pulex	46682	NR3B	
Homo sapiens	X03225	NR3C1	GR
Homo sapiens	M16801	NR3C2	MR
Homo sapiens	M15716	NR3C3	PR
Homo sapiens	M20132	NR3C4	AR
Petromyzon marinus	AAK20930.1	NR3C5	CR
Myxine glutinosa	Q1KXY6	NR3C5	CR
Petromyzon marinus	AY028458.2	NR3C6	PR
Myxine glutinosa	Q1KXY5	NR3C6	SR2
Branchiostoma floridae	201600	NR3C	SR
Branchiostoma belcheri	BAI59768.1	NR3C	SR
Capitella teleta	170275	NR3D1	NR3A, ER
Platynereis dumerili	C0IR13	NR3D1	NR3A, ER
Octopus vulgaris	Q19AB0	NR3D1	NR3A, ER
Lottia gigantea	132166	NR3D1	NR3A, ER
Crassostrea gigas	BAF45381.1	NR3D1	NR3A, ER
Trichoplax adhaerens	16711	NR3E1	ERR
Homo sapiens	L13740	NR4A1	NGFIB
Rattus norvegicus	P22829.2	NR4A1	
Homo sapiens	X75918	NR4A2	NURR1
Homo sapiens	D38530	NR4A3	NOR1
Strongylocentrotus purpuratus	XP_786266.2	NR4A	Sp-nurr1
Branchiostoma floridae	223708	NR4A	
Drosophila melanogaster	U36762	NR4A4	HR38
Ixodes scapularis	XP_002400899.1	NR4A	
Caenorhabditis elegans	AAD03682.1	NR4A	NHR-6
Capitella teleta	167161	NR4A	
Lottia gigantea	136520	NR4A	
Homo sapiens	D88155	NR5A1	SF1
Homo sapiens	U93553	NR5A2	LRH1
Mus musculus	AAI37846.1	NR5A2	

Sheet1

Branchiostoma floridae	XP_002596353.1	NR5A	
Strongylocentrotus purpuratus	XP_791919.1	NR5A	
Saccoglossus kowalevskii	NP_001158442.1	NR5A	
Capitella teleta	186691	NR5A	
Lottia gigantea	196914	NR5A	
Drosophila melanogaster	M63711	NR5A3	FTZ-F1
Daphnia pulex	305379	NR5A	
Caenorhabditis elegans	Q19345.1	NR5A	NHR-25
Brugia malayi	EDP33144.1	NR5A	5a3Bruma
Brugia malayi	EDP36743.1	NR5A	5A4bBruma
Drosophila melanogaster	M63711	NR5B1	HR39
Lottia gigantea	120156	NR5B	
Capitella teleta	153528	NR5B	
Schistosoma mansoni	AF158103_1	NR5B	
Schistosoma japonicum	CAX73127.1	NR5B	
Schmidtea mediterranea	Contig915.2	NR5B	
Branchiostoma floridae	255231	NR5B	
Homo sapiens	U14666	NR6A1	GCNF1
Branchiostoma floridae	96828	NR6A	
Strongylocentrotus purpuratus	NP_001020384.1	NR6A	
Lottia gigantea	120003	NR6A	
Caenorhabditis elegans	Q9U2R6.2	NR6A	NHR-91
Ixodes scapularis	XP_002415570.1	NR6A	
Branchiostoma floridae	123436	NR7A	
Strongylocentrotus purpuratus	XP_784447.1	NR7	Sp-nr2C
Saccoglossus kowalevskii	ACY92467.1	NR7A	
Capitella teleta	224945	NR7A	
Helobdella robusta	103307	NR7A	
Lottia gigantea	104793	NR7A	
Acropora millepora	AAL29194.1	NR7A	RXR_ACRM1
Nematostella vectensis	108851	NR7A	NvNR3
Nematostella vectensis	101676	NR8A1	NvNR1
Acropora millepora	AAL29199.1	NR8A1	AmNR6
Nematostella vectensis	93844	NR8A2	
Nematostella vectensis	134436	NR9A	

** sequences that are no longer available in GenBank*

Part III

Evolution of steroidogenic enzymes

Independent elaboration of steroid hormone signaling pathways in metazoans

Gabriel V. Markov[a,b], Raquel Tavares[c,d], Chantal Dauphin-Villemant[e], Barbara A. Demeneix[b], Michael E. Baker[f], and Vincent Laudet[a,1]

[a]Molecular Zoology Team, Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Université Lyon 1 and Centre National de la Recherche Scientifique, Institut National de la Recherche Agronomique, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France; [b]Unité Mixte de Recherche 7221, Evolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, 75005 Paris, France; [c]Université de Lyon, Université Lyon 1, F-69000 Lyon, France; [d]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France; [e]Unité Mixte de Recherche 7622, Biologie du Developpement, Equipe Biogenèse des Signaux Hormonaux, Centre National de la Recherche Scientifique, Université Paris 6, 75005 Paris, France; and [f]Department of Medicine, University of California at San Diego, La Jolla, CA 92093-0693

Edited by Jan-Åke Gustafsson, Karolinska Institutet, Stockholm, Sweden, and approved May 26, 2009 (received for review December 4, 2008)

Steroid hormones regulate many physiological processes in vertebrates, nematodes, and arthropods through binding to nuclear receptors (NR), a metazoan-specific family of ligand-activated transcription factors. The main steps controlling the diversification of this family are now well-understood. In contrast, the origin and evolution of steroid ligands remain mysterious, although this is crucial for understanding the emergence of modern endocrine systems. Using a comparative genomic approach, we analyzed complete metazoan genomes to provide a comprehensive view of the evolution of major enzymatic players implicated in steroidogenesis at the whole metazoan scale. Our analysis reveals that steroidogenesis has been independently elaborated in the 3 main bilaterian lineages, and that steroidogenic cytochrome P450 enzymes descended from those that detoxify xenobiotics.

evolution | nuclear-receptor ligand | steroidogenesis

Multicellular organisms have complex endocrine systems, allowing responses to environmental stimuli, regulation of development, reproduction, and homeostasis. Nuclear receptors (NRs), a metazoan-specific family of ligand-activated transcription factors, play central roles in endocrine responses, as intermediates between signaling molecules and target genes (1). The NR family includes ligand-bound and orphan receptors, that is, receptors with no known ligand or for which there is no ligand pocket (2). Understanding NR evolution has been further improved by comparison of several completed genomes, particularly those of deuterostomes and ecdysozoans (3–6).

In contrast, evolution of NR ligands is still much debated. One hypothesis proposes that several independent gains and losses of ligand-binding ability in NRs occurred in protostomes and deuterostomes (7–9). A second hypothesis, pertaining to the NR3 subfamily (vertebrate steroid hormone receptors and estrogen-related receptor), proposes that before the divergence of protostomes and deuterostomes, there was an ancestral steroid receptor (AncSR) that was ligand-activated and that orphan receptors secondarily lost the ability to bind a ligand (10, 11). Phylogenetic analyses indicate that AncSR was able to bind estrogens (10, 11), which formed the basis for an intriguing "ligand exploitation model" (10, 12) for the evolution of vertebrate steroid receptors. In this model, estradiol (E2), a terminal product of the steroid biosynthetic pathway, was the first ligand for AncSR. Synthesis of E2 also requires the synthesis of steroid intermediates (Fig. 1). However, receptors for these intermediate steroids had not yet evolved. It was only after duplication of AncSR that NR3 receptors for these intermediate steroids evolved. The "ligand exploitation model" explains divergence in ligand specificity seen in steroid receptors, namely AR/NR3C4, GR/NR3C1, MR/NR3C2, PR/NR3C3, and ERs/NR3A (10, 12, Fig. 1*A* and *B*).

The ligand exploitation model is based mainly on NR data. But it has implications for the evolution of ligand synthesis. For example, it implies that 17ß-estradiol (E2) was a ligand for an ER in Urbilateria, the common ancestor of protostomes and deuterostomes (10, 12, 13). Such a hypothesis can be tested by searching for the origins of the enzymes involved in the synthesis of vertebrate adrenal and sex steroids.

As to steroid hormones in metazoans, there are major structural differences among different classes of steroids synthesized in vertebrates, insects and nematodes (Fig. S1). In insects and nematodes, the active steroid hormones retain all or most of the C17 side chain of cholesterol, with selective hydroxylations providing specificity for a given NR (Fig. 1*C*) (14–16). In contrast, in vertebrates, such as humans, synthesis of the main active steroids [estradiol for ERs, dihydroxytestosterone (DHT) for AR, progesterone (P4) for PR, cortisol for GR, and aldosterone for MR] begins with cleavage of the C17 side chain at C20 by CYP11A1 to yield pregnenolone (P5) (Fig. 1*B*) (17). Further enzymatic modifications involving selective hydroxylations, oxido-reductions and isomerizations of P5 and its metabolites yield ligands for adrenal and sex steroid receptors (Fig. 1*B*).

Many searches for "human"-type steroid hormones such as E2 or P4, throughout metazoan groups have been prone to artefacts and/or misidentification. To date, biochemical evidence (immunological and/or chromatographic methods linked to mass spectrometry) for presence of vertebrate steroids in lophotrochozoans, ecdysozoans, and cnidarians have not been substantiated by molecular characterization of enzymes directly involved in their de novo biosynthesis (18, 19). Thus, the presence of human-type steroids in protostomes remains an open question.

With this in mind, we investigated origins of enzymes in the pathways leading to steroid hormones in vertebrates. Our phylogenetic analyses of all enzymes known to be implicated in vertebrate (Fig. 1*B*) or ecdysozoan (Fig. 1*C*) steroid biosynthesis [belonging to the cytochrome P450 (CYP, 20, 21), short-chain dehydrogenase/reductase (SDR, 22), 3-ß hydroxysteroid dehydrogenase (HSD3B, 23) and steroid 5-α reductase (SRD5A) families] suggest that steroidogenesis was independently elaborated in vertebrates and protostomes, partly through recruitment of xenobiotic-metabolising CYPs. This has important implications on our views about the

Author contributions: G.V.M. and V.L. designed research; G.V.M. performed research; G.V.M., R.T., and V.L. analyzed data; and G.V.M., C.D.-V., B.A.D., M.E.B., and V.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed at: Université de Lyon, Université Lyon 1, Ecole Normale Supérieure de Lyon, Institut de Génomique Fonctionnelle de Lyon, Equipe de Zoologie Moléculaire, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5242, Institut National de la Recherche Agronomique, Unité Mixte de Recherche 1237, 46 Allée d'Italie, 69364 Lyon Cedex 07, France. E-mail: vincent.laudet@ens-lyon.fr.

This article contains supporting information online at www.pnas.org/cgi/content/full/0812138106/DCSupplemental.

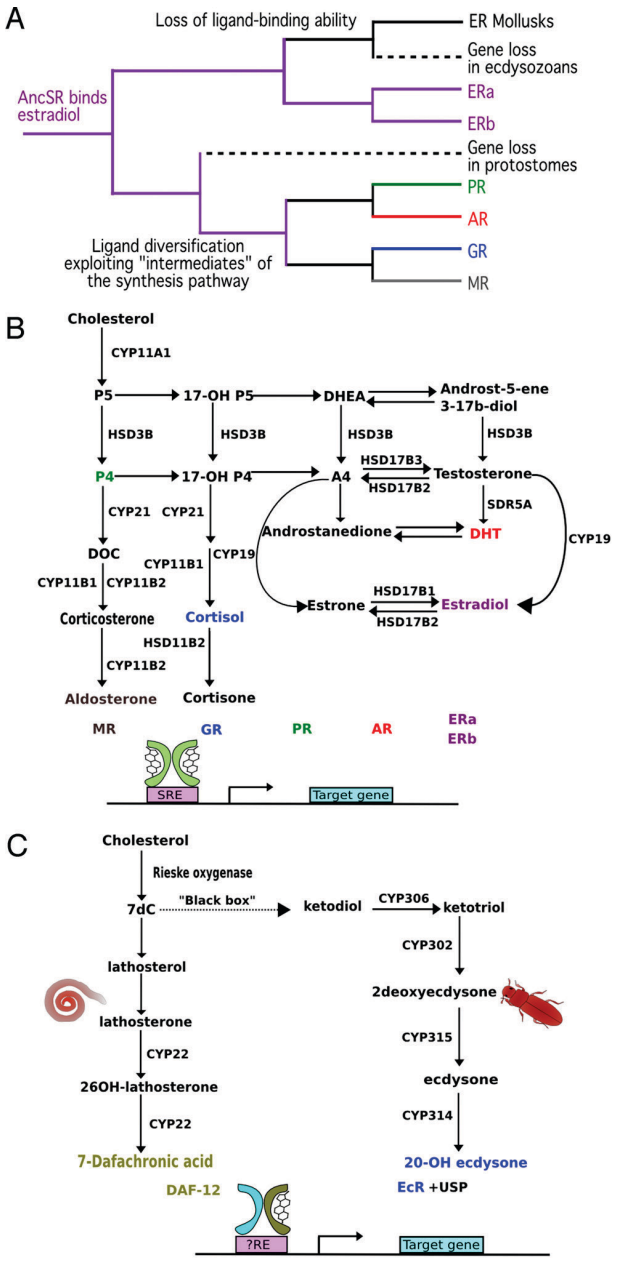


Fig. 1. Study background. (*A*) The ligand exploitation hypothesis. The ancestral receptor, that is supposed to bind estradiol, should have been lost in ecdysozoans, have lost its ligand-binding ability in mollusks, and have undergone ligand diversification through gene duplications in vertebrates. (*B*) The human steroid signaling pathway. (*C*) The steroid signaling pathways in ecdysozoans.

ligand-binding abilities of AncSR. Our analyses also show that there are pitfalls in extrapolating about the role in steroidogenesis of human or tetrapod genes to homologs in protostomes and other distant metazoans.

Results

General Strategy. To date, the best characterized steroidogenic enzymes belong to the CYP, SDR, HSD3B, and SRD5A families in human, mouse, *Drosophila*, and *C. elegans*. We screened recently completed metazoan genomes (Fig. S2) looking for orthologs of these steroidogenic enzymes. The retrieved sequences were used for phylogenetic reconstruction (using maximum likelihood coupled with bootstrapping) to determine their orthology with vertebrate sequences.

Orthology was defined on the basis of robust branches containing a consistent phylogenetic sampling (that is only protostome sequences for example) and/or high (>90%) bootstrap values. The large sequence variability present in some families such as CYPs precluded phylogenetic reconstruction, but our systematic survey revealed clear orthology in specific cases relevant to steroidogenesis evolution (24, 25).

Twenty-one complete genomes were screened, including 6 recently sequenced lophotrochozoan genomes, making it highly probable that not finding a given protein in a given zoological group (e.g., protostomes) indicates a real absence. Fig. 1*A* exemplifies our reasoning for estrogen receptors.

Our strategy successfully identified clear orthologs of known steroidogenic enzymes. For example, we identified in *Daphnia*, a crustacean, orthologs of enzymes that metabolize insect steroids (i.e., CYP302, CYP314, CYP315, and CYP306) known in *Drosophila* and other insects (Fig. 2*A* and Fig. 3 and Figs. S3 and S4).

Fig. 2 provides a general overview of the phylogeny of the 4 protein families analyzed. For more complete versions of these phylogenies, and the relevant specific branches see Figs. S3–S7. These phylogenies, based on several sequenced genomes, are in good agreement with published studies (25).

Polyphyletic Origin for Steroidogenic CYPs. The metazoan CYP family is currently divided into 11 clans (21, 22, Fig. 2*A* and Fig. S3), including the mito clan that clusters mitochondrial proteins in vertebrates and insects (Fig. 3 and Fig. S4). To date, all mitochondrial CYPs identified in vertebrates are involved in metabolism of endogenous compounds (e.g., CYP27A1 for bile acids) or hormone biosynthesis (CYP11A and CYP11B for steroid hormones), with CYP11A catalyzing cleavage of the cholesterol side chain at C20 (Fig. 1*B*). In contrast, arthropod mitochondrial CYPs include several xenobiotic-metabolizing proteins (e.g., CYP12) and enzymes catalyzing steroid biosynthesis (CYP302, CYP314, and CYP315) (Fig. 3). We observed that the vertebrate (CYP11A and CYP11B) and arthropod steroidogenic enzymes (CYP302, CYP314, and CYP315) do not form a monophyletic clade, and are rather dispersed at various places in the tree, often linked to non-steroidogenic proteins (Fig. 2*A*).

The most parsimonious scenario for these different activities within family- and lineage-specific duplications, is that these different steroidogenic activities arose independently in arthropod and vertebrate mitochondrial CYPs. This scenario implies that, if the substrate for an ancestral mito CYP was a steroid, then it probably was not a vertebrate steroid found in present-day organisms.

Similar conclusions can be drawn for other CYP clans. For example, in the CYP2 clan, steroidogenic activity seems to have appeared independently at least 3 times, in vertebrates, in insects, and in nematodes (Fig. 2*A* and Fig. S3). An important point is that many of the vertebrate members of this clan are known to be xenobiotic-metabolizing enzymes that are correlated with a high rate of lineage-specific duplications (24). Lineage-specific duplications are also abundant within lophotrochozoan members of this clan, thus indicating that these enzymes may be xenobiotic metabolisers.

SDR: Convergent Acquisition of the same Biochemical Activity. Short-chain dehydrogenase/reductase (SDRs) enzymes display a wide substrate spectrum, ranging from steroids, retinoids, alcohols, sugars, and aromatic compounds to xenobiotics (22). In terms of steroidogenesis, this family contains proteins with 17ß-hydroxysteroid dehydrogenase (17ßHSD) activity as well as 11ßHSD activity (26, 27), characterized as steroidogenic enzymes in vertebrates (HSD17B1, −2, and −3; HSD11B1, −2, and −3). Previous reports

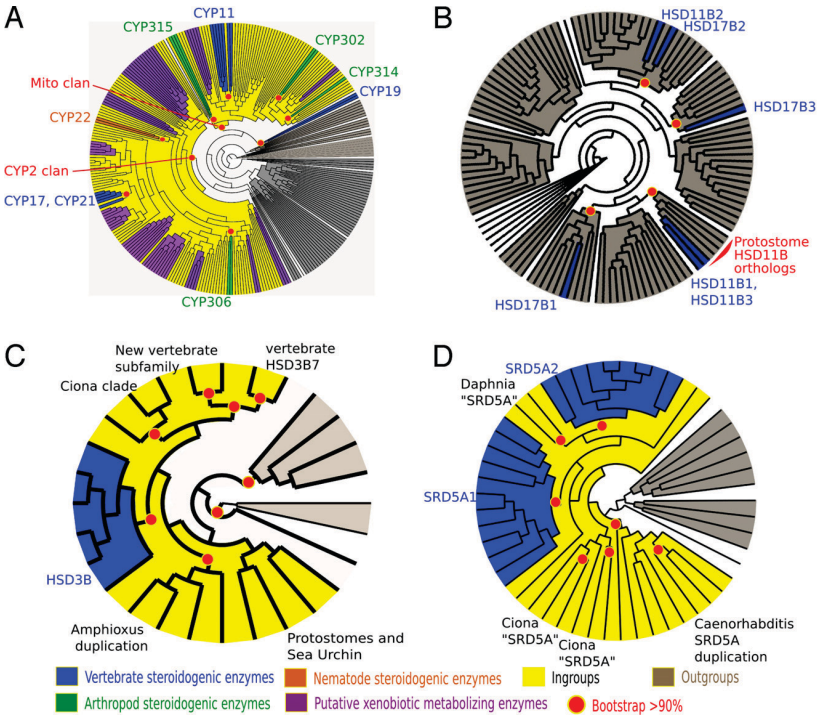


Fig. 2. Simplified Maximum-likelihood phylogenies of the CYP, SDR, HSD3B, and SRD5A families in metazoans. (*A*) CYP family. (*B*) SDR family. (*C*) HSD3B family. (*D*) SRD5A family. Steroidogenic proteins are highlighted in different colors. This clearly illustrates that in most cases the steroidogenic enzymes are dispersed in the evolutionary trees, suggesting independent acquisition of their steroid specificity.

(26, 28) noted that 17ßHSD and 11ßHSD activities arose independently many times in the SDR family. We confirm and extend this notion by finding that among the vertebrate steroidogenic proteins, only 1, HSD11B1, that is involved in the synthesis of cortisol from cortisone in vertebrates, has clear orthologues in lophotrochozoans (Fig. 2*B*). All of the other proteins, and especially those implicated in estrogen synthesis (HSD17B1, −2, and −3), arose from vertebrate-specific duplications and have no orthologues in protostomes.

The subfamily 3 of SDR (Fig. S5) illustrates this notion. It contains 1 human enzyme, HSD17B1 that clusters with a group containing the human RDH8, a photoreceptor-associated retinol dehydrogenase, as well as many vertebrate paralogs with uncharacterized activities. All of these vertebrate proteins cluster with proteins found in the cnidarian *Nematostella* whose activities are unknown. These data are consistent with the hypothesis that an ancestral HSD17B1 acquired the 17ßHSD biological function for synthesis of estradiol late during vertebrate evolution (8, 26).

HSD3B and SRD5A: Independent Lineage-Specific Duplications Within Chordates. The HSD3B family contains 5 robust clades (Fig. 2*C* and Fig. S6) that are the products of lineage-specific duplications in deuterostomes (23). The protostome sequences are external to these groups. According to the topology of this tree, *Ciona*, amphioxus and protostome proteins may have a HSD3B activity, but it is not possible to infer whether the function of the *Ciona* and protostome proteins is to metabolize vertebrate steroid hormones, bile acids, or other molecules. Similarly, in the SRD5A family (Fig. 2*D* and Fig. S7), lineage-specific duplications also occurred in vertebrates, *Ciona*, *Daphnia*, and *Caenorhabditis*, whereas the gene was lost in insects. Thus, in these 2 gene families, lineage-specific elaboration of steroidogenic enzymes occurred in vertebrates.

Two Key Enzymes Necessary to Generate Vertebrate Steroids Are Specific to Vertebrates. The first step of vertebrate steroid synthesis is the cleavage of the side chain present in cholesterol (17). This activity is catalyzed by CYP11A, which is, as discussed above, specific to vertebrates. This clearly shows that vertebrate-type steroids either may not be present outside vertebrates or, if present, are generated using enzymes of different phylogenetic origins. The latter case is an example of evolutionary convergence.

Interestingly, the very last step of estrogen synthesis, namely aromatization of testosterone or androstenedione, is catalyzed by CYP19, an aromatase, which arose in chordates. The phylogenetic analyses of CYP11A and CYP19 support our model that steroidogenic enzymes for adrenal and sex steroids arose in the deuterostome line, in which we also propose arose their cognate steroid receptors (7, 8).

Discussion

Independent Elaboration of Steroidogenesis in the 3 Main Bilaterian Lineages. Except for vertebrate SRD5A and HSD11B1, for which orthologous genes were found in protostomes and/or cnidarians (even if their biochemical activity is not known), other enzymes known to be involved in steroidogenesis in arthropods, nematodes, or vertebrates have no clear orthologues outside their respective metazoan phyla. This indicates that the steroidogenic enzymes have evolved independently within each phylum, through lineage-specific duplications, and subsequent neofunctionalization. Such convergent evolution of synthesis pathways for complex molecules is not unique: examples include morphine synthesis in plants and animals (29) and gibberellin in plants and fungi (30).

An important point is that the major active steroid hormones identified so far in vertebrates, arthropods, and nematodes have important differences in their structures (Figs. 3 and 4 and Fig. S1), which is consistent with our phylogenetic analyses of steroidogenic enzymes and argues for independent evolution of the steroidogenic pathway in these metazoan groups.

To clarify the fundamentally different characteristics of the steroid hormones across metazoan phyla and to highlight their

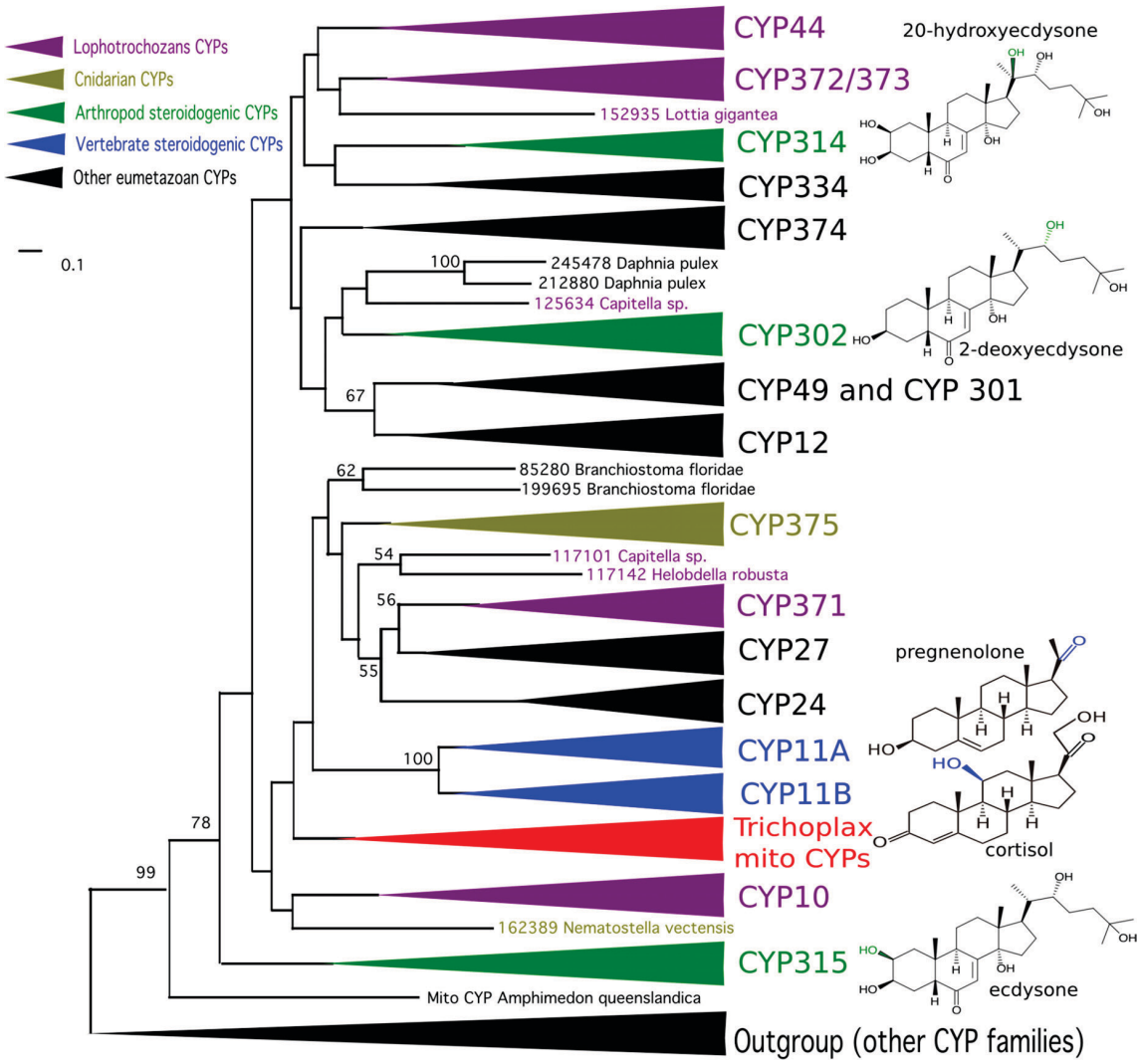


Fig. 3. A simplified Maximum-Likelihood phylogeny of the mitochondrial clan. Vertebrate and arthropod steroidogenic enzymes are highlighted in blue and green, respectively, and the molecules they produce are indicated. These molecules are 20-OH Ecdysone for CYP314, Ecdysone for CYP315, 2deoxyecdysone for CYP302, pregnenolone for CYP11A, and cortisol for CYP11B. Colored residues in the chemical formulas are those that are modified by the catalytic reaction.

independent evolutionary elaboration, one could apply a taxonomic based nomenclature, namely lophosteroids, ecdysosteroids, vertebrosteroids, and cnidosteroids (Fig. 4 and *SI Text*). Each of these compounds has a defined structural feature; for example, vertebrosteroids exhibit a characteristic cleavage of the long C17 side chain found in cholesterol. It is only when more biochemical and functional data become available in non-model taxa such as lophotrochozoans that a clear and unambiguous nomenclature can be defined.

Caution Is Needed in Assigning a Function Solely from Sequence Data. The CYP and SDR family members are known to exhibit a huge variation of substrate specificity, even at the subfamily level. This indicates that one must exercise caution in attributing vertebrate-like steroidogenic activities to homologs in protostomes and cnidarians. For example, although it was convincingly shown that LET-767 is able to transform androgens into estrogens in mammalian cell cultures, as HSD17B3 does, and that this substrate-specificity can be altered by selective mutations (31), it does not necessarily follow that LET-767 and HSD17B3 have similar functions in vivo. Ecdysozoans have cholesterol-like steroids, in which there is a side-chain at C-17. Thus, there is no C17 alcohol or ketone for modification by a 17ß-HSD in nematode cells. Future characterization of the biological activity of LET-767 in *C. elegans* is necessary to provide insights into the evolution of substrate specificity in 17ß-HSD and its paralogs.

CYP19 Is a Chordate Aromatase. The only non-vertebrate to contain a CYP19 ortholog is amphioxus, a chordate that is a close relative of vertebrates. Thus, our analysis (Fig. 2*A*) shows that, in contrast to recent claims (32), there is no support for the presence of an ortholog of vertebrate CYP19 in protostomes and cnidarians. This could be explained either by long-branch attraction in chordates CYP19 (which would be consistent with a functional shift) or by secondary loss of the CYP19 genes in protostomes and cnidarians. Since this is observed for other CYP families, for example CYP20, which seems to be orthologous to the sponge CYP38, with no counterparts in cnidarians and protostomes, we favor the hypothesis of secondary loss of an ancestral gene with no aromatizing activity. If an aromatization reaction really occurs in some lophotrochozoans (33), our analysis indicates that this reaction is carried out by a protein that is not a member of the chordate CYP19 family.

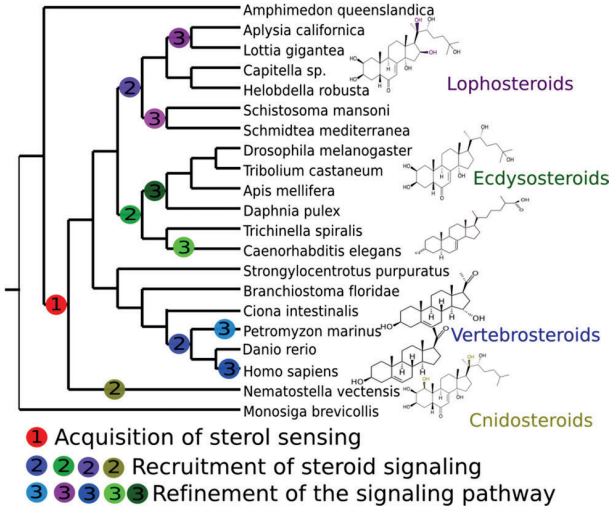


Fig. 4. A hypothesis about the acquisition of steroidogenic pathways in metazoans. We propose that steroid sensing by NR was already present in the last common ancestor of all eumetazoans (step 1), but that steroid signaling was independently recruited many times from slightly different molecules (step 2), with subsequent refinement in some lineages (e.g., lamprey; step 3).

This could be a CYP from the CYP1 and CYP3 clan (especially CYP3A4), which can aromatize indoline (34), or it may even be a protein from another family. Such an example of convergent evolution has already been described in the case of allene oxide synthase of a coral being able to metabolize a fatty acid peroxide in a way that was previously thought to be specific to CYPs (35). However, aromatization of steroids by these enzymes or a lophotrochozoan enzyme has not been reported.

Implications for the Presence of Vertebrate Steroids in Protostomes and Cnidarians. Our phylogenetic analyses are relevant to studies in the comparative endocrinology field, which discuss the presence of vertebrate-type steroids and steroidogenic activities in non-model species, especially in protostomes or cnidarians (33, 36). In non-vertebrate species, the presence of steroids is usually monitored by radioimmunoassay (RIA) using antibodies generated against vertebrate hormones. Most importantly, vertebrate antibodies may cross-react with other steroids, including non-vertebrate steroids.

The limits of such approaches were clearly demonstrated in sea lamprey. Whereas classical RIA studies led to the identification of vertebrate-type steroids (37), more recent experiments, based on high performance liquid chromatography (HPLC) with 2 different solvents showed that the main circulating steroids are 15α-hydroxylated steroids and not vertebrate-type steroids as determined through RIA (38).

We show in this paper that genes orthologous to the vertebrate steroidogenic enzymes are not present in lophotrochozoans and cnidarians. Thus, there is no reason "a priori," other than a residual anthropomorphism, to search specifically for the presence of vertebrate-type steroids in lophotrochozoans and cnidarians, and to imagine that those vertebrate steroids—if present—would be more likely to act as hormones, through vertebrate-like transduction pathways, than other steroids that are supposed to be present in non-vertebrate animals.

In our opinion, the first step in characterizing new steroidogenic pathways in non-model organisms should be identification of all steroids with sensitive methods such as GC-MS, and verification that these molecules are synthesized de novo from a defined sterol precursor. The physiological effects of these molecules should be tested, and the potential enzymes capable of catalyzing the different steps of the synthesis pathway identified. This is of course very challenging experimentally, but the only way to progress and to build knowledge on solid ground.

Xenobiotic-Metabolizing CYPs Are Ancestors of Steroidogenic Enzymes. Many CYPs hydroxylate xenobiotics, which increases their solubility, facilitating excretion of the hydroxylated metabolite, and in optimal situation, leading to metabolites with reduced toxicity due to a lower affinity for enzymes and/or receptors (20, 21). On the other hand, CYPs can also lead to an increased affinity of lipophilic molecules for enzymes and/or receptors. Thus, hydroxylation of various lipophilic molecules, such as cholesterol, ergosterol, bile acids, retinoids, and vertebrate steroids, by CYPs can yield ligands that activate nuclear receptors (39–42). Indeed, selective expression of CYPs in specific tissues is an important mechanism for regulating the actions of vertebrate steroids and other ligands. Phylogenetic analysis of CYPs indicates that they are ancient and found in bacteria, yeast, and basal metazoans (20, 21, Fig. 2 and Figs. S3 and S4), preceding the evolution of steroid receptors in arthropods and deuterostomes. The broad substrate specificity and micromolar affinity of CYPs for xenobiotics allows a few CYPs to protect their host organism. Our phylogenetic analyses indicate that key steroidogenic enzymes, such as CYP11A, CYP11B, and CYP19 (aromatase) in vertebrates, CYP22 in nematodes, or CYP314, CYP315, CYP306, and CYP302 in arthropods, arose late in animal evolution and are most likely descended from CYPs that metabolize xenobiotics. These steroidogenic CYPs have evolved increasing their specificity for different steroids regarding hydroxylation, aromatization or cleavage of the C17 side chain. This specificity is an important mechanism for regulating steroid hormone action.

Like xenobiotic-metabolizing CYPs, some nuclear receptors are xenobiotic sensors, in that these transcription factors bind a wide range of molecules with micromolar affinity (42). An example of an ancient liganded receptor system is the NR1H/NR1I/NR1J group containing FXR, LXR, ECR, PXR, CAR, and VDR in vertebrates and also DHR96 in *Drosophila* and DAF12 in nematodes. Some of these receptors (FXR, PXR, CAR, and VDR) regulate CYPs and other transcription factors that detoxify xenobiotics. VDR, ECR, and LXR also are steroid-regulated transcription factors (43, 44). A characteristic of the nuclear receptors that respond to xenobiotics is their broad substrate specificity (43, 45), which is important in protection from the effects of xenobiotics. In contrast, chordate steroid receptors have nanomolar affinity for different adrenal and sex steroids, which is important in selective activation of endocrine pathways. Interestingly, 17ß -ethynylestradiol and the xenoestrogen, 4-nonylphenol, activate responses for detoxification of xenobiotics (46, 47), which suggests that the vertebrate ER activates some responses to xenobiotics.

AncSR Was Not a Hormone Receptor, but More Likely a Sensor. Our phylogenetic analysis of steroidogenic enzymes favors the independent elaboration of different steroid synthesis pathways in metazoan groups. These data support the hypothesis that the responses of nuclear receptors in vertebrates and arthropods evolved independently (7). This model differs from the "ligand exploitation" model (10), in which the first active steroid would be estradiol, which would act through the AncSR in all bilaterians. Only later on, other "intermediate" steroids (androgens, corticosteroids, progestins, etc.) would have become ligands after gene duplication of the AncSR gave rise to new receptors that could exploit these intermediates (see Fig. 1*A* and *B*). This model indeed implies that the whole pathway governing estrogen production evolved in an ancestral bilaterian and that enzymes involved in estrogen synthesis (the ancestral ligand) are evolutionarily conserved in metazoans, and this is not what we observed.

To date, all of the binding data on ancestral SR were interpreted in the framework of vertebrate steroids being present in all metazoans and opposing an unliganded AncSR to an hormone-binding AncSR. Given the fact that vertebrate steroid hormones are not

synthesized in other metazoans and that there are many possible crosstalks between the hormone synthesis and xenobiotic detoxification pathways, we propose that AncSR was able to bind estrogen with micromolar affinity but that it was not an hormone receptor, but rather a sensor, that was able to bind a broad range of various metabolites, such as sterol food derivatives and xenobiotics. Indeed, some current sensors, like PXR, are able to bind both xenobiotics and estradiol (48).

Materials and Methods

Protein sequences were retrieved in various public databases (Dataset S1), aligned with muscle (49), and alignments were checked by eye and edited with Seaview (50). Phylogenetic trees were made using PHYML (51), a fast and accurate maximum likelihood heuristic method, under the JTT substitution model (52), with 100 bootstrap replicates. The trees were first made with sequences of experimentally characterized proteins, for which a cDNA was cloned. Then the sampling was completed with EST-based or ab initio predictions to check the presence of the studied genes in non-model organisms. For additional details see *SI Text*.

ACKNOWLEDGMENTS. We thank Stéphanie Bertrand, Pascale Chevret, Ferdinand Marlétaz, and Loïc Ponger for technical advice; François Bonneton, Frédéric Brunet, Guillaume Lecointre, Mathilde Paris, Bruno Querat, Marc Robinson-Rechavi, and Michael Schubert for useful discussions; David Nelson for naming new CYP families; and the reviewers and editor for constructive comments. This work was supported by Ministère de l'Education Nationale, de la Recherche et de la Technologie, the Cascade Network of Excellence (FOOD-CT-2003–506319), Ecole Normale Supérieur e de Lyon, and Centre National de la Recherche Scientifique.

1. Gronemeyer H, Gustafsson JA, Laudet V (2004) Principles for modulation of the nuclear receptor superfamily. *Nat Rev Drug Discov* 3:950–964.
2. Benoit G, et al. (2006) International Union of Pharmacology. LXVI. Orphan nuclear receptors. *Pharmacol Rev* 58:798–836.
3. Laudet V, Hanni C, Coll J, Catzeflis F, Stehelin D (1992) Evolution of the nuclear receptor gene superfamily. *EMBO J* 11:1003–1013.
4. Laudet V (1997) Evolution of the nuclear receptor superfamily: Early diversification from an ancestral orphan receptor. *J Mol Endocrinol* 19:207–226.
5. Escriva H, et al. (1997) Ligand binding was acquired during evolution of nuclear receptors. *Proc Natl Acad Sci USA* 94:6803–6808.
6. Bertrand S, et al. (2004) Evolutionary genomics of nuclear receptors: From twenty-five ancestral genes to derived endocrine systems. *Mol Biol Evol* 21:1923–1937.
7. Escriva H, Delaunay F, Laudet V (2000) Ligand binding and nuclear receptor evolution. *Bioessays* 22:717–727.
8. Baker ME (2003) Evolution of adrenal and sex steroid action in vertebrates: A ligand-based mechanism for complexity. *Bioessays* 25:396 400.
9. Iwema T, et al. (2007) Structural and functional characterization of a novel type of ligand-independent rxr-usp receptor. *EMBO J* 26:3770–3782.
10. Thornton JW (2001) Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci USA* 98:5671–5676.
11. Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301:1714–1717.
12. Bridgham JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312:97–101.
13. Bridgham JT, Brown JE, Rodríguez-Marí A, Catchen JM, Thornton JW (2008) Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet* 4:e1000191.
14. Motola DL, et al. (2006) Identification of ligands for DAF-12 that govern dauer formation and reproduction in *C. elegans. Cell* 124:1209–1223.
15. Gerisch B, et al. (2007) A bile acid-like steroid modulates *Caenorhabditis elegans* lifespan through nuclear receptor signaling. *Proc Natl Acad Sci USA* 104:5014–5019.
16. Rewitz KF, Rybczynski R, Warren JT, Gilbert LI (2006) The Halloween genes code for cytochrome P450 enzymes mediating synthesis of the insect molting hormone. *Biochem Soc Trans* 34:1256–1260.
17. Payne AH, Hales DB (2004) Overview of steroidogenic enzymes in the pathway from cholesterol to active steroid hormones. *Endocr Rev* 25:947–970.
18. Lafont R, Mathieu M (2007) Steroids in aquatic invertebrates. *Ecotoxicology* 16:109–130.
19. Markov GV, Paris M, Bertrand S, Laudet V (2008) The evolution of the ligand/receptor couple: A long road from comparative endocrinology to comparative genomics. *Mol Cell Endocrinol* 293:5–16.
20. Nelson DR (1998) Metazoan cytochrome P450 evolution. *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol* 121:15–22.
21. Nelson DR (2006) Cytochrome P450 nomenclature, 2004. *Methods Mol Biol* 320:1–10.
22. Jornvall H, et al. (1995) Short-chain dehydrogenases/reductases (SDR). *Biochemistry* 34:6003–6013.
23. Simard J, et al. (2005) Molecular biology of the 3beta-hydroxysteroid dehydrogenase/delta5-delta4 isomerase gene family. *Endocr Rev* 26:525–582.
24. Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet* 3:e67.
25. Rewitz KF, Gilbert LI (2008) *Daphnia* Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone: evolutionary implications. *BMC Evol Biol* 8:60.
26. Baker ME (2001) Evolution of 17beta-hydroxysteroid dehydrogenases and their role in androgen, estrogen and retinoid action. *Mol Cell Endocrinol* 171:211–215.
27. Baker ME (2004) Evolutionary analysis of 11beta-hydroxysteroid dehydrogenase-type 1, -type 2, -type 3 and 17beta-hydroxysteroid dehydrogenase-type 2 in fish. *FEBS Lett* 574:167–170.
28. Belyaeva OV, Kedishvili NY (2006) Comparative genomic and phylogenetic analysis of short-chain dehydrogenases/reductases with dual retinol/sterol substrate specificity. *Genomics* 88:820–830.
29. Boettcher C, Fellermeier M, Boettcher C, Dräger B, Zenk MH (2005) How human neuroblastoma cells make morphine. *Proc Natl Acad Sci USA* 102:8495–8500.
30. Kawaide H (2006) Biochemical and molecular analyses of gibberellin biosynthesis in fungi. *Biosci Biotechnol Biochem* 70:583–590.
31. Desnoyers S, et al. (2007) Caenorhabditis elegans LET-767 is able to metabolize androgens and estrogens and likely shares common ancestor with human types 3 and 12 17beta-hydroxysteroid dehydrogenases. *J Endocrinol* 195:271–279.
32. Tiwary BK, Li W (2009) Parallel evolution between aromatase and androgen receptor in the animal kingdom. *Mol Biol Evol* 26:123–129.
33. Osada M, Tawarayama H, Mori K (2004) Estrogen synthesis in relation to gonadal development of Japanese scallop, *Patinopecten yessoensis*: gonadal profile and immunolocalization of P450 aromatase and estrogen *Comp Biochem Physiol B Biochem Mol Biol* 139:123–128.
34. Sun H, et al. (2007) Dehydrogenation of indoline by cytochrome P450 enzymes: A novel ''aromatase'' process. *J Pharmacol Exp Ther* 322:843–851.
35. Oldham ML, Brash AR, Newcomer ME (2005) The structure of coral allene oxide synthase reveals a catalase adapted for metabolism of a fatty acid hydroperoxide. *Proc Natl Acad Sci USA* 102:297–302.
36. Twan WH, Hwang JS, Chang CF (2003) Sex steroids in scleractinian coral, *Euphyllia ancora*: implication in mass spawning *Biol Reprod* 68:2255–2260.
37. Kime DE, Larsen LO (1987) Effect of gonadectomy and hypophysectomy on plasma steroid levels in male and female lampreys (*Lampetra fluviatilis*, L.). *Gen Comp Endocrinol* 68:189–196.
38. Lowartz S, et al. (2003) Blood steroid profile and in vitro steroidogenesis by ovarian follicles and testis fragments of adult sea lamprey, *Petromyzon marinus. Comp Biochem Physiol A Mol Integr Physiol* 134:365–376.
39. Baker ME (2005) Xenobiotics and the evolution of multicellular animals: Emergence and diversification of ligand-activated transcription factors. *Integr Comp Biol* 45:172–178.
40. Gilbert LI, Rybczynski R, Warren JT (2002) Control and biochemical nature of the ecdysteroidogenic pathway. *Annu Rev Entomol* 47:883–916.
41. Nebert DW, Russell DW (2002) Clinical importance of the cytochromes P450. *Lancet* 360:1155–1162.
42. Moore DD, et al. (2006) International Union of Pharmacology. LXII. The NR1H and NR1I receptors: constitutive androstane receptor, pregnene X receptor, farnesoid X receptor alpha, farnesoid X receptor beta, liver X receptor alpha, liver X receptor beta, and vitamin D receptor. *Pharmacol Rev* 58:742–759.
43. Xie W, Evans RM (2001) Orphan nuclear receptors: The exotics of xenobiotics. *J Biol Chem* 276:37739–37742.
44. Chawla A, Repa JJ, Evans RM, Mangelsdorf DJ (2001) Nuclear receptors and lipid physiology: Opening the X-files. *Science* 294:1866–1870.
45. Blumberg B, et al. (1998) SXR, a novel steroid and xenobiotic-sensing nuclear receptor. *Genes Dev* 12:3195–3205.
46. Mortensen AS, Arukwe A (2007) Effects of 17alpha-ethynylestradiol on hormonal responses and xenobiotic biotransformation system of Atlantic salmon (*Salmo salar*). *Aquat Toxicol* 85:113–123.
47. Meucci V, Arukwe A (2006) The xenoestrogen 4-nonylphenol modulates hepatic gene expression of pregnane X receptor, aryl hydrocarbon receptor, CYP3A, and CYP1A1 in juvenile Atlantic salmon (*Salmo salar*). *Comp Biochem Physiol C Toxicol Pharmacol* 142:142–150.
48. Xue Y, et al. (2007) Crystal structure of the pregnane X receptor-estradiol complex provides insights into endobiotic recognition. *Mol Endocrinol* 21:1028–1038.
49. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
50. Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12:543–548.
51. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
52. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.

Supporting Information

Markov et al. 10.1073/pnas.0812138106

SI Text

Searching Strategy. Orthologs from the proteins of interest (i.e., all mentioned vertebrate, arthropod, and nematode steroidogenic enzymes) were searched by blasting again the mentioned databases. The phylogenetic position of the organisms that were screened in this study is indicated in Fig. S2, with information on the data quality. All sequence hits were retrieved, and the dataset was cleaned using the following criteria: sequences lacking a conserved family motif (e.g., SDR cofactor binding site TG***G*G) were discarded, and truncated sequences were also discarded when there were found to be members of paralog groups. The majority of sequences from *Schmidtea mediterranea*, *Trichinella spiralis*, and *Aplysia californica* were eliminated during this step, because ab initio predictions were shown to be less accurate than the EST-based predictions that were available for other species. Sequences were checked by eye in SEAVIEW (2) to eliminate too divergent positions and unaligned regions before phylogenetic reconstruction.

Protein Sequences. Additionally, as *SI Text* we provide expanded trees corresponding to Figs. 2 and 3. For each tree, all sequence accession numbers are grouped in Dataset S1, (1 sheet per tree). The given accession numbers are GenBank IDs, jgi IDs (only numbers, mainly for *Capitella*, *Nematostella*, *Helobdella*, *Lottia*, *Trichoplax, Daphnia*, and some sequences from *Branchiostoma floridae*), or Ensembl IDs (sequences beginning with *ENS0000*). Two additional sequences are not in those databases:

- Pr5 from *Aplysia californica* is a homemade GENSCAN (1) prediction from the GenBank contig AASC01065054.1.
- 002 from *Amphimedon queenslandica*, which is a manually annotated prediction based upon traces, that is available on the online website from David Nelson:
http://drnelson.utmem.edu/biblioC.html.

Proteins that are marked with * are those for which corrected intron-exon boundaries were manually performed by D. Nelson and that are available on the indicated Web page.

Detection of Annotation Errors. Illustrating the difficulties in assessing orthology when partial data sets are used, the recently cloned *Branchiostoma belcheri* protein BAF61103.1, originally described as CYP11 (3) is in fact a member of CYP374, a distant paralog group of deuterostome CYPs, which was lost in vertebrates (Fig. S4). This shows that experimental data concerning the enzymatic activities of the CYPs can be biased by a wrong identification linked to a partial phylogenetic analysis (4). Similarly, the *Branchiostoma belcheri* BAF61104.1, that is described in the same paper as a CYP17 is clearly not an ortholog of the vertebrate and *Branchiostoma floridae* CYP17s, but a paralog from a subfamily where the gene may have been lost in vertebrates too.

A Nomenclature Note About Fig. 4. We propose the name "ecdysosteroid" to name steroids from ecdysozoans because the classically used name "ecdysteroids" is used to describe steroids from arthropods and steroids from plants that have the same structure.

1. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Bio* 268:78–94.
2. Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12:543–548.
3. Mizuta T, Kubokawa K (2007) Presence of sex steroids and cytochrome P450 (CYP) genes in amphioxus. *Endocrinology* 148:3554–3565.
4. Markov G, Lecointre G, Demeneix B, Laudet V (2008) The "street light syndrome", or how protein taxonomy can bias experimental manipulations. *Bioessays* 30:349–357.

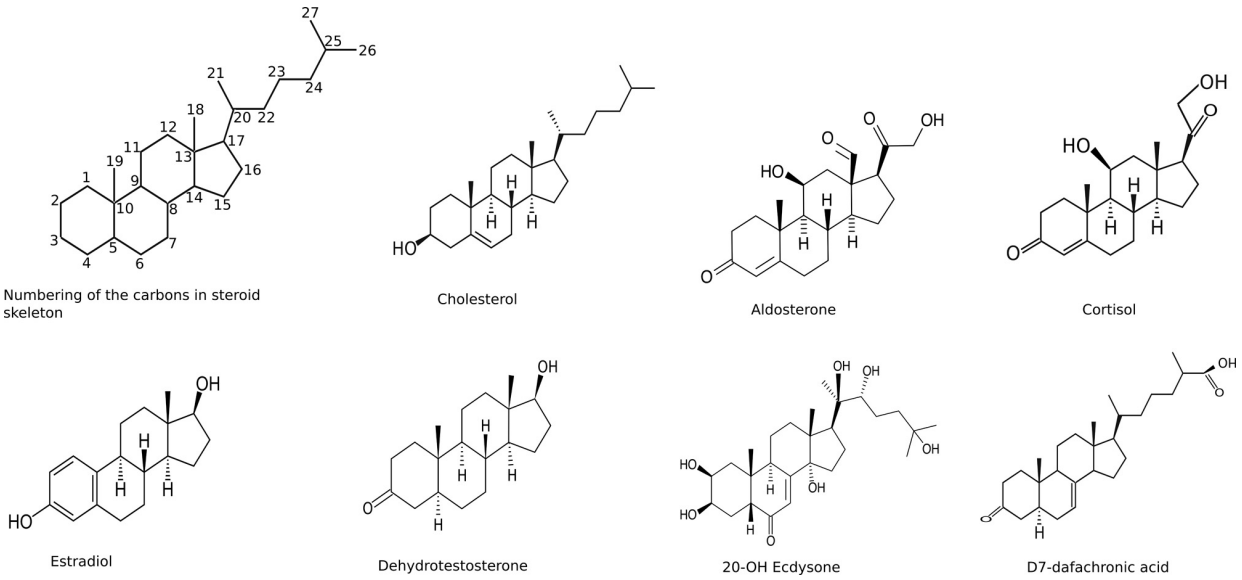


Fig. S1. Structure of the sterol ring, of cholesterol, and of the steroid hormones of human, *Drosophila melanogaster* and *Caenorhabditis elegans*. The sterol ring numbering is indicated. Aldosterone, cortisone, estradiol and dihydrotestosterone are human steroid hormones. 20-OH ecdysone is the main steroid hormone in *Drosophila melanogaster*, whereas delta-7-dafachronic acid is 1 of the 2 steroids in *Caenorhabditis elegans*.

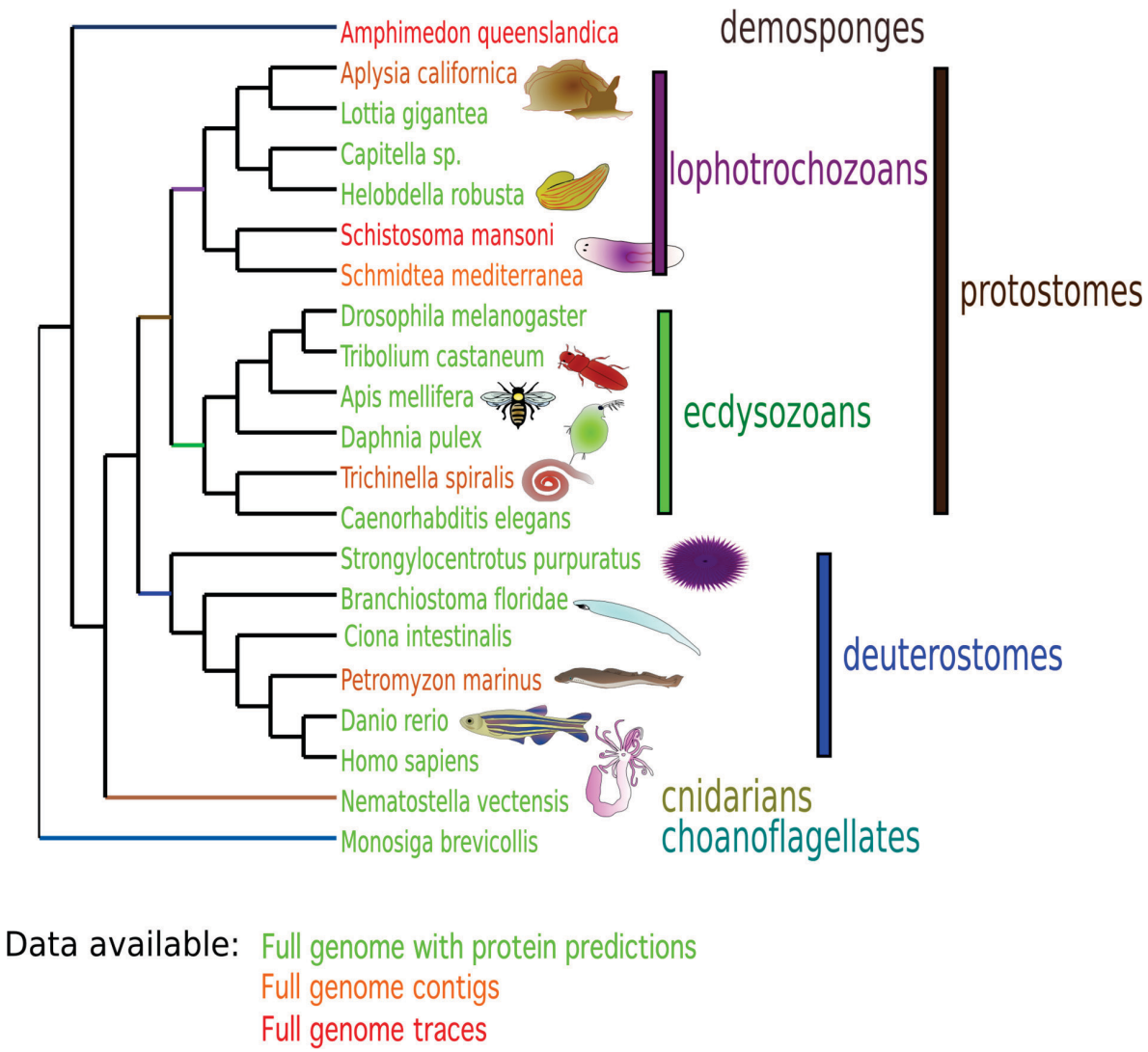


Fig. S2. Genomic data used in this study. The genomic model species that were screened in this study are indicated, with complementary information about their phylogenetic relationships and about the quality of their genome data. Species in green are those for which EST-based gene predictions are available. The genome of species in orange is provided as contigs, that were used for ab initio predictions. The genome of species in red is available only as traces.

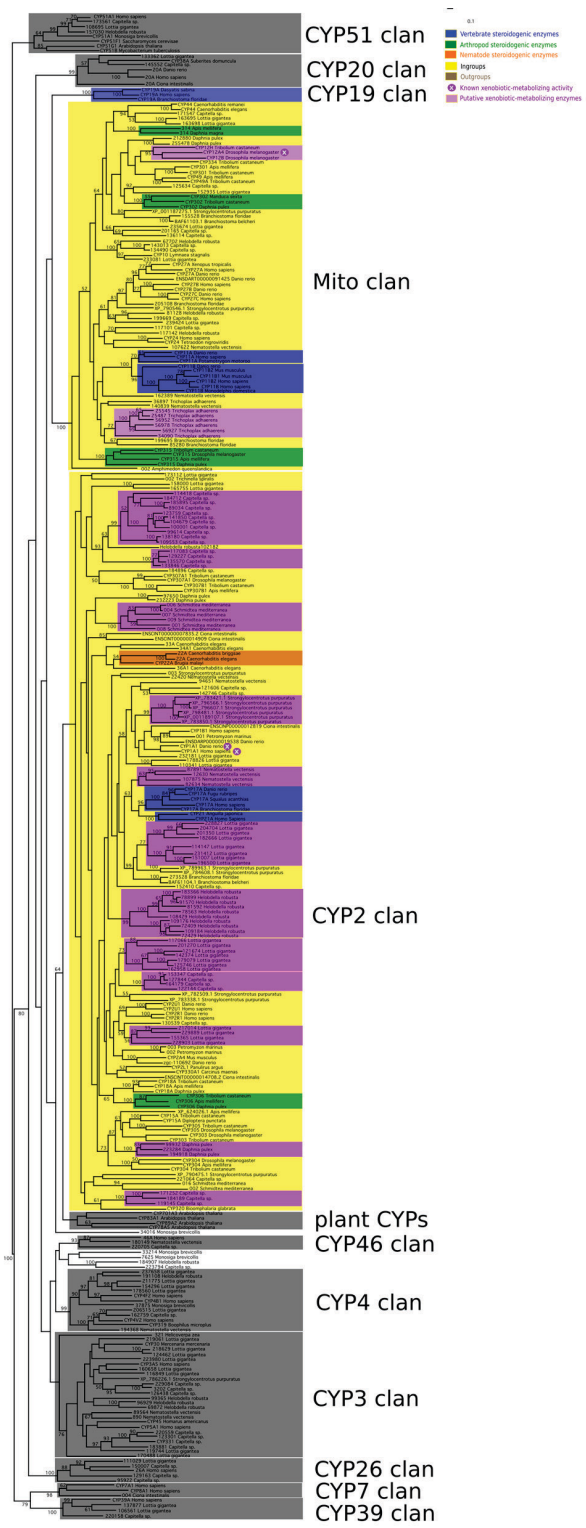


Fig. S3. Phylogeny of the CYP family. A maximum-likelihood analysis of the CYP family. Vertebrate steroidogenic proteins are highlighted in blue, arthropod steroidogenic proteins are in green and nematode steroidogenic proteins are in orange. Enzymes with known xenobiotic-metabolizing activity are indicated by circled ''X'', and proteins resulting from abundant lineage-specific duplication, that are thus candidate xenobiotic-metabolising enzymes, are highlighted by purple boxes. For details about the mito clan, see also Fig. S4.

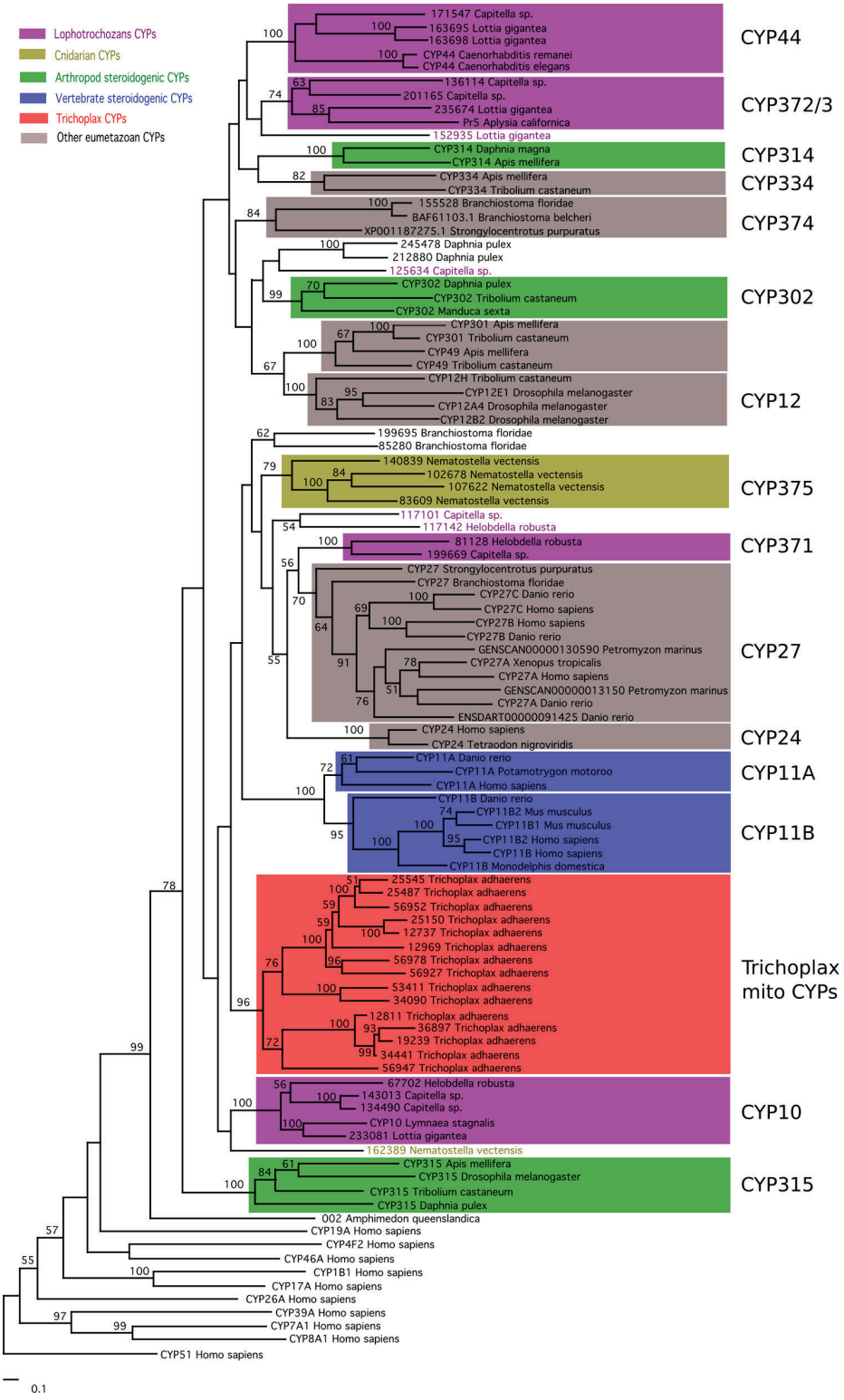


Fig. S4. Phylogeny of the mitochondrial CYP clan. A maximum-likelihood analysis of the mitochondrial CYP clan. Proteins are named according to classical CYP nomenclature when an official name exists. At least 3 groups of paralogs with unknown activity were found in lophotrochozoans (CYP10, CYP372/CYP373, and CYP371). Vertebrate steroidogenic CYPs are highlighted in blue, arthropod steroidogenic CYPs are in green, lophotrochozoan mito CYPs are in purple, cnidarian mito CYPs are in light brown, Trichoplax mito CYPs are in red, and other supported mito CYP clades are in gray.

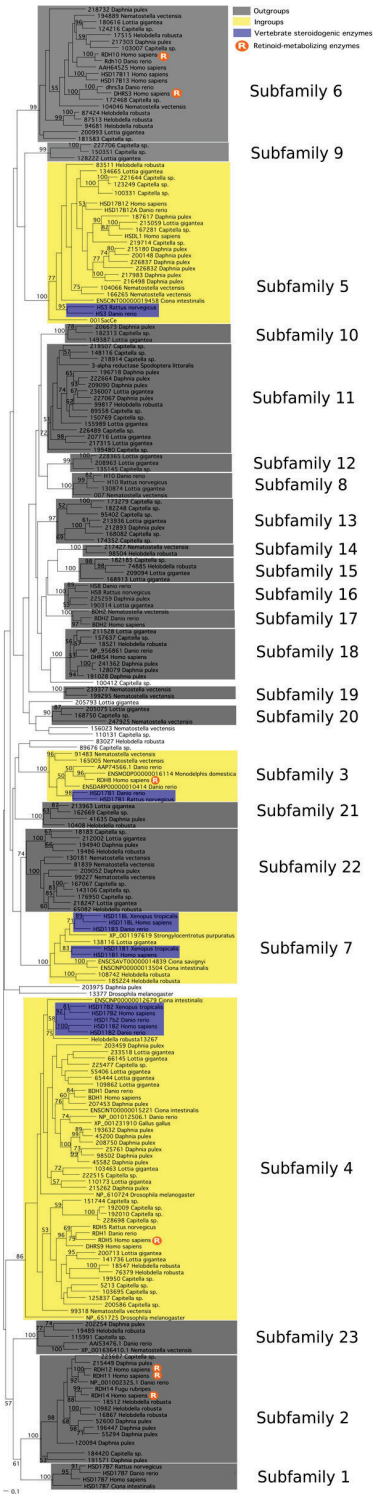


Fig. S5. Phylogeny of the SDR family. Our analysis reveals the existence of 23 strongly supported families that were arbitrarily named from 1 to 23, and are highlighted in gray. Groups of vertebrate steroidogenic enzymes are in blue, and other member of the same subfamily are in yellow. Enzymes with known retinoid-metabolizing activities are indicated by circled ''R.''

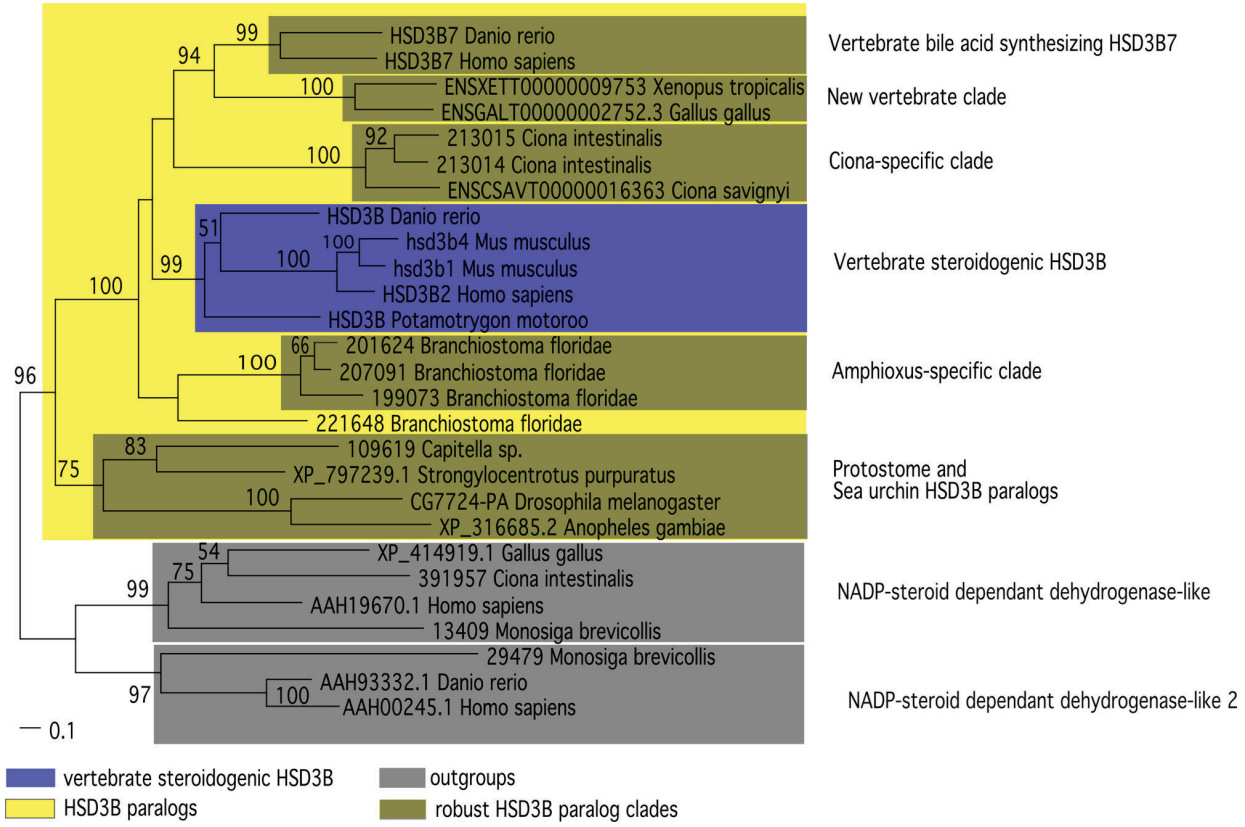


Fig. S6. Phylogeny of the HSD3B family. A maximum-likelihood analysis of the HSD3B family. Groups of vertebrate steroidogenic enzymes are in blue, other members of the same subfamily are in yellow. Outgroups are in gray. Robust HSD3B paralogs clades (those who are indicated by red dots in Fig. 2) are highlighted in yellow+gray.

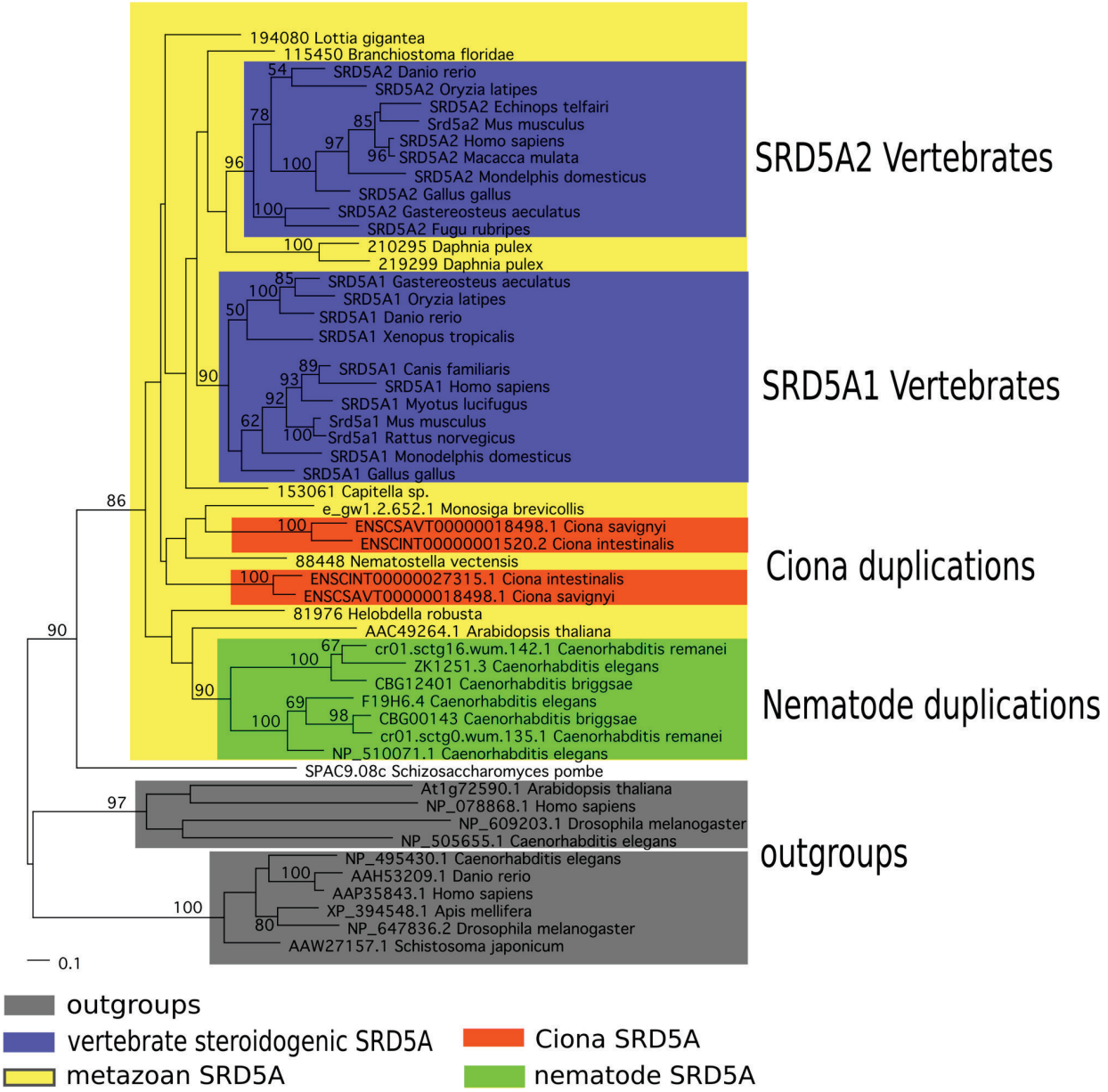


Fig. S7. Phylogeny of the SRD5A family. A maximum-likelihood analysis of the SRD5A family. Vertebrate SRD5A are highlighted in blue, nematode duplicated SRD5 are highlighted in green and ciona duplications are highlighted in orange. The metazoan SDR5A family is in yellow and the outgroups in gray.

Other Supporting Information Files

Dataset S1 (XLS)

Part IV

Comparative anatomy of steroidogenic pathways

Title page

Running head: Steroids are Domesticated Cholesterol Metabolites

Steroid Hormones are Domesticated Cholesterol Metabolites

Gabriel V. Markov[1,2], Guillaume Lecointre[3], Vincent Laudet[1*]

1 Molecular Zoology Team; Institut de Génomique Fonctionnelle de Lyon; Université de Lyon; Université Lyon 1; CNRS; INRA; Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France.
2 UMR 7221 - Evolution des Régulations Endocriniennes. Muséum National d'Histoire Naturelle, Paris, France.
3 UMR 7138 CNRS-UPMC-IRD-MNHN-ENS, CP39 Département Systématique et Evolution, Muséum National d'Histoire Naturelle, Paris, France.
* email: Vincent.Laudet@ens-lyon.fr

The evolutionary origin of metabolic pathways is a major issue that has mainly been adressed at a theoretical level. If we assume that metabolic pathways have been inherited by descent with modifications, then it should be possible to compare their structure in the same way that we compare anatomical structures, using cladistic analysis. We thus coded the enzyme similarities shared by different pathways using standard parsimony analyses, which does not use sequence comparisons. Here, we reconstructed the evolutionary history of metazoan steroidogenesis, a pathway whose origins are still elusive. This method allows the relative and absolute dating of various steps in the evolution of a metabolic pathway. Our analysis reveals that the cholesterol side-chain cleavage enzyme activity, today carried by CYP11A, is a unique vertebrate feature that appeared between – 643 Mya and – 500 Mya. We thus are able to propose a predicted structure for the ancestral chordate steroid. We also show that metazoan steroidogenesis, classically considered as a complex set of anabolic pathways, evolved in fact similarly to a catabolic pathway. We therefore suggest that animal steroids are domesticated cholesterol metabolites.

Keywords: steroids, metabolic pathways, biochemical evolution, metazoans

Introduction

Understanding animal evolution requires that we decipher the timing of the appearance and modification of morpho-anatomical or genomic characters, but also that we link these changes to modifications of physiological processes. Due to their peculiar tetracyclic structure, sterols are important modulators of cell membrane fluidity and flexibility in all eucaryotes, and they also regulate cellular oxygen levels [1]. In animals, cholesterol is also a precursor for the synthesis of

steroid hormones, that are key regulators of reproduction, development and homeostasis. The origin of animal steroid hormone signaling is debated. Several unique origins have been proposed, at different nodes of the metazoan tree: eumetazoan, bilaterian, chordate or vertebrate [2]. Alternatively, based on genomic analysis, it was proposed recently that steroid hormone signaling has evolved independently several times in various metazoan phyla from a common sterol-sensing background [3]. This has led to the hypothesis that steroid hormone synthesis may be homologous to the first step of xenobiotic detoxification [4]. However, protein comparisons take just a part of the available information on the synthesis pathways into account, because they are solely able to cope with biochemical reactions for which both the substrate and the product, but also the enzymes involved are known. Currently, in some vertebrates [5-7], arthropods [8] and nematodes [9-10], there are a number of steps in various steroidogenic pathways where the substrates and the products of a given reaction are known, but not the enzyme. The evolutionary signal of these incomplete data can be analysed through standard parsimony (cladistic) methods. This approach assumes that metabolic pathways are the result of an evolutionary process of descent with modification and that their evolutionary history can be inferred using parsimony analysis. In this approach, biochemical pathways are considered as taxa, and the individual reactions (or types of reactions) catalyzed by identified enzymes shared by pathways are the compared character states (Figure 1A-C). This type of analysis has been previously used with success to trace back the origin of the universal metabolism [11-13].

Here, we addressed the question of the relationships between synthesis pathways of vertebrate sex and adrenal steroids, bile acids, oxysterols and vitamin D, nematode dafachronic acids and arthropod ecdysteroids. We tested whether phylogenetic reconstruction by « comparative anatomy » of these pathways is able to order events of biochemical evolution of steroids and cholesterol. Except for vitamin D synthesis, that starts from 7-dehydrocholesterol, all these pathways start from the same precursor – cholesterol, a major metabolic hub - and end with a final product which is a ligand for a nuclear receptor, being in this way implicated in steroid cell signalling (Figure 1A-B). Resolving the main steps in the origins and diversification of steroids will also be important to better understand the evolution of the liganded Nuclear Receptors, that are major endocrinological players [14].

Using this approach, we are able to root steroid evolution in an absolute temporal framework. We also show that vertebrate sex and adrenal steroid synthesis pathways share a specific biochemical reaction type, and that, quite surprisingly, steroid synthesis pathways display the pattern of a catabolic pathway.

Results and Discussion

In order to elucidate the relative timing of the appearance of different metazoan steroidogenesis pathways, we reconstructed their relationships by comparing the enzymatic reactions that are involved in each of them. In our data set, each pathway from cholesterol to a steroid is a taxon (e.g. cortisol or chenodeoxicholic acid, Figure 1A) and enzymatic reactions along this pathway are the characters that are coded for each taxon and compared between the various taxa in a data matrix (Figure 1A-C, see also the complete data matrix in Figure S1). We thus constructed a data matrix in which the various taxons are systematically compared in terms of enzymes and enzymatic functions controlling the production of each ligand (Figure 1C, see also Methods in [13]). Characters are defined according to four criteria of homologies: type I homologies, and three subtypes of type II homologies (Figure 1E-H). Type I homologies are defined when two reactions in different metabolic pathways share the same enzyme with the same specificity for a substrate and same product (e. g. side-chain cleavage at carbon 20 in the synthesis of cortisol by two different pathways, Figure 1D-E). Type II homologies are cases in which pathways share enzymatic functions without sharing the specificity for a substrate in a more (Figure 1F) or less (Figure 1G) relaxed way (see also Methods section), or when they share an enzyme able to add the same residue at different positions on the carbon skeleton (Figure 1H).

Side-chain cleavage is the unique synapomorphy of vertebrate sex and adrenal steroidogenic pathways

A consensus tree was computed from the data matrix, which is shown in Figure 2. It groups together the synthesis of vertebrate bile acids with the synthesis of vertebrate sex and adrenal steroids, the synthesis of ecdysteroids, vitamin D (calcitriol), oxysterol and dafachronic acids being sister groups. The unique synapomorphy that unites vertebrate-type steroids is the side-chain cleavage of cholesterol on carbon 20 (Figure 1E; Figure 2, node A, character 31). Within vertebrate sex and adrenal steroids, steroids from various organisms are mixed, and even the synthesis pathway leading to the same molecules do not group together, suggesting that some enzymatic activities arose convergently in different pathways. The 5α-reduction of testosterone in the synthesis of dihydrotestosterone (DHT1-8) seems to have been recruited four times independently (Figure 2, character 130 at nodes F, G, H, I), and this is also the case for steroid aromatization (Figure 2, character 137). This may indicate that enzymes able to perform 5α-reduction and aromatization were already present before the apparition of enzymes performing side-chain cleavage. At that time they were likely to be acting on non-cleaved steroids and they were later recruited in the synthesis of steroids with a cleaved side-chain. This is consistent with the biochemical observation that these reactions can occur *in vitro* in amphioxus, in contrast to side-chain cleavage [4,15]. Thus we hypothesize that the ancestral chordate steroid was a molecule with a side-chain, an aromatic A-ring

and/or 3β and 17β-oxidated and 5α-reduced residues (Figure 2, node A). This would confirm that the reported binding ability of estrogens by a reconstructed ancestral bilaterian or chordate sex steroid receptor [16,17], is a by-product of the binding of an other steroid molecule, whose structure is still elusive for the moment [18]. This is also consistent with genomic analyses showing, outside vertebrates, the absence of an orthologue of side-chain cleavage enzymes [3].

Absolute dating of biochemical evolutionary events

Using the tree topology, we used synapomorphies related to families of reactions to assign different time periods to the tree branches (Figure 2, see also [13]). Our approach assumes that enzyme specificity has evolved from low-specificity proteins catalyzing a whole range of activities at low levels, to enzyme subfamilies with potent and highly specialized activities [19-20]. If this assumption is correct, then the putative common ancestry of pathways can be postulated not only on the basis of shared enzymes with high specificities (such as reaction 31 at node A), but also on the basis of very similar reactions (type II homologies). Thus if a branch of the tree is followed by downstream branches that do not bear changes in the type II homologies, the downstream branches are of the same period. When a new type II homology occurs on a branch this defines the next period (e.g. node B, Figure 2, see also Methods). According to this view, type II homologies distinguish between specific time periods during which innovations - in terms of enzymatic involvement in steroidogenesis - have occured.

To date, cladistic analysis of biochemical pathways has been carried out on the universal metabolism [13], which is common to all living organisms. By contrast, steroid metabolism is highly divergent among metazoans. This presents a unique way to order the various synthesis pathways that diverged since the Ediacaran [21]. Indeed, the availability of data regarding this variability in several metazoan phyla has made it possible to adopt the method developed on universal metabolism by managing both newly discovered interrelationships of steroidogenesis pathways and previously known interrelationships and divergence times of animal taxa. Here we can identify tree precise chronological boundaries. The first bilaterian steroid synthesis pathways here appear after the divergence between protostomes and deuterostomes (Figure 2, node E), 643 Mya ago [21]. After that, the presence of a clade of vertebrate side-chain steroid synthesis pathways containing gnathostome sex and adrenal steroids synthesis pathways and lamprey sex steroids synthesis pathways (Figure 2, node A) indicates that the last common ancestor of gnathostomes and lampreys, which lived between 500 and 475 Mya [22], was already able to synthesize at least one sex and/or adrenal steroid, not necessarily identical to the present day molecules. Finally, the elasmobranch-specific ability to make 1α-hydroxysteroid appeared after the divergence between chondrichtyians and osteichytians, 423 Mya ago [23]. We thus conclude that our method allows to put the appearance of new enzymatic activities

and new synthesis pathways in an absolute temporal framework.

Steroid synthesis has a catabolic evolutionary pattern

The definitions of periods allowed us to decipher the order of appearance of the different steroidogenic pathways (Figure 3, see also Figure S2 and S3). Starting from the cholesterol synthesis backbone (Figure 3, period 1), the first pathway to appear was the synthesis of ecdysteroids (Figure 3, period 2), followed by oxysterols (Figure 3, period 3). Then came the synthesis of vitamin D3 and Δ7-dafachronic acid (Figure 3, period 4). Of note, the synthesis of Δ4-dafachronic acid (Figure 2, period 5) appeared later than Δ7-dafachronic acid (Figure 2, period 4), which is consistent with the fact that Δ7-dafachronic acid seems to be a more ancient ligand for the nuclear receptor DAF-12 than Δ4-dafachronic acid [24]. Within vertebrate sex and adrenal steroids, pregnenolone appeared first (Figure 3, period 5), followed by progesterone, its 15α-derivatives (15α-hydroxyprogesterone and 15,17-dihydroxyprogesterone) and 11-deoxycorticosterone (Figure 3, period 6). A later phase of diversification led to the synthesis of other major steroids, such as estrogens, androgens, cortisol and aldosterone (Figure 3, period 7), and the very last synthesis pathways to appear were specific to shark and teleost steroids and some bile acids (Figure 3, period 8). As the more upstream reactions in the pathways are the first to appear, steroidogenesis thus evolved in a forward direction.

Theories on biochemical pathway evolution predict that anabolic pathways evolve backwards [25] whereas catabolic pathways evolve forwards [26]. But such simple patterns may be blurred by late opportunistic connexions of different patterns, as presumed for the metabolism of some amino-acids [13]. Surprisingly, we observed that steroid synthesis appears to develop forward, with a « catabolic » pattern, because upstream compounds such as oxysterols (Figure 3, period 3) and progesterone (Figure 3, period 6) appear before the more downstream bile acids (Figure 3, periods 6 and 8) and sex steroids (Figure 3, periods 7 and 8) respectively. Thus steroid synthesis should rather be viewed as a cholesterol degradation pathway, as it has already been suggested, at least for bile acids [27]. This view is in accordance with the observation that in phylogenetical trees based on genomic data, steroidogenic enzymes are nested within detoxification enzymes [3]. Therefore, steroidogenesis could be a derivative of catabolic pathways that are implicated in xenobiotic detoxification. This is also consistent with the biochemical definition of catabolism as the oxidative breakdown of organic molecules that releases energy. As a current working hypothesis we thus propose (Figure 4) that steroids were primarily metabolites of cholesterol degradation, and that some of them were recruited as a ligand for a sensor, which is a receptor able to bind various molecules with micromolar affinity and low specificity. This would have allowed cells to sense their environment and to regulate the expression of their metabolic machinery depending on the nutritional conditions. In a second phase, some receptors gained increasing affinity and specificity to a given

metabolite. Thus these « domesticated » metabolites may have secondarily acquired a more integrated hormonal function, by coupling nutritional status with reproductive cycle [28].

In conclusion, our work has allowed, by using an original way of reconstructing the evolution of steroids, to trace back the origins of a major component of a signaling pathway. We are convinced that the cladistic analysis of metabolic pathways could shed light on other open questions regarding the evolution of signaling molecules and will thus complement the concepts based mainly on the study of hormone receptors.

Methods

The general aim of the method employed in this paper is to determine interrelationships among biochemical pathways, each of them being defined by a starting point and a ending point (e.g. synthesis of estradiol from cholesterol). Such pathways are considered as taxa, and the characters are the enzymatic reactions (or the enzyme used to perform it) catalysing each step of the pathway.

Taxonomic Sampling

The present work focuses on the metabolic evolution pathways leading to the synthesis of vertebrate sex and adrenal steroids, bile acids, oxysterols and vitamin D, nematode dafachronic acids and arthropod ecdysteroids that are all end product ligands of nuclear hormone receptors. For example (Figure 1A-C), F1 and F2 are the sets of enzymatic activities involved in the synthesis of cortisol from cholesterol in two different ways. All these molecules can be synthesized from a common precursor, which is cholesterol, except from vitamin D synthesis that starts from 7-deoxycholesterol, and except from outgroups, that are the synthesis pathways from squalene to ergosterol, cholesterol and sitosterol. A complete taxon listed is given in Figure S1*A*.

Characters and Homologies

A complete list of characters is given in Figure S1*B*. They are defined according to four homology criteria, as ilustrated in Figure1.

Type hI homologies

Primary homologies of « type hI » were defined as sharing the same reaction, with absolute specificity for the substrate, for two or more different pathways. For example, the cholesterol side-chain cleavage by CYP11A1 (coded by character 31) is used by many pathways, which include the pathway transforming cholesterol into cortisol through a progesterone intermediate (Figure 1A) and the one transforming cholesterol into cortisol through a 17α-hydroxypregnenolone intermediate (Figure 1B). The CYP11A1 enzyme is shared by these two pathways without difference in specificity

for its substrate (character 31 on Figure 1E). In other words, the specificity is taken into account when defining type I homologies.

Type hIIa homologies

Primary homologies of « type hIIa » occur when two taxa share similar enzymatic function without considering their respective specificity for a substrate. This is for example the case for 3β-hydroxy-Δ5-steroid dehydrogenation on two steroids with a cleaved side chain, here pregnenolone or 17α-hydroxypregnenolone that illustrate a case of type hIIa homologies shared by taxons F1 and F2 among others (character 127 on Figure 1F).

Type hIIc homologies

Primary homologies of « type hIIc » are for shared functional family of enzymatic reaction, for example 3β-hydroxy-Δ5-steroid dehydrogenations on a steroid with or without a lateral side chain on the sterol skeleton (character 150 on Figure 1G).

Type hIIe homologies

Type « hIIb » and type « hIId » homologies as previously described [13] are not relevant here, but we defined a fifth type of secondary homology (type « hIIe ») for the sharing of the same enzyme which performs two slightly different reactions. Here the case appears for character 151, which describes the fact that the CYP27A1 enzyme is able to catalyse the hydroxylation of either carbon 25 or carbon 26 on the side chain of cholesterol during the synthesis of bile acids and calcitriol (Figure 1H). There is a risk associated with this type of homology: the enzyme could have been recruited by one of the two pathways. However the above reaction is situated very early (first step of cholesterol hydroxylation), favouring the hypothesis of an initial versatility of the enzyme.

The majority of characters corresponds to a single enzymatic reaction and a single enzyme. However, some characters correspond to multiple enzymes. This is the case for bile acid ligation and oxydation, which is a highly conserved process in which all the steps occur exactly in the same order for all bile acid precursors (reactions 93, 100, 110 and 116 on Figure 3, see also Figure S1*B*). So coding these steps separately would have lead to overwheighting this part of the matrix.

A number of characters referring to homologies of type II contain question marks. These are included when the taxon-pathway does not exhibit the appropriate substrate for the reaction, or when coding the reaction is meaningless for the taxon. For example, coding the possibility of a reduction from a ketone in ß on carbon 17 (reaction 129) has a meaning only for pathways where there is, at one step or another, a ketone on this carbon, and is meaningless for pathways were carbon 17 is the junction point of the side chain (for example in synthesis of ecdysone or dafachronic acids).

Phylogenetic Reconstruction

Tree Search

The matrix contains 71 taxa and 151 characters (Figure S1*C*). Characters were treated as unordered and unweighted in the search of the most parsimonious tree. Heuristic searches were conducted with TNT [29], using TBR branch swapping. CI=M/S, where M is the minimum number of character changes on the most parsimonious phylogeny possible (here 151), and S is the actual number required on the given phylogeny (here 240) [30].

Rooting

As outgroups we used the synthesis pathways from squalene to cholesterol, ergosterol and sitosterol, assuming that, due to the important role played by sterol molecules in the stability of eukaryotic membranes, the synthesis of sterol precursors by various eukaryotes is more ancient that the synthesis of steroid hormones from those precursors in animals.

Defining Time Spans Criteria

From the root to the tip of branches, phylogenetic trees provide a relative order of transformations (here enzymatic innovations) through time. We call an « upstream node » a deep, more inclusive internal branch or clade, and a « downstream node » a more terminal, less inclusive internal branch or clade. Time spans in metabolism are defined as the time along the tree separating two type II character changes. The use of this criterion is empirical and based on the supposition that the apparition of type II homologies correlates with the apparition of new kinds of enzymatic reactions catalysed by the same enzyme. To define periods the following criteria must be taken into account:

The order of nodes. In the absence of other criteria, two sister-nodes are of the same relative period.

The nature of enzymatic changes. If a branch is followed by downstream branches that do not bear changes in the type II homologies, the downstream branches are of the same period. When a new type II homology occurs on a branch, it defines the next period, except when it is already present in an earlier branch (homoplasy). For example, the node B is purple (Figure 2, period 6) due to character 137. This character also appears at node C, where there is no period change in comparison with the upstream node D because the period of node D (period 7) is already posterior to the period when character 137 appeared. Losses of type II characters were not used to define periods because this loss can be the result of very different processes (mutations, changes in cellular localisation, in expression regulations...).

No homoplasy. When a character exhibits homoplasy, only transformations with unambiguous localisations were taken into account in the use of the second criterion. For example, character 130 appears independently in synthesis of Δ7-dafachronic acid in nematodes and in synthesis of dehydrotestosterone in vertebrates. Due to the very high divergence time between these two lineages and the absence of this reaction in other taxa, we hypothesize that these reactions appeared convergently and can be used in the two taxa to define a new period. Similarly, character 147, that appears with a very complicated pattern, is used to infer period change only at the level of its first

occurrence at node E. The same reasoning was carried for character 129.

Highly specific type I homologies. In one case (character 31), a single enzyme with high specificity innovates new enzymatic mechanisms (the cleavage of the cholesterol side-chain). It is recorded as a type I homology, but as it also corresponds to enzymatic changes used to define type II homologies, it is taken into account for defining periods.

Acknowledgements

We thank Michael Baker, François Bonneton, Joanne Burden and Chantal Dauphin-Villemant for their valuable comments on this manuscript and David Russel for checking the bile acid pathways.

Funding

This work was supported by MENRT, the Cascade European Network of Excellence (FP6), and the integrated project Crescendo (FP6), ENS Lyon, CNRS and FRM. We also thank the Company Saint-Gobain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

The authors have made the following declarations about their contributions: Conceived and designed the experiments: GVM GL VL. Performed the experiments: GVM. Analyzed the data: GVM GL VL. Wrote the paper: GVM GL VL.

References

1. Brown AJ, Galea AM (2010) Cholesterol as an evolutionary response to living with oxygen. Evolution 64:2179-2183.
2. Lafont R, Mathieu M (2007) Steroids in aquatic invertebrates. Ecotoxicology 16: 109-30.
3. Markov GV, Tavares R, Dauphin-Villemant C, Demeneix BA, Baker ME, et al. (2009) Independent elaboration of steroid hormone signaling pathways in metazoans. Proc Natl Acad Sci U S A 106: 11913-8.
4. Markov GV, Laudet V (2011) Origin and evolution of the ligand-binding ability of nuclear receptors. Mol Cell Endocrinol 344: 21-30.
5. Payne AH, Hales DB (2004) Overview of steroidogenic enzymes in the pathway from cholesterol to active steroid hormones. Endocr Rev 25: 947-970.
6. Bury NR, Sturm A (2007) Evolution of the corticosteroid receptor signalling pathway in fish. Gen Comp Endocrinol 153: 47-56.
7. Lowartz S, Petkam R, Renaud R, Beamish FWH, Kime DE, et al. (2003) Blood steroid profile and in vitro steroidogenesis by ovarian follicles and testis fragments of adult sea lamprey, *Petromyzon marinus*. Comp

Biochem Physiol A Mol Integr Physiol 134: 365-376.

8. Huang X, Warren JT, Gilbert LI (2008) New players in the regulation of ecdysone biosynthesis. J Genet Genomics 35: 1-10.

9. Motola DL, Cummins CL, Rottiers V, Sharma KK, Li T, et al. (2006) Identification of ligands for DAF-12 that govern dauer formation and reproduction in *C. elegans*. Cell 124: 1209-1223.

10. Patel DS, Fang LL, Svy DK, Ruvkun G, Li W (2008) Genetic identification of HSD-1, a conserved steroidogenic enzyme that directs larval development in *Caenorhabditis elegans*. Development 135: 2239-2249.

11. Cunchillos C, Lecointre G (2002) Early steps of metabolism evolution inferred by cladistic analysis of amnio acid catabolic pathways. C R Biologies 325: 119-129.

12. Cunchillos C, Lecointre G (2005) Integrating the universal metabolism into a phylogenetic analysis. Mol Biol Evol 22: 1-11.

13. Cunchillos C, Lecointre G (2007) Ordering events of biochemical evolution. Biochimie 89: 555-573.

14. Bridgham JT, Eick GN, Larroux C, Deshpande K, Harms MJ, et al. (2010) Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. PLoS Biol 8.

15. Mizuta T, Asahina K, Suzuki M, Kubokawa K (2008) In vitro conversion of sex steroids and expression of sex steroidogenic enzyme genes in amphioxus ovary. J Exp Zool 309A: 83-93.

16. Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301: 1714-1717.

17. Bridgham JT, Brown JE, Rodríguez-Marí A, Catchen JM, Thornton JW (2008) Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. PLoS Genet. 4, e1000191.

18. Eick GN, Thornton JW (2011) Evolution of steroid receptors from an estrogen-sensitive ancestral receptor. Mol Cell Endocrinol 334: 31-38.

19. Jensen RA (1976) Enzyme recruitment in evolution of new function. Annual review of microbiology 30: 409-425.

20. Khersonsky O, Roodveldt C, Tawfik DS (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol 10: 498-508.

21. Peterson KJ, Cotton JA, Gehling JG, Pisani D (2008) The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos Trans R Soc Lond B Biol Sci 363: 1435-1443.

22. Janvier P (2008) Primitive fishes and fishes from Deep Time. In: McKenzie DJ, Farrell AP, Brauner CJ, editors. Primitive Fishes. Fish Physiology 26, Academic Press. pp.1-51.

23. Zhu M, Zhao W, Jia L, Lu J, Qiao T, et al. (2009). The oldest articulated osteichthyan reveals mosaic gnathostome characters. Nature 458: 469-474.

24. Ogawa A, Streit A, Antebi A, Sommer RJ (2009) A conserved endocrine mechanism controls the formation of dauer and infective larvae in nematodes. Curr Biol 19: 67-71.

25. Horowitz NH (1945) On the evolution of biochemical syntheses. Proc Natl Acad Sci U S A 31: 153-157.

26. Cordon F (1990) *Tratado Evolucionista De Biologia.* Aguilar, Madrid.

27. Russell DW (2009) Fifty years of advances in bile acid synthesis and metabolism. J Lipid Res 50 Suppl: S120-S125.

28. Della Torre S, Rando G, Meda C, Stell A, Chambon P et al. (2011) Amino Acid-Dependent Activation of Liver Estrogen Receptor Alpha Integrates Metabolic and Reproductive Functions via IGF-1. Cell Metab 13: 205-214.

29. Goloboff P (1999) Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15:

415-428.

30. Farris JS (1989) The retention index and the rescaled consistency index. Cladistics 5: 417-419.

Figure legends

Figure 1. Methodological principles followed in character coding.

(A) Extract from the total steroidogenic pathways, showing three of the four pathway taxa studied in this example. The pathway starting and ending points are boxed. F1 taxon, synthesis of cortisol from cholesterol through progesterone is highlighted in blue. The two other taxa are CALCITRIOL, synthesis of 1,25-dihydroxyvitamin D3 from 7-dehydrocholesterol, and CHENOAC2, one of the pathways from cholesterol to chenodeoxycholic acid. For these two taxa the coding is given in panel C. (B) Extract from the pathway showed in panel A, highlighting in orange the fourth pathway, the F2 taxon, which represents synthesis of cortisol from cholesterol through 17α-hydroxypregnenolone. (C) Extract from the data matrix, showing the coding for the four taxa. The colours of the columns refer to the four types of homologies that are explained in panels E to H. (D) Carbon numbering of the sterol skeleton. Panels E to H show the character coding and the various types of homology used in our analysis. (E) Reaction 31 (side-chain cleavage on cholesterol, highlighted in purple) occurs both in taxon F1 in taxon F2. This is a type I homology (hI), coded by a 1 in the 31st column of the data matrix for both taxa (see panel C). (F) Both pregnenolone in taxon F1 and 17α-hydroxypregnenolone in taxon F2 undergo a 3β-hydroxy-Δ5-steroid dehydrogenation (highlighted in blue), but the two substrate molecules are different, because there an hydroxyl group on carbon 17 for 17α-hydroxypregnenolone and a hydrogen atom at this position for pregnenolone: this is a type IIa homology (hIIa). (G) Pregnenolone in taxon F1 and 3β,7α-dihydroxy-5-cholestenoic acid in taxon CHENOAC2 both undergo a 3β-hydroxy-Δ5-steroid dehydrogenation (highlighted in green). They differ more than pregnenolone and 17α-hydroxypregnenolone, because additional to the differences in the presence of an hydroxyl group on the carbon residue, pregnenolone has no side-chain whereas 3β,7α-dihydroxy-5-cholestenoic acid has one. This more relaxed similarity reflects a type IIc homology (hIIc). (H) Cholesterol in taxon CHENOAC2 and vitamin D3 in taxon CALCITRIOL both have their side-chain hydroxylated by the CYP27A1 enzyme, but on two different carbons: C25 and C26 (highlighted in red). This is a type IIe homology (hIIe).

Figure 2. Interrelationships among animal steroidogenic pathways obtained from standard parsimony approach.

The zoological groups the various biochemical taxa belong to are mapped on a 65%-majority-rule

consensus tree. We obtained a consensus of 25 equiparsimonious trees, each of 240 steps. Each tree has a CI of 0.63, which indicates a rather low level of homoplasy, with regard to the number of taxa. Red dots indicate the nodes for which a minimum age for the split between two animal groups can be inferred. The inferred age is indicated in the scale at the bottom of the figure. Type II homologies are mapped on the tree, and used for the definition of time periods, for which the colour code is explained in the caption box at the bottom. For example, the node A is purple (period 6) due to character 137. This character also appears at node C, where there is no period change in comparison with the upstream node B because the period of node B (period 7) is already posterior to the period when character 137 appeared. Losses of type II characters (in red) were not used to define periods because these losses can be the results of very different processes, such as mutations, changes in cellular localisation or in expression regulation. A proposed ancestral chordate steroid is shown at node L. Other nodes are discussed in text.

Figure 3. A tentative chronology about the appearance of animal steroid metabolic pathways. Colours indicate successive time spans (or periods) as inferred from the tree on Figure 2. Starting and ending points of metabolic pathways are boxed. For those ending points that are nuclear receptor ligands, the official and trivial name of their main receptor is indicated in black around the box. Reaction numbers refer to enzymatic reactions that are described in Figure S1*B,* and a fully detailed version of the pathways is provided in Figure S2.

Figure 4. A model about the origin of steroid hormone signaling through nuclear receptors. We propose that the enzymatic machinery implicated in xenobiotic oxidation was also used for cholesterol degradation, and that cholesterol and its metabolites became ligands for an ancestral nuclear receptor sensor, binding to it with micromolar affinity (thin arrow) and low specificity. Following gene duplication events, some receptors gained higher affinity in the nanomolar range (thick arrow) and higher specificity for some of these metabolites, thus becoming steroid hormones, with new physiological properties.

Supporting Information

Figure S1: Data matrix containing 71 taxons (rows) and 151 characters (columns) ***A.*** definition of the compared taxons. As almost all taxa are enzymes synthesized from cholesterol, taxons names such as « P4 » mean « synthesis of progesterone from cholesterol). The only four exceptions are sqERG, sqC and sqSIT that represent synthesis of ergosterol, cholesterol and sitosterol from squalene, and CALCITRIOL, which represents synthesis of vitamin D3 from 7dehydrocholesterol. When there

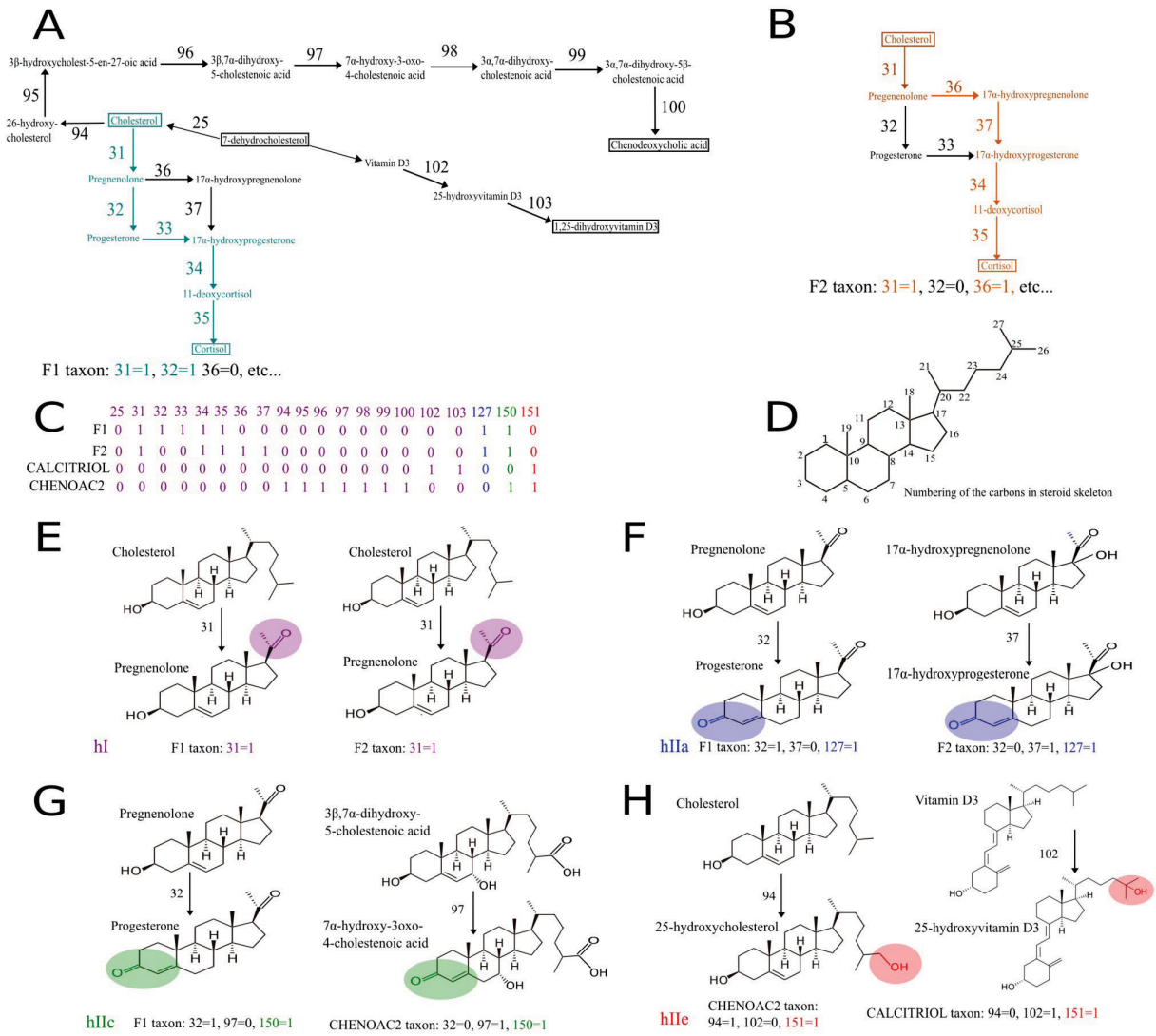
are several manners to synthesize a product, the indication « via X» indicates reaction numbers (see Figure 3) that are specific to each pathway. Thus it is possible to discriminate between various possibilities. Each taxon corresponds to one of the lines of the data matrix shown in *C*. ***B.*** character names, with the corresponding number of international nomenclature, when possible, and homology types defined in the text (see Figure 1). Each character corresponds to one of the column of the data matrix shown in *C*. Characters are colored according to the following convention: type I homologies (hI), type IIa homologies (hIIa), type IIc homologies (hIIc) and type IIe homology (hIIe). Names of compounds that are shortened on Figure 3 are here given in full, with the abbreviation between brackets. ***C.*** matrix. Each line is a taxon, i.e. a given pathway, as detailed in *A*. Each column is a character, i.e. either a "type I" homology (when an enzyme is shared by several pathways with the same specificity for its substrate in them) or a "type II" homology (when a type of reaction is performed in several pathways without considering specificity), as detailed in *B*. Dots « . » and « 1 » refer to the character state found in the corresponding taxon. « 1 » indicates the presence of the character and « . » its absence. Question marks are assigned to character states when the enzyme or the enzymatic function is not applicable to the taxon: the required substrate is not available in this pathway. Characters (columns) are numbered following their order from the left to right: the character number one is the first column; the character number 36 is the 36th column. The matrix columns are ordered as follows: 116 type I homologies (hI), 29 type IIa homologies (hIIa), 5 type IIc homologies (hIIc) and one type IIe homology (hIIe).

Figure S2. A detailed view of animal steroid metabolic pathways.

This figure is a more detailed version of Figure 3. Colours indicate successive time spans (or periods) as inferred from the tree on Figure 2. Metabolic pathways are drawn, all intermediates being shown, with reaction numbers referring to the enzymatic reactions that are described in Figure S1*B*. Starting and ending points are boxed.

Figure S3. Sequential appearance of animal steroid metabolic pathways

This is a simplified version of the pathways presented in Figure 2, without any taxon name and any reaction number. The eight numbered boxes show the eight steps in appearance of the various pathways. Thus, each box corresponds to a defined period, with the same colour code as in Figure 2 and 3 as well as Figure S2.


A
3β-hydroxycholest-5-en-27-oic acid
96
3β,7α-dihydroxy-5-cholestenoic acid
97
7α-hydroxy-3-oxo-4-cholestenoic acid
98
3α,7α-dihydroxy-cholestenoic acid
99
3α,7α-dihydroxy-5β-cholestenoic acid
100
Chenodeoxycholic acid
95
26-hydroxy-cholesterol
94
Cholesterol
25
7-dehydrocholesterol
Vitamin D3
102
25-hydroxyvitamin D3
103
1,25-dihydroxyvitamin D3
31
Pregnenolone
36
17α-hydroxypregnenolone
32
Progesterone
33
17α-hydroxyprogesterone
37
34
11-deoxycortisol
35
Cortisol
F1 taxon: 31=1, 32=1 36=0, etc...
B
Cholesterol
31
Pregenenolone
36
17α-hydroxypregnenolone
32
Progesterone
33
17α-hydroxyprogesterone
37
34
11-deoxycortisol
35
Cortisol
F2 taxon: 31=1, 32=0, 36=1, etc...
C
25 31 32 33 34 35 36 37 94 95 96 97 98 99 100 102 103 127 150 151
F1 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0
F2 0 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1 0
CALCITRIOL 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1
CHENOAC2 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 1 1
D
Numbering of the carbons in steroid skeleton
E
Cholesterol
31
Pregnenolone
hI
F1 taxon: 31=1
Cholesterol
31
Pregnenolone
F2 taxon: 31=1
F
Pregnenolone
32
Progesterone
hIIa
F1 taxon: 32=1, 37=0, 127=1
17α-hydroxypregnenolone
37
17α-hydroxyprogesterone
F2 taxon: 32=0, 37=1, 127=1
G
Pregnenolone
32
Progesterone
hIIc
F1 taxon: 32=1, 97=0, 150=1
3β,7α-dihydroxy-5-cholestenoic acid
97
7α-hydroxy-3oxo-4-cholestenoic acid
CHENOAC2 taxon: 32=0, 97=1, 150=1
H
Cholesterol
94
25-hydroxycholesterol
hIIe
CHENOAC2 taxon: 94=1, 102=0, 151=1
Vitamin D3
102
25-hydroxyvitamin D3
CALCITRIOL taxon: 94=0, 102=1, 151=1

470 Fig. 1

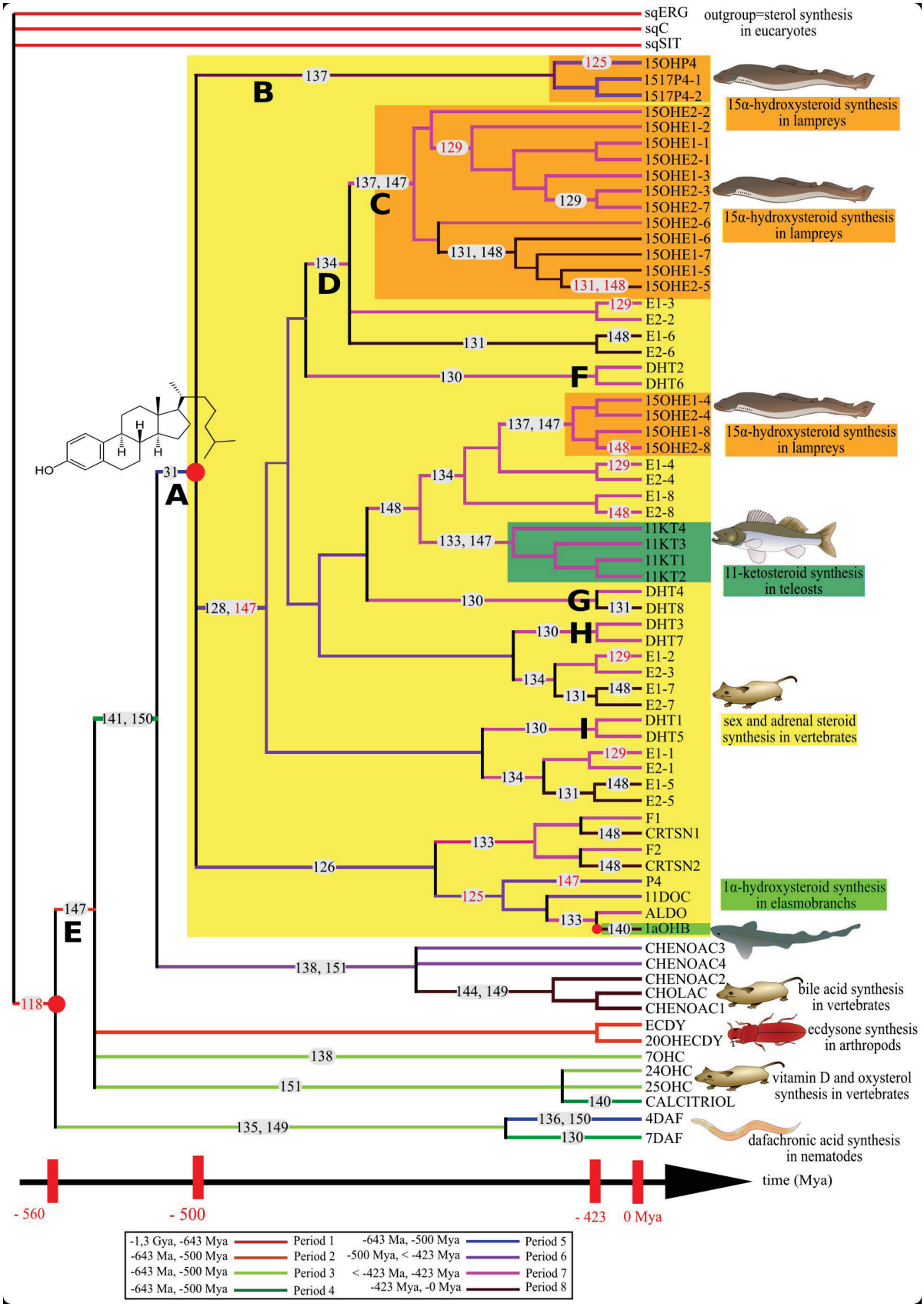

sqERG
sqC
sqSIT
outgroup=sterol synthesis in eucaryotes
15OHP4
1517P4-1
1517P4-2
15α-hydroxysteroid synthesis in lampreys
15OHE2-2
15OHE1-2
15OHE1-1
15OHE2-1
15OHE1-3
15OHE2-3
15OHE2-7
15OHE2-6
15OHE1-6
15OHE1-7
15OHE1-5
15OHE2-5
15α-hydroxysteroid synthesis in lampreys
E1-3
E2-2
E1-6
E2-6
DHT2
DHT6
15OHE1-4
15OHE2-4
15OHE1-8
15OHE2-8
15α-hydroxysteroid synthesis in lampreys
E1-4
E2-4
E1-8
E2-8
11KT4
11KT3
11KT1
11KT2
11-ketosteroid synthesis in teleosts
DHT4
DHT8
DHT3
DHT7
E1-2
E2-3
E1-7
E2-7
sex and adrenal steroid synthesis in vertebrates
DHT1
DHT5
E1-1
E2-1
E1-5
E2-5
F1
CRTSN1
F2
CRTSN2
P4
11DOC
ALDO
1aOHB
1α-hydroxysteroid synthesis in elasmobranchs
CHENOAC3
CHENOAC4
CHENOAC2
CHOLAC
CHENOAC1
bile acid synthesis in vertebrates
ECDY
20OHECDY
ecdysone synthesis in arthropods
7OHC
24OHC
25OHC
CALCITRIOL
vitamin D and oxysterol synthesis in vertebrates
4DAF
7DAF
dafachronic acid synthesis in nematodes
A
B
C
D
E
F
G
H
I
time (Mya)
- 560
- 500
- 423
0 Mya
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
-643 Ma, -500 Mya Period 5
-500 Mya, < -423 Mya Period 6
< -423 Ma, -423 Mya Period 7
-423 Mya, -0 Mya Period 8

Fig. 2

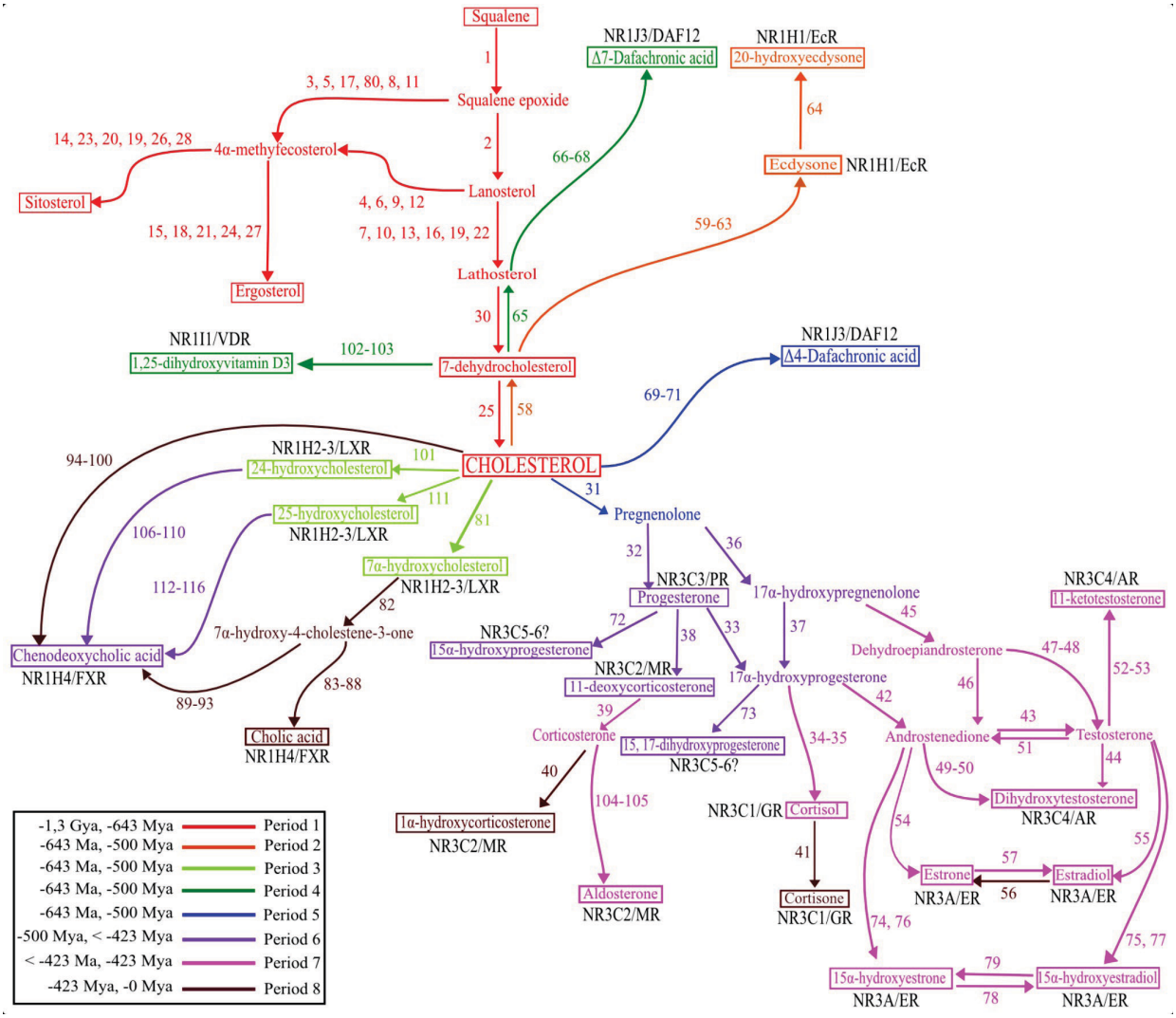

Squalene
1
Squalene epoxide
2
Lanosterol
Lathosterol
7-dehydrocholesterol
CHOLESTEROL
Pregnenolone
Progesterone
11-deoxycorticosterone
Corticosterone
1α-hydroxycorticosterone
NR3C2/MR
Aldosterone
17α-hydroxypregnenolone
17α-hydroxyprogesterone
15, 17-dihydroxyprogesterone
NR3C5-6?
15α-hydroxyprogesterone
Cortisol
NR3C1/GR
Cortisone
Androstenedione
Testosterone
Dihydroxytestosterone
NR3C4/AR
11-ketotestosterone
Dehydroepiandrosterone
Estrone
Estradiol
NR3A/ER
15α-hydroxyestrone
15α-hydroxyestradiol
4α-methylfecosterol
Sitosterol
Ergosterol
24-hydroxycholesterol
25-hydroxycholesterol
7α-hydroxycholesterol
7α-hydroxy-4-cholestene-3-one
Chenodeoxycholic acid
NR1H4/FXR
Cholic acid
NR1H2-3/LXR
NR1I1/VDR
1,25-dihydroxyvitamin D3
NR1J3/DAF12
Δ7-Dafachronic acid
Δ4-Dafachronic acid
NR1H1/EcR
20-hydroxyecdysone
Ecdysone
3, 5, 17, 80, 8, 11
14, 23, 20, 19, 26, 28
15, 18, 21, 24, 27
4, 6, 9, 12
7, 10, 13, 16, 19, 22
66-68
59-63
64
30
65
102-103
25
58
69-71
94-100
101
111
81
106-110
112-116
82
83-88
89-93
31
32
36
72
38
33
37
45
47-48
52-53
46
43
51
44
42
39
73
34-35
40
104-105
41
54
55
57
56
49-50
74, 76
75, 77
79
78
-1,3 Gya, -643 Mya
-643 Ma, -500 Mya
-643 Ma, -500 Mya
-643 Ma, -500 Mya
-643 Ma, -500 Mya
-500 Mya, < -423 Mya
< -423 Ma, -423 Mya
-423 Mya, -0 Mya
Period 1
Period 2
Period 3
Period 4
Period 5
Period 6
Period 7
Period 8

Fig. 3

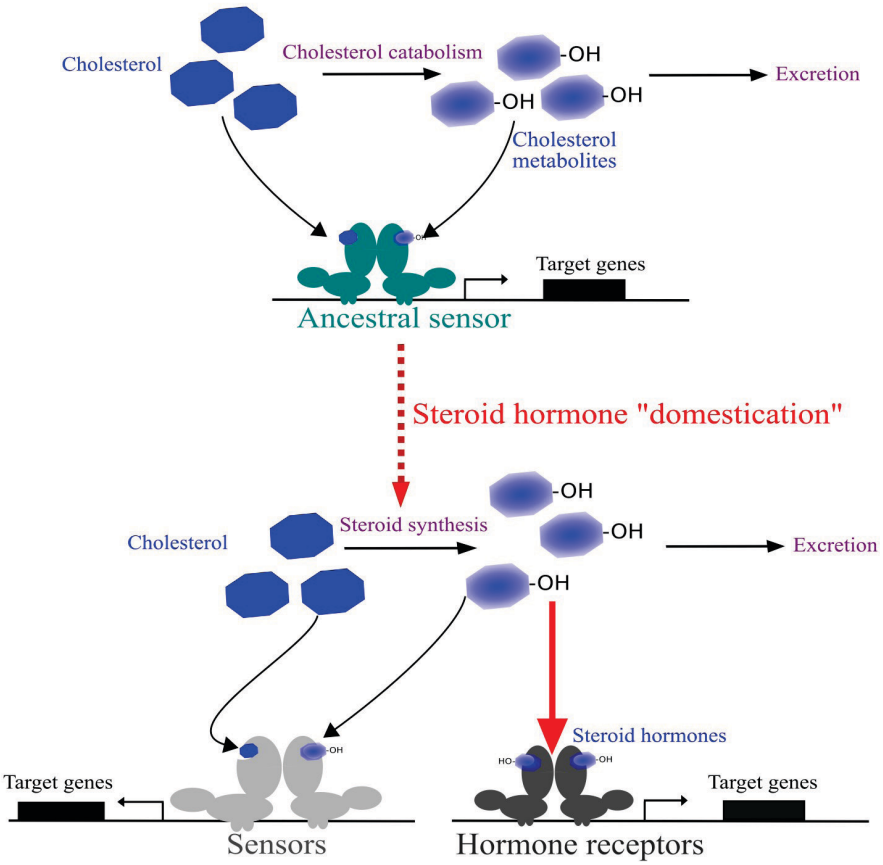


Fig. 4

Supplemental Figure S1

A

sqERG: synthesis of ergosterol from squalene
sqC: synthesis of cholesterol from squalene
sqSIT: synthesis of sitosterol from squalene
15OHP4: synthesis of 15α-hydroxyprogesterone
1517P4-1: synthesis of 15,17-hydroxyprogesterone via 32
1517P4-2: synthesis of 15,17-hydroxyprogesterone via 37
15OHE1-1: synthesis of 15α-hydroxyestrone via 32 and 76
15OHE1-2: synthesis of 15α-hydroxyestrone via 37 and 76
15OHE1-3: synthesis of 15α-hydroxyestrone via 46 and 76
15OHE1-4: synthesis of 15α-hydroxyestrone via 48 and 76
15OHE1-5: synthesis of 15α-hydroxyestrone via 32 and 77
15OHE1-6: synthesis of 15α-hydroxyestrone via 37 and 77
15OHE1-7: synthesis of 15α-hydroxyestrone via 46 and 77
15OHE1-8: synthesis of 15α-hydroxyestrone via 48 and 77
15OHE2-1: synthesis of 15α-hydroxyestradiol via 32 and 76
15OHE2-2: synthesis of 15α-hydroxyestradiol via 37 and 76
15OHE2-3: synthesis of 15α-hydroxyestradiol via 46 and 76
15OHE2-4: synthesis of 15α-hydroxyestradiol via 48 and 76
15OHE2-5: synthesis of 15α-hydroxyestradiol via 32 and 77
15OHE2-6: synthesis of 15α-hydroxyestradiol via 37 and 77

15OHE2-7: synthesis of 15α-hydroxyestradiol via 46 and 77
15OHE2-8: synthesis of 15α-hydroxyestradiol via 48 and 77
4DAF: synthesis of Δ4-dafachronic acid
7DAF: synthesis of Δ7-dafachronic acid
ECDY: synthesis of ecdysone
20OHECDY: synthesis of 20-hydroxyecdysone
F1: synthesis of cortisol via 32
F2: synthesis of cortisol via 37
ALDO: synthesis of aldosterone
11DOC: synthesis of 11-deoxycorticosterone
1aOHB: synthesis of 1α-hydroxycorticosterone
CRTSN1: synthesis of cortisone via 32
CRTSN2: synthesis of cortisone via 37
P4: synthesis of progesterone
DHT1: synthesis of dehydrotestosterone via 32 and 44
DHT2: synthesis of dehydrotestosterone via 37 and 44
DHT3: synthesis of dehydrotestosterone via 46 and 44
DHT4: synthesis of dehydrotestosterone via 47 and 48
DHT5: synthesis of dehydrotestosterone via 32 and 49
DHT6: synthesis of dehydrotestosterone via 37 and 49
DHT7: synthesis of dehydrotestosterone via 46 and 49
DHT8: synthesis of dehydrotestosterone via 48 and 49
11KT1: synthesis of 11-ketotestosterone via 32
11KT2: synthesis of 11-ketotestosterone via 37
11KT3: synthesis of 11-ketotestosterone via 46
11KT4: synthesis of 11-ketotestosterone via 48
E1-1: synthesis of estrone via 32 and 54
E1-2: synthesis of estrone via 37 and 54
E1-3: synthesis of estrone via 46 and 54
E1-4: synthesis of estrone via 48 and 54
E1-5: synthesis of estrone via 32 and 55
E1-6: synthesis of estrone via 37 and 55
E1-7: synthesis of estrone via 46 and 55
E1-8: synthesis of estrone via 48 and 55
E2-1: synthesis of estradiol via 32 and 54
E2-2: synthesis of estradiol via 37 and 54
E2-3: synthesis of estradiol via 46 and 54
E2-4: synthesis of estradiol via 48 and 54
E2-5: synthesis of estradiol via 32 and 55
E2-6: synthesis of estradiol via 37 and 55
E2-7: synthesis of estradiol via 46 and 55
E2-8: synthesis of estradiol via 48 and 55
CHOLAC: synthesis of cholic acid
CHENOAC1: synthesis of chenodeoxycholic acid via 81
7OHC: synthesis of 7α-hydroxycholesterol
24OHC: synthesis of 24-hydroxycholesterol
CALCITRIOL: synthesis of vitamin D3 from 7-dehydrocholesterol
25OHC: synthesis of 25-hydroxycholesterol
CHENOAC2: synthesis of chenodeoxycholic acid via 94
CHENOAC3: synthesis of chenodeoxycholic acid via 101
CHENOAC4: synthesis of chenodeoxycholic acid via 111

B

[1] *Squalene epoxidase:* 1.14.99.7; I.
[2] *Lanosterol synthase:* 5.4.99.7; I.
[3] *Cycloartenol synthase:* 5.4.99.8; I.
[4] *Sterol 24-C-methyltransferase (smt1) on lanosterol:* 2.1.1.41; I.
[5] *Sterol 24-C-methyltransferase (smt1) on cycloartenol:* 2.1.1.41; I.
[6] *Sterol 14-demethylase (CYP51) on lanosterol:* 1.14.13.70; I.
[7] *Sterol 14-demethylase (CYP51) on eburicol:* 1.14.13.70; I.
[8] *Sterol 14-demethylase (CYP51) on obtusifoliol:* 1.14.13.70 ; I.
[9] *Delta(14)-sterol reductase on 4,4-dimethyl-5α-ergosta-8,14,24(28)-trien-3β-ol:* 1.3.1.70; I.
[10] *Delta(14)-sterol reductase on 4,4-dimethyl-5α-cholesta-8,14,24-trien-3β-ol:* 1.3.1.70 ; I.
[11] *Delta(14)-sterol reductase on 4α-methyl-5α-ergosta-8,14,24(28)-trien-3β-ol:* 1.3.1.70; I.
[12] *Methylsterol monooxygenase, sterol-4-alpha-carboxylate 3-dehydrogenase, 3-keto-steroid reductase on 4,4-dimethylfecosterol:* 1.14.13.72, 1.1.1.170, 1.1.1.270; I.
[13] *Methylsterol monooxygenase, sterol-4-alpha-carboxylate 3-dehydrogenase, 3-keto-steroid reductase on 4,4-dimethyl-5α-cholesta-8,24(28)-dien-3β-ol:* 1.14.13.72, 1.1.1.170, 1.1.1.270; I.
[14] *Cholestenol Delta-isomerase on 4α-methylfecosterol:* 5.3.3.5; I.
[15] *Methylsterol monooxygenase, sterol-4α-carboxylate 3-dehydrogenase, 3-keto-steroid reductase on 4α-methylfecosterol:* 1.14.13.72, 1.1.1.170, 1.1.1.270; I.
[16] *Methylsterol monooxygenase, sterol-4α-carboxylate 3-dehydrogenase, 3-keto-steroid reductase on 4α-methylzymosterol:* 1.14.13.72, 1.1.1.170, 1.1.1.270; I.
[17] *Sterol-4α-methyl oxidase 1 (smo1); sterol-4α-carboxylate 3-dehydrogenase, 3-keto-steroid reductase on 4α-methylfecosterol on 24-methylene cycloartenol:* smo1, 1.1.1.170, 1.1.1.270; I.
[18] *C-8 sterol isomerase (erg2):* 5.-.-.-; I.
[19] *Cholestenol Delta-isomerase on zymosterol:* 5.3.3.5; I.
[20] *Sterol-4α-methyl oxidase 2 (smo2), sterol-4α-carboxylate 3-dehydrogenase, 3-keto-steroid reductase on 4α-methylfecosterol on 24-ethylenelophenol:* smo2, 1.1.1.170, 1.1.1.270; I.
[21] *C-5 sterol desaturase (erg 3):* 1.3.3.-; I.
[22] *Delta24-sterol reductase (DHCR24):* 1.3.1.72; I.
[23] *24-methylenesterol C-methyltransferase (smt2): 2.1.1.143*; I.
[24] *C-22 sterol desaturase (erg5):* 1.14.14.-; I.
[25] *7-dehydrocholesterol reductase (DHCR7) on 7-dehydrocholesterol:* 1.3.1.21; I.
[26] *7-dehydrocholesterol reductase (dwf5) on 24-methylene 5-dehydroepisterol:* 1.3.1.21; I.
[27] *Delta24(24(1))-sterol reductase (erg4) on 5,7,22,24(28)-ergostatetraenol:* 1.3.1.71; I.
[28] *Delta24-sterol reductase (dwf1) on isofucosterol:* 1.3.1.72; I.
[29] *Lathosterol oxidase (SC5DL) on 24-ethylenelathosterol:* 1.14.21.6; I.
[30] *Lathosterol oxidase (SC5DL) on lathosterol:* 1.14.21.6; I.
[31] *Cholesterol side-chain cleavage enzyme (CYP11A1):* 1.14.15.6; I.
[32] *3β-hydroxy-delta5-steroid dehydrogenase / steroid delta-isomerase (HSD3B) on pregnenolone:* 1.1.1.145 5.3.3.1 ; I.
[33] *Steroid 17α-monooxygenase (CYP17A1) on progesterone:* 1.14.99.9; I.
[34] *Steroid 21-monooxygenase (CYP21A1) on 17α-hydroxyprogesterone:* 1.14.99.10; I.
[35] *Steroid 11β-monooxygenase (CYP11B) on 11-deoxycortisol:* 1.14.15.4; I.
[36] *Steroid 17α-monooxygenase (CYP17A1) on pregnenolone:* 1.14.99.9; I.
[37] *3β-hydroxy-Δ5-steroid dehydrogenase / steroid delta-isomerase (HSD3B) on 17α-hydroxypregnenolone:* 1.1.1.145 5.3.3.1 ; I.
[38] *Steroid 21-monooxygenase (CYP21A1) on progesterone:* 1.14.99.10; I.
[39] *Steroid 11β-monooxygenase (CYP11B1) on 11-deoxycorticosterone:* 1.14.15.4; I.
[40] *Steroid 1α-monooxygenase on corticosterone:* 1.14.-.-; I.
[41] *11β-hydroxysteroid dehydrogenase (HSD11B2) on cortisol:* 1.1.1.146; I.

[42] *17α-hydroxyprogesterone aldolase (CYP17A1) on 17α-hydroxyprogesterone:* 4.1.2.30; I.
[43] *Hydroxysteroid (17-β) dehydrogenase 3 (HSD17B3) on androstenedione:* 1.1.1.64; I.
[44] *3-oxo-5α-steroid 4-dehydrogenase 1 (SRD5A) on testosterone:* 1.3.99.5; I.
[45] *17α-hydroxyprogesterone aldolase (CYP17A1) on 17α-hydroxypregenolone:* 4.1.2.30; I.
[46] *3β-hydroxy-Δ5-steroid dehydrogenase / steroid delta-isomerase (HSD3B) on dehydroepiandrosterone*; 1.1.1.145 5.3.3.1; I.
[47] *17β-hydroxysteroid dehydrogenase (HSD17B) on dehydroepiandrosterone:* 1.1.1.51; I.
[48] *3β-hydroxy-Δ5-steroid dehydrogenase / steroid delta-isomerase (HSD3B) on androst-5ene-3β,17β-diol:* 1.1.1.145 5.3.3.1; I.
[49] *3-oxo-5-α-steroid 4-dehydrogenase 1 (SRD5A) on androstenedione: 1.3.99.5*; I.
[50] *17-ketoreductase (HSD17B3) on androstanedione:* 1.1.1.64; I.
[51] *Hydroxysteroid (17-β) dehydrogenase 2 (HSD17B2) on testosterone:* 1.1.1.63; I.
[52] *Steroid 11β-monooxygenase (CYP11B) on testosterone:* 1.14.15.4; I.
[53] *11β-hydroxysteroid dehydrogenase on 11β-hydroxytestosterone:* 1.1.1.146; I.
[54] *Aromatase (CYP19) on androstenedione:* 1.14.14.1; I.
[55] *Aromatase (CYP19) on testosterone:* 1.14.14.1; I.
[56] *Hydroxysteroid (17-β) dehydrogenase 2 (HSD17B2) on estradiol:* 1.1.1.62; I.
[57] *17-ketoreductase (HSD17B) on estrone:* 1.1.1.51; I.
[58] *Cholesterol 7,8-dehydrogenase (Nvd/DAF36)*; I.
[59] *Ecdysteroid 14-hydroxylase*; I.
[60] *Other blackbox enzymes*; I.
[61] *Ecdysteroid 25-hydroxylase (CYP306A1)*; I.
[62] *Ecdysteroid 22-hydroxylase (CYP302A1)*; I.
[63] *Ecdysteroid 2-hydroxylase (CYP315A1)*; I.
[64] *20-hydroxylase (CYP314A1) on ecdysone:* 1.14.99.22; I.
[65] *5α-steroid 4-dehydrogenase on 7-dehydrocholesterol*; I.
[66] *3β-hydroxy-Δ5-steroid dehydrogenase on lathosterol*; I.
[67] *Steroid 26-hydroxylase (CYP22A) on lathosterone*; I.
[68] *Steroid further 26-hydroxylase (CYP22A) on 26-hydroxylathosterone*; I.
[69] *3β-hydroxy-Δ5-steroid dehydrogenase on cholesterol (hsd-1)*; I.
[70] *Steroid 26-hydroxylase (CYP22) on 4-cholestene-3-one*; I.
[71] *Steroid 26-carboxylase (CYP22) on 26-hydroxy-4-cholestene-3-one*; I.
[72] *Progesterone 15α-hydroxylase*; I.
[73] *17-hydroxyprogesterone 15α-hydroxylase*; I.
[74] *Androstenedione 15α-hydroxylase*; I.
[75] *Testosterone 15α-hydroxylase*; I.
[76] *Aromatase on 15α-hydroxyandrostenedione;* I.
[77] *Aromatase on 15α-hydroxytestosterone;* I.
[78] *17-ketoreductase on 15α-hydroxyestrone*; I.
[79] *Hydroxysteroid (17-β) dehydrogenase on 15α-hydroxyestradiol*; I.
[80] *Cycloeucalenol cycloisomerase:* 5.5.1.9; I.
[81] *Cholesterol 7α-hydroxylase (CYP7A1):* 1.14.13.17; I.
[82] *Cholest-5-ene-3β,7α-diol 3β-dehydrogenase (HSD3B7):* 1.1.1.181; I.
[83] *Sterol 12α-hydroxylase (CYP8B1):*1.14.13.95; I.
[84] *Δ4-3-oxosteroid 5β-reductase (AKR1D1) on 7α,12α-dihydroxy-4-cholesten-3-one:* 1.3.1.3; I.
[85] *3α-hydroxysteroid dehydrogenase (AKR1C4) on 7α,12α-dihydroxy-4-cholestan-3-one:* 1.1.1.50.; I.
[86] *5β-cholestane-3α,7α,12α-triol 26-hydroxylase (CYP27A1):* 1.14.13.15; I.
[87] *5β-cholestane-3α,7α,12α,26-tetrol 26-carboxylase (CYP27A1):* 1.14.13.15; I.
[88] *Bile acid coenzyme A ligase and β-oxydation enzymes in peroxisomes on 3α,7α,12α-trihydroxy-5β-cholestenoic acid: 6.2.1.7, 5.1.99.4, 1.17.99.3, 4.2.1.107, 1.1.1.35, 2.3.1.176, 6.2.1.7,* I.
[89] *Δ4-3-oxosteroid 5β-reductase (AKR1D1) on 7α-hydroxy-4-cholesten-3-one:* 1.3.1.3; I.

[90] *3α-hydroxysteroid dehydrogenase (AKR1C4) on 7α-hydroxy-5β-cholestan-3-one:* 1.1.1.50.; I.
[91] *Sterol 26-hydroxylase (CYP27A1) on 3α,7 -dihydroxy-5β-cholestane:* 1.14.13.15; I.
[92] *Sterol further 26-hydroxylase (CYP27A1) on 3α,7α,26-trihydroxy-5β-cholestane:* 1.14.13.15; I.
[93] *Bile acid coenzyme A ligase and β-oxydation enzymes in peroxisomes on 3α,7α-dihydroxy-5β-cholestenoic acid: 6.2.1.7, 5.1.99.4, 1.17.99.3, 4.2.1.107, 1.1.1.35, 2.3.1.176, 6.2.1.7,* I.
[94] *Cholesterol 26-hydroxylase (CYP27A1):* 1.14.13.15; I.
[95] *26-hydroxysterol further 26-hydroxylase (CYP27A1):* 1.14.13.15; I.
[96] *Oxysterol 7α-hydroxylase (CYP7B1) on 3β-hydroxy-5-cholestenoate:* 1.14.13.100; I.
[97] *Cholest-5-ene-3β,7α-diol 3β-dehydrogenase (HSDB7) on 3β,7α-dihydroxy-5-cholestenoic acid:* 1.1.1.181; I.
[98] *Δ4-3-oxosteroid 5β-reductase on 7α-hydroxy-3-oxo-4-cholestenoic acid:* 1.3.1.3; I.
[99] *3α-hydroxysteroid dehydrogenase on 3α,7α-dihydroxy-cholestenoic acid:* 1.1.1.50; I.
[100] *Bile acid coenzyme A ligase and β-oxydation enzymes in peroxisomes on 3α,7α-dihydroxy-5β-cholestenoic acid: 6.2.1.7, 5.1.99.4, 1.17.99.3, 4.2.1.107, 1.1.1.35, 2.3.1.176, 6.2.1.7;* I.
[101] *Cholesterol 24-hydroxylase (CYP46A1):* 1.14.13.98; I.
[102] *Vitamin D3 25-hydroxylase (CYP2R1 or CYP27A1):* 1.14.15.- ; I.
[103] *25-hydroxyvitamin D3 1α-hydroxylase (CYP27B1):* 1.14.13.13; I.
[104] *Corticosterone 18-hydroxylase (CYP11B):* 1.14.15.5; I.
[105] *18-hydroxycorticosterone further 18-hydroxylase (CYP11B):* 1.14.15.5; I.
[106] *24-hydroxycholesterol 7α-hydroxylase (CYP39):* 1.14.13.99
[107] *(24S)-cholest-5-ene-3β,7α,24-triol 3β-dehydrogenase (HSD3B7):* 1.1.1.181; I.
[108] *Δ4-3-oxosteroid 5β-reductase on 7α,24-dihydroxy-4-cholesten-3-one:* 1.3.1.3; I.
[109] *3α-hydroxysteroid dehydrogenase on 3α,7α,24-trihydroxy-5β-cholestane:* 1.1.1.50.; I.
[110] *Bile acid coenzyme A ligase and β-oxydation enzymes in peroxisomes on 3α,7α,24-trihydroxy-5β-cholestane: 6.2.1.7, 5.1.99.4, 1.17.99.3, 4.2.1.107, 1.1.1.35, 2.3.1.176, 6.2.1.7;* I.
[111] *Cholesterol 25-hydroxylation (CH25H):* 1.14.99.38; I.
[112] *25-hydroxycholesterol 7α-hydroxylase (CYP7B1):* 1.14.13.100; I.
[113] *7α,25-dihydroxycholesterol 3β-oxidase (HSD3B7):* 1.1.1.181; I.
[114] *Δ4-3-oxosteroid 5β-reductase on 7α,25-dihydroxy-4-cholesten-3-one:* 1.3.1.3; I.
[115] *3α-hydroxysteroid dehydrogenase on 7α,25-dihydroxy-5β-cholestan-3-one:* 1.1.1.50; I.
[116] *Bile acid coenzyme A ligase and β-oxydation enzymes in peroxisomes on 3α,7α,25-trihydroxy-5β-cholestane;* I.
[117] Sterol 24-C-methyltransfer (smt1); IIa.
[118] Sterol 14-demethylation (CYP51); IIa.
[119] Delta(14)-sterol reduction; IIa.
[120] Methylsterol monooxygenation, sterol-4α-carboxylate 3-dehydrogenation, 3-keto-steroid reduction on 4,4 dimethylsterol; IIa.
[121] Methylsterol monooxygenation, sterol-4α-carboxylate 3-dehydrogenation, 3-keto-steroid reduction on 4α-methylsterol; IIa.
[122] C-8 sterol isomeration; IIa.
[123] Delta(7)-sterol 5-desaturation; IIa.
[124] Delta(24)-sterol reduction; IIa.
[125] Steroid 17α-hydroxylation on a steroid with cleaved side chain (CYP17A1); IIa.
[126] 21-hydroxylation on carbon with cleaved side chain (CYP21); IIa.
[127] 3β-hydroxy-Δ5-steroid dehydrogenation on a steroid with cleaved side chain (HSD3B); IIa.
[128] 17α-hydroxysteroid aldolisation; IIa.
[129] 17-ketoreduction; IIa.
[130] 3-oxo-5α-steroid 4-dehydrogenation; IIa.
[131] 17β-hydroxysteroid dehydrogenation; IIa.
[132] 11β-hydroxysteroid dehydrogenation; IIa.

[133] Steroid 11α or 18α-hydroxylation (CYP11B); IIa.
[134] Aromatisation; IIa.
[135] Sterol 26-hydroxylation (CYP22); IIa.
[136] 26-hydroxysterol further hydroxlation (CYP22); IIa.
[137] Steroid 15α-hydroxylation; IIa.
[138] Steroid 7α-hydroxylation; IIa.
[139] Steroid 12α-hydroxylation; IIa.
[140] Steroid 1α-hydroxylation; IIa.
[141] 3β-hydroxy-Δ5-steroid dehydrogenation on a steroid with a side-chain (HSD3B7); IIa.
[142] Δ4-3-oxosteroid 5β-reduction (AKR1D1); IIa.
[143] 3α-hydroxysteroid dehydrogenation (AKR1C4); IIa.
[144] 27carboxylation; IIa.
[145] Coenzyme A ligation followed by β-oxidation sequence; IIa.
[146] Cyclisation; IIc.
[147] Steroid hydroxylation (CYP); IIc.
[148] Hydroxysteroid dehydrogenation on 11β or 17β-hydroxysteroid; IIc.
[149] 26 or 27 carboxylation; IIc.
[150] 3β-hydroxy-Δ5-steroid dehydrogenation; IIc.
[151] Steroid 25 or 26 hydroxylation performed by CYP27A1; IIe.

C

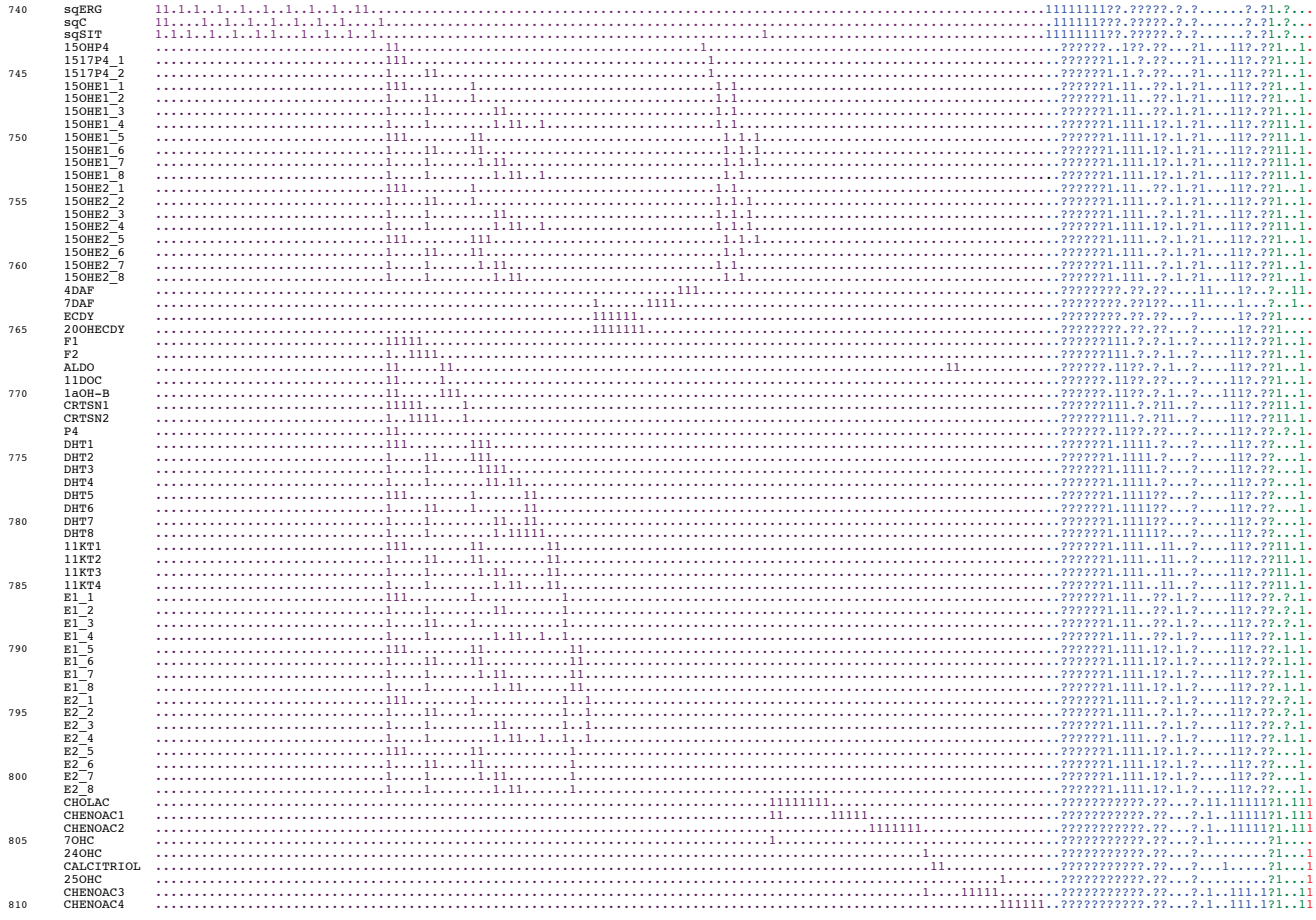


Figure S1: Data matrix containing 71 taxons (rows) and 151 characters (columns) ***A.*** definition of the compared taxons. As almost all taxa are enzymes synthesized from cholesterol, taxons names such as « F1 » mean « synthesis of F1 from cholesterol). The only four exceptions are sqERG, sqC and sqSIT that are synthesized from squalene, and CALCITRIOL, which is synthesized from 7dehydrocholesterol. When there are several manners to synthesize a product, the indication « via X» indicates reaction numbers (see Fig. 3) that are specific to each pathway. Thus it is possible to discriminate between various possibilities. Each taxon corresponds to one of the lines of the data matrix shown in *C*. ***B.*** character names, with the corresponding number of international nomenclature, when possible, and homology types defined in the text (see Fig. 1). Each character corresponds to one of the column of the data matrix shown in *C*. Characters are colored according to the following convention: type I homologies (hI), type IIa homologies (hIIa), type IIc homologies (hIIc) and type IIe homology (hIIe). Names of compounds that are shortened on Fig. 3 are here given in full, with the abbreviation between brackets. ***C.*** matrix. Each line is a taxon, i.e. a given pathway, as detailed in *A*. Each column is a character, i.e. either a "type I" homology (when an enzyme is shared by several pathways with the same specificity for its substrate in them) or a "type II" homology (when a type of reaction is performed in several pathways without considering specificity), as detailed in *B*. Dots « . » and « 1 » refer to the character state found in the corresponding taxon. « 1 » indicates the presence of the character and « . » its absence. Question marks are assigned to character states when the enzyme or the enzymatic function is not applicable to the taxon: the required substrate is not available in this pathway. Characters (columns) are numbered following their order from the left to right: the character number one is the first column; the character number 36 is the 36th column. The matrix columns are ordered as follows: 116 type I homologies (hI), 29 type IIa homologies (hIIa), 5 type IIc homologies (hIIc) and one type IIe homology (hIIe).

Fig. S1

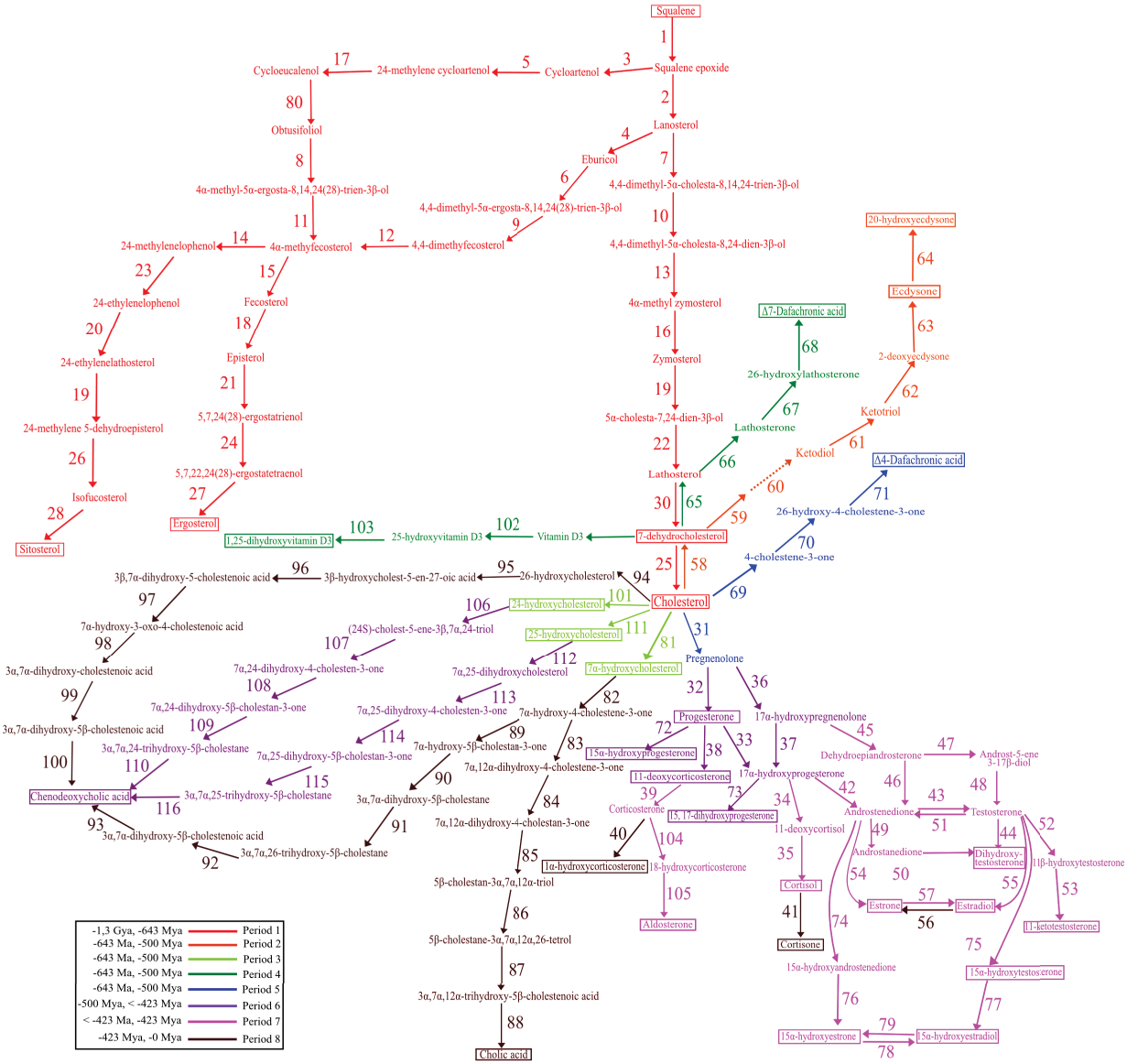

Squalene
Squalene epoxide
Cycloartenol
24-methylene cycloartenol
Cycloeucalenol
Obtusifoliol
Lanosterol
Eburicol
4,4-dimethyl-5α-cholesta-8,14,24-trien-3β-ol
4α-methyl-5α-ergosta-8,14,24(28)-trien-3β-ol
4,4-dimethyl-5α-ergosta-8,14,24(28)-trien-3β-ol
24-methylenelophenol
4α-methylfecosterol
4,4-dimethylfecosterol
4,4-dimethyl-5α-cholesta-8,24-dien-3β-ol
24-ethylenelophenol
Fecosterol
4α-methyl zymosterol
Δ7-Dafachronic acid
20-hydroxyecdysone
Ecdysone
2-deoxyecdysone
24-ethylenelathosterol
Episterol
Zymosterol
26-hydroxylathosterone
Ketotriol
24-methylene 5-dehydroepisterol
5,7,24(28)-ergostatrienol
5α-cholesta-7,24-dien-3β-ol
Lathosterone
Ketodiol
Δ4-Dafachronic acid
Isofucosterol
5,7,22,24(28)-ergostatetraenol
Lathosterol
26-hydroxy-4-cholestene-3-one
Sitosterol
Ergosterol
1,25-dihydroxyvitamin D3
25-hydroxyvitamin D3
Vitamin D3
7-dehydrocholesterol
4-cholestene-3-one
3β,7α-dihydroxy-5-cholestenoic acid
3β-hydroxycholest-5-en-27-oic acid
26-hydroxycholesterol
Cholesterol
7α-hydroxy-3-oxo-4-cholestenoic acid
24-hydroxycholesterol
(24S)-cholest-5-ene-3β,7α,24-triol
25-hydroxycholesterol
3α,7α-dihydroxy-cholestenoic acid
7α,24-dihydroxy-4-cholesten-3-one
7α,25-dihydroxycholesterol
7α-hydroxycholesterol
Pregnenolone
3α,7α-dihydroxy-5β-cholestenoic acid
7α,24-dihydroxy-5β-cholestan-3-one
7α,25-dihydroxy-4-cholesten-3-one
7α-hydroxy-4-cholestene-3-one
Progesterone
17α-hydroxypregnenolone
3α,7α,24-trihydroxy-5β-cholestane
7α,25-dihydroxy-5β-cholestan-3-one
7α-hydroxy-5β-cholestan-3-one
15α-hydroxyprogesterone
Dehydroepiandrosterone
Androst-5-ene 3-17β-diol
Chenodeoxycholic acid
3α,7α,25-trihydroxy-5β-cholestane
7α,12α-dihydroxy-4-cholestene-3-one
11-deoxycorticosterone
17α-hydroxyprogesterone
3α,7α-dihydroxy-5β-cholestane
Androstenedione
Testosterone
3α,7α-dihydroxy-5β-cholestenoic acid
7α,12α-dihydroxy-4-cholestan-3-one
Corticosterone
15, 17-dihydroxyprogesterone
11-deoxycortisol
Androstanedione
Dihydro-testosterone
11β-hydroxytestosterone
3α,7α,26-trihydroxy-5β-cholestane
1α-hydroxycorticosterone
18-hydroxycorticosterone
5β-cholestan-3α,7α,12α-triol
Cortisol
Estrone
Estradiol
11-ketotestosterone
Aldosterone
5β-cholestane-3α,7α,12α,26-tetrol
Cortisone
15α-hydroxyandrostenedione
15α-hydroxytestosterone
3α,7α,12α-trihydroxy-5β-cholestenoic acid
15α-hydroxyestrone
15α-hydroxyestradiol
Cholic acid
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
-643 Ma, -500 Mya Period 5
-500 Mya, < -423 Mya Period 6
< -423 Ma, -423 Mya Period 7
-423 Mya, -0 Mya Period 8

Fig. S2

840

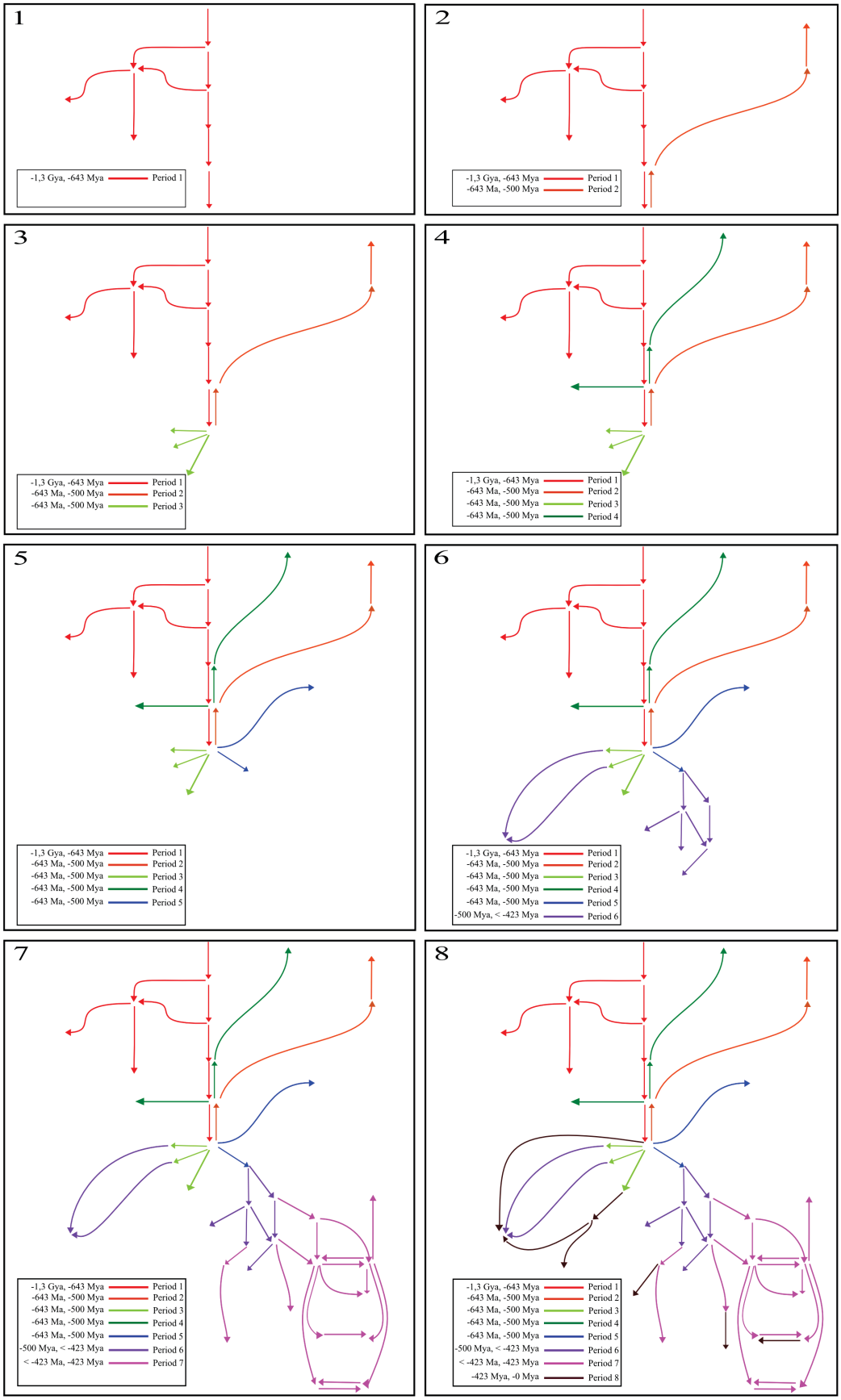

1
-1,3 Gya, -643 Mya Period 1
2
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
3
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
4
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
5
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
-643 Ma, -500 Mya Period 5
6
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
-643 Ma, -500 Mya Period 5
-500 Mya, < -423 Mya Period 6
7
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
-643 Ma, -500 Mya Period 5
-500 Mya, < -423 Mya Period 6
< -423 Ma, -423 Mya Period 7
8
-1,3 Gya, -643 Mya Period 1
-643 Ma, -500 Mya Period 2
-643 Ma, -500 Mya Period 3
-643 Ma, -500 Mya Period 4
-643 Ma, -500 Mya Period 5
-500 Mya, < -423 Mya Period 6
< -423 Ma, -423 Mya Period 7
-423 Mya, -0 Mya Period 8

Fig. S3

Part V

Discussion

Chapter 6

Some cues on the origin of steroid signaling from a dietary background

Here we will briefly sum up the major conclusions of our three papers, and we will try to put all these observations in a more precise zoological and paleontological framework. The update of the NR phylogeny taking into account genomic data from all metazoan groups makes possible to reassess the distribution of ligand-binding abilities along the NR tree (Fig. 22). There is still an important bias in the available data, with most of the ligand-binding data being available for vertebrates. But with this in mind, we think it is worth to propose a general analysis of the distribution pattern.

6.1 The ancestral nuclear receptor may have been a fatty acid sensor

Receptors branching at the basalmost nodes of the NR tree, such as the metazoan NR2A, the sponge NR2I and the eumetazoan NR2B are all able to bind fatty acids with micromolar affinity. The biological effect of such a binding is nevertheless quite variable. Some of them unambiguously activate gene transcription, such as the sponge NR2A (Bridgham et al., 2010) whereas in some other cases, the binding of a fatty acid does not seem to modulate transcriptional activation. This is the case for the sponge NR2I (Bridgham et al., 2010) or for the mammalian NR2A1 (Yuan et al., 2009). This observation has led to the proposal that the ancestral receptor was permanently bound to a fatty acid, that would play the role of a structural ligand, necessary to give the receptor its native conformation but not to trigger transcriptional activity (Sladek, 2002). However, one has to take into account that all these receptors, even if they branch at basal nodes, have evolved as much time as the others, with the possibility to acquire specific features. In particular, concerning mammals, other receptors, the PPARs, that belong to the NR1C group, are very efficient fatty acid sensors. Thus it is possible that they have at least partially outcompeted the NR2A as regulators of fatty acid metabolism, thus making possible a transformation from an ancestral fatty acid sensor into a transcription factor with a structural ligand.

Of note, there is fully possible that fatty acids were not the only molecules the ancestral NR was able to sense. Hemes may be good candidates, thus making possible the connection between light perception and transcriptional activity, which is the basis for circadian rhythm, a physiological process that is probably as universal as the regulation of fatty acid synthesis. But for the moment, because data on heme-binding ability are too scarce, we think it is not meaningful to speculate on that for the moment.

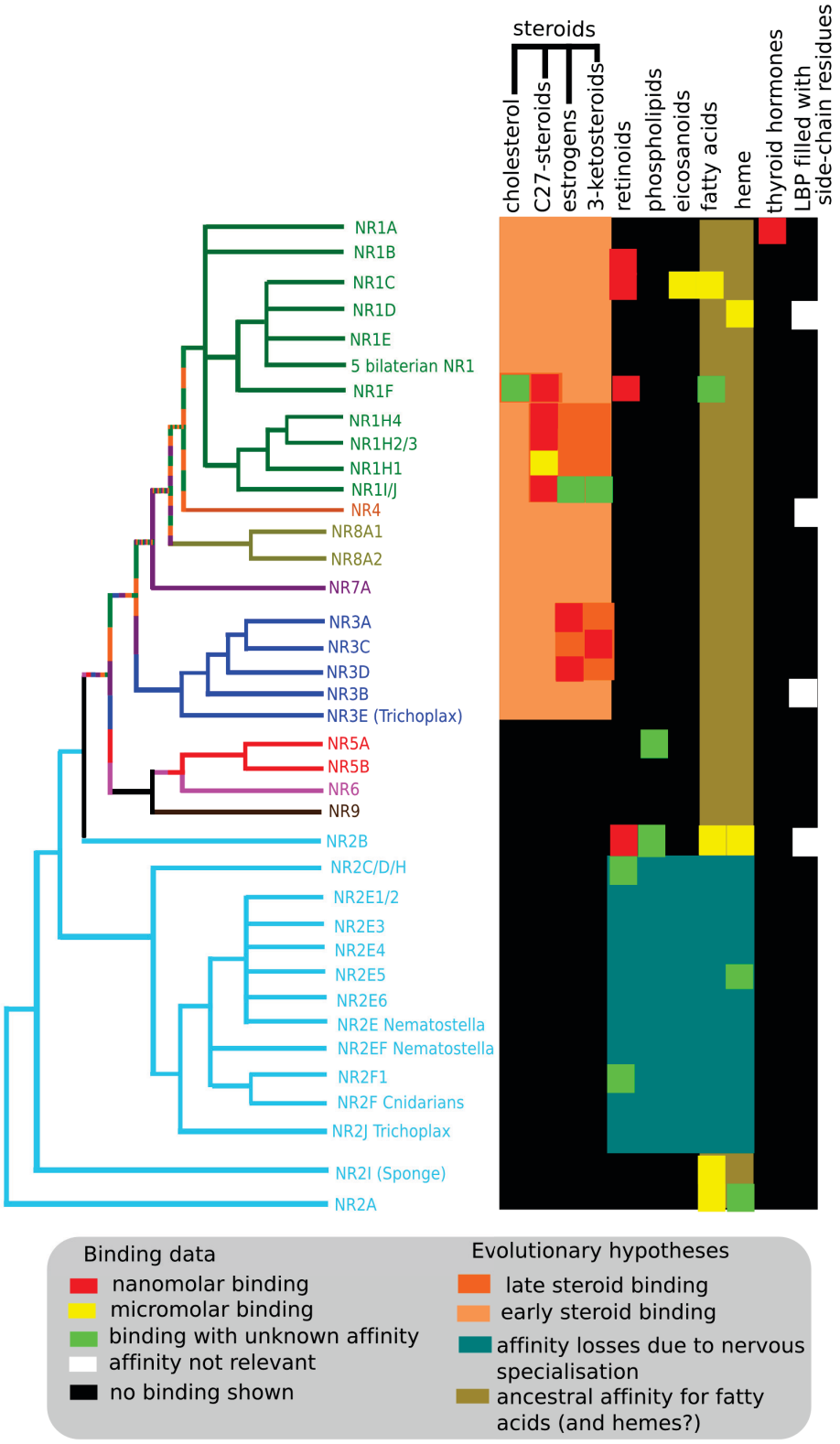


Figure 22: Summary about ligand-binding abilities of NRs in the completed metazoan framework. Hypotheses on evolution of these abilities are also indicated, and discussed extensively in text.

Another open question is whether other transcription factors were already able to modulate transcription in response to small hydrophobic molecules in the ancestor of metazoans or if this ability was acquired with the first nuclear receptor. Indeed, in fungi, many different transcription factors were recently shown to be ligand-activated (Näär and Thakur, 2009). And concerning metazoans, other transcription factors that can be activated by a ligand are the bHLH-PAS, that are known to activate transcription triggered by the binding of xenobiotics (Hahn, 2002) or gases bound to their hemes (Mukaiyama et al., 2006). The phylogeny of this family is not studied as deeply as this of NRs, but preliminary analyses show that it has diversified earlier than NRs, with at least three members in sponges (Simionato et al., 2007). Additionally, it also remains possible that modulation of transcriptional activity by small lipophilic molecules was carried by allosteric regulation of other proteins than transcription factors.

If the ancestral NR was a fatty acid sensor, there are two main observations that need to be interpreted concerning the ligand distribution pattern. The first is that members of the NR2C/D/H and NR2E/F groups have only limited binding activities (Fig. 22). The second is that, on the contrary, the other receptors underwent a widening of their ligand spectrum. Both processes occurred in parallel during the early diversification of eumetazoans.

6.2 NR-mediated steroid signaling in eumetazoans may be a by-product of extracellular digestion

An important characteristic of the transition between the ediacaran[1] word of filtrating animals and the cambrian word of bilaterian animals is the apparition of macropredation (Conway Morris, 2000). Until this time, metazoans are supposed to be in majority filter feeders, filtering non-metazoan particles such as bacterias and unicellular algae. The ”cambrian explosion” is often viewed under the angle of a mechanical arms race, where the apparition of claws, jaws and other crunching devices selected the apparition of protective shells and carapaces. But there was probably an other dimension in this arms race. Shells are not the only protective devices in animals. An other side of the protection is the synthesis of chemical molecules that act as food repellents. This phenomenon is quite well described in the case of plant-herbivore interactions in terrestrial ecosystems, where many plants are able to disrupt the reproductive regulation of their predators (Coley et al., 1985). But this should also be the case in sessile metazoan animals, that have to resist the predation pressure of some bilaterians. This would be consistent with the high concentrations of ecdysteroids in these animals, that are not compatible with an hormonal role. In fact, such a chemodefensive role is already clearly demonstrated for some epidermal ecdysteroids in some arthropods (Lafont and Mathieu, 2007).

A major physiological difference between sponges and eumetazoans concerns the feeding mechanism. In sponges, each cell digests food particles through endocytosis, so the entire ingested material is confined to very specific cell compartments, such as endosomal vesicles that fusions with lysosomes (Fig. 23A). On the contrary, in eumetazoans, the apparition of the gut means that digestion becames mainly extracellular (Fig. 23B), even if it can be completed by endocytosis, at least in cnidarians. This increased specialisation, linked with the presence of true epithelia that strictly delimit extra- and intraorganismal compartments, also implies an increased division of labour between cell types. One striking example of cellular specialisation is the case of the

[1] Ediacaran is a geological period that predates the cambrian, between 630 and 540 Million years.

nervous cells. In sponges, various cells exhibit some sensory abilities (Renard et al., 2009), but an anatomically recognisable nervous system does not exist. Since many of the nuclear receptors from the NR2C/D/H and NR2E/F groups diversified in animals with a nervous system and are known to be important regulators of nervous system differenciation, we hypothesize that this was their primary function. It is possible that in such a process, that is strongly controlled by endogenous parameters, the ability to be modulated by small lipophilic molecules was not functionally important, and that this allowed the loss or weakening of ligand-binding ability in these groups. One other possibility is that there was a division of labour even in the nervous system between NR2 and members of the other NR families that diversified in parallel with a widening of their ligand-binding abilities. Such a general widening is not contradictory with the later acquisition of high-affinity ligands, such as thyroid hormones for the chordate NR1A. The exploration of various ligand-receptor couple due to natural variation may have given rise to very specific associations linked with a peculiar physiological advantage that could have triggered an increase in specificity.

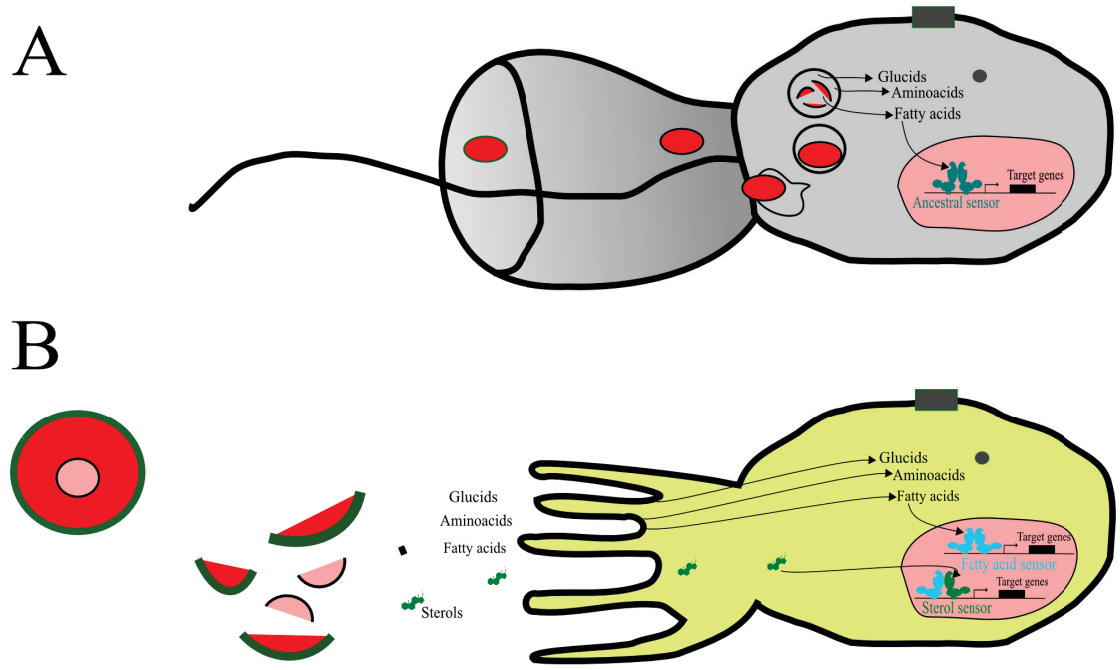


Figure 23: Implications of extracellular digestion on nuclear receptor signaling. (A) Digestion through endocytosis by an individual choanocyte. Bacterias (in red) are entirely absorbed by each cell through endocytosis. (B) Absorption of extracellularly digested material by an enterocyte. Here the ingested eucaryotic cells, possibly from other metazoans, are mechanically and chemically destroyed in the digestive tube, and the produced nutrients are directly absorbed through passive or facilitated diffusion. But the absorption barrier thus can also be crossed by various toxins, among which they could be steroid molecules. Event if both types digestive mechanisms are present here separated for clarity purposes, they are not mutually exclusive.

Extracellular digestion has important implications in terms of ligand ability for nuclear receptors (Fig. 23). In intracellular digestion through endocytosis, the food particles - that are of limited size, such as bacterial cells - are uptaken by endocytosis vesicles and are digested when the vesicles fusions with lysosomes (Fig. 23A). This

means that only digested metabolites, such as aminoacids, glucids and fatty acids go out of the lysosome. Between these metabolites, only fatty acids are primarily nuclear receptor ligands. Of note, thyroid hormones are aminoacid derivatives, but their presence and physiological role outside metazoans are poorly known for the moment. On the contrary, extracellular digestion implies that the ingested food particles - that can be eucaryotic cells, or even multicellular organisms - are destroyed in the gut, and that nutrients diffuse to the animal through the membrane of the cells that consist the digestive epithelium (Fig. 23B). With the possibility to eat other eucaryotes, the first eumetazoans also became to eat sterol molecules. The ability to digest extracellularly provided such animals with increased food quantities could have been important in a world shifting from bacteria-dominated biomass to metazoan-dominated biomass (Butterfield, 2011). But this had also some side effects. With the nutrients, other components from the ingested food could cross the epithelial barrier. Among them where the sterols from the membranes of the eaten eucaryotes, that could act as modulators for the intracellular nuclear receptors of the enterocytes. This is well illustrated by the known pleiotropic effects of genistein, a component of soy, in humans (Henley and Korach, 2006). The presence of an NR as a sensing transcription factor would have enabled the possibility to efficiently regulate the production of detoxifying enzymes in response to endocrine perturbation due to food nutrients (Baker, 2005). Moreover, dietary sterols or steroids may have allowed the coordination of the reproduction cycle with food availability through NR-mediated activation of reproductive maturation.

6.3 Independent acquisition of steroidogenic synthesis pathways in bilaterians

One critical aspect to understand the context of ancestral NR binding is to decipher the timing of the appearance of steroids. Steroids are reported in almost all animal groups (Fig. 24), the most recent example being progesterone in rotifers (Stout et al., 2010) and even in plants. Unfortunately, many searches for “human”-type steroid hormones such as estradiol or progesterone throughout metazoan groups have been prone to artifacts and/or misidentification. To date, biochemical evidence (immunological and/or chromatographic methods linked to mass spectrometry) for presence of vertebrate steroids in lophotrochozoans, ecdysozoans and cnidarians have not been substantiated by molecular characterization of enzymes directly involved in their de novo biosynthesis (Lafont and Mathieu, 2007). In the part III of this work, we show that the enzymes implicated in the synthesis of vertebrate sex and adrenal steroids, as well as the enzymes implicated in the synthesis of ecdysone and dafachronic acids appeared specifically following gene duplications in the vertebrate, arthropod and nematode lineages, respectively. This is also true for some key enzymes that are involved in the synthesis of oxysterols, bile acids and vitamin D. All these enzymes are members of multigenic families where the phylogeny is far from being fully elucidated, but there are nevertheless some robust nodes, in particular those concerning two key enzymes in vertebrate sex and adrenal steroid synthesis: CYP19 and CYP11A.

The CYP19, also named aromatase, is a chordate-specific enzyme (Reitzel and Tarrant, 2010), that may either have been secondarily lost in other animals or, following the same reasoning as presented above (section 4.2.1) for vertebrate NR3C, may be a highly derived chordate protein that branches basally to other metazoan CYPs due to long-branch attraction. However, in both cases, this indicates that there is no evidence for a CYP19 ortholog that would perform aromatase activity outside chordates. This

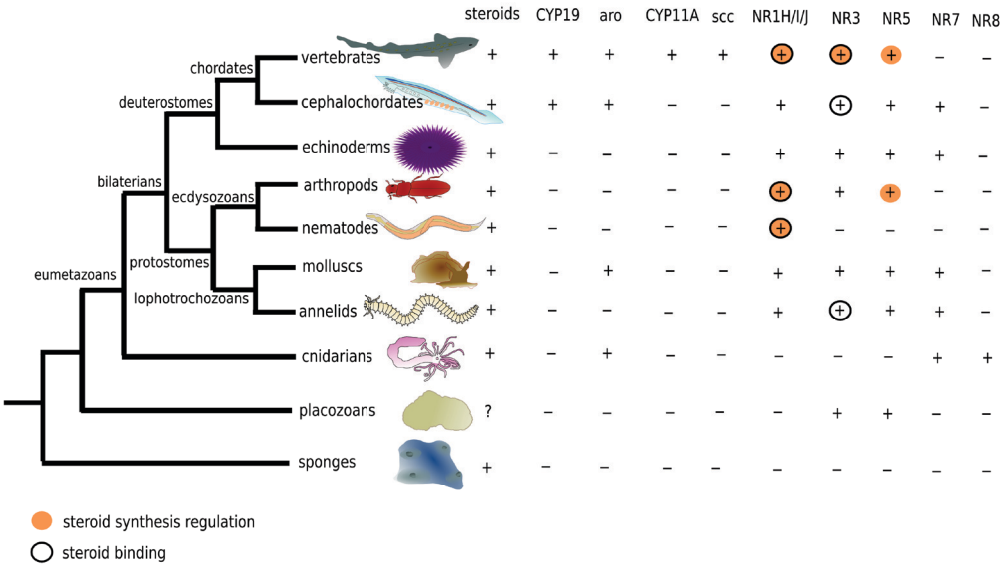


Figure 24: Distribution of steroids, steroidogenic receptors, steroidogenic enzymes and nuclear receptors involved in regulation of steroidogenesis in metazoans.

should be put in perspective with the reports from aromatase activities in cnidarians or mollusks (Fig. 24). This means that this aromatase activity may be performed by another enzyme than CYP19, and on an endogenous substrate that is not necessarily testosterone, as in vertebrates. Moreover, because all cnidarians studied to date lack any NR1I/J/H or any NR3 (Reitzel and Tarrant, 2009), if there is any NR-mediated steroid signaling in these animals, it would not be mediated by a classical bilaterian-type NR. Even the NR5, that are important regulators of steroidogenesis at least in vertebrates and arthropods, are not present in cnidarians. Thus, there is no reason to suppose that the physiologically active steroids in cnidarians, if they exist, would be of vertebrate-type, instead of being one of the various steroids that were identified in some cnidarians (Lafont and Mathieu, 2007). Concerning mollusks, that lack steroid-binding NR3, candidate steroid-binding receptors may be found in the NR1H/I/J group or more widely in the NR1 family and possibly also in the NR7 subfamily, but once again, that would not be in favour of a vertebrate-type estrogen ligand, rather than one of the various hydroxylated steroids once identified in some mollusks. Concerning annelids, where NR3D was shown to bind estrogens (Keay and Thornton, 2009), evidence of biochemical activity implicated in estrogen synthesis is poorly documented, and most steroid reports concern ecdysteroids (Lafont and Mathieu, 2007). In a recent report trying to link «vertebrate-like» steroid hormone levels to sexual maturity indexes in *Nereis diversicolor*, the authors acknowledge in the discussion that «*Quantification of steroid hormones in worms cannot permit [them] to take into account exogenous compounds accumulated in [worm tissues] and which can act as endocrine disruptors. The presence of these latter contaminants is well documented in the Seine estuary*» (Durou and Mouneyrac, 2007). More importantly, receptors from the NR7 family are absent from vertebrates and ecdysozoans, both groups where steroidogenesis and NR-mediated steroid-signaling pathway are genetically characterized. So this leaves fully open the possibility that in cnidarians, lophotrochozoans and maybe cephalochordates, the NR-mediated steroid signaling pathway uses a molecular machinery that is very different from this of vertebrates and ecdysozoans.

The late appearance of steroidogenic enzymes within highly multigenic and promiscuous families is in good agreement with the growing evidence that enzyme specificity should have evolved from low-specificity proteins catalyzing a whole range of activities at low levels, to subfamilies with potent and highly specialized activities (Khersonsky et al., 2006). This is also consistent with the striking similarity between the xenobiotic response pathway in vertebrates and the sex and adrenal steroid metabolism (Fig. 25). Xenobiotic response is divided in phase I (hydroxylations by CYPs) and phase II (addition of further hydrophilic residues on hydroxylated carbon), before transport outside the cell (Wada et al., 2009). Another fate of these xenobiotics is conjugation with fatty acids and storage in fat (Jandacek and Tso, 2001). Sex steroids are synthesized from cholesterol, mainly by a succession of hydroxylations catalysed by CYPs, and are directly degraded throughout a mechanism similar to the phase II xenobiotic response, involving enzymes from the SULT and UGT families (He et al., 2010). This similarity led us to the hypothesis that both phase I of xenobiotic detoxification and steroid synthesis are homologous, meaning that steroid hormones may be recruited cholesterol metabolites. This is in agreement with the proposition that thyroid hormones (T3 and T4) were dietary metabolites before becoming hormones endogenously synthesized by the thyroid gland in vertebrates (Miller and Heyland, 2010), and with the fact that other high affinity NR ligands, such as retinoids and eicosanoids, are also derivatives from food components. This hypothesis is further reinforced by the results we present in part IV of this manuscript, showing that steroid hormone synthesis evolved with the pattern of a cholesterol degradation pathway.

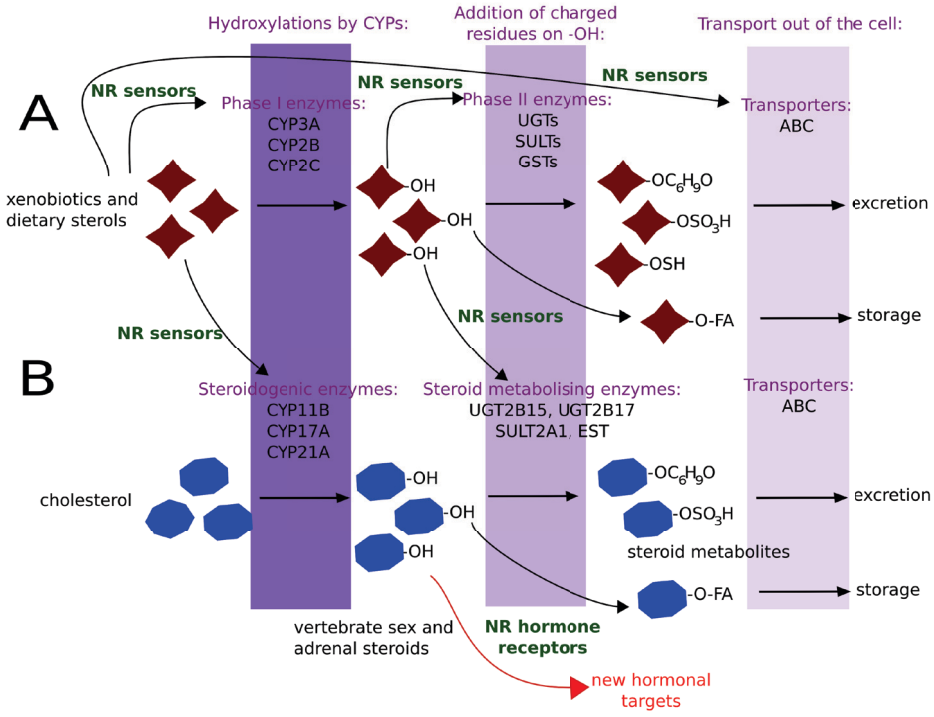


Figure 25: Similarities between xenobiotic response and steroid metabolism in vertebrates. (A) Xenobiotic response. (B) Steroid synthesis. In both case, the first step is hydroxylation by CYPs, and we hypothesize that these steps are thus homologous. After that, both steroids and xenobiotics are conjugated either to fatty acids, and then accumulated in fat, or conjugated to hydrophilic molecules and excreted out of the body.

The recruitement of endogenous synthesis of steroids may have been advantageous in comparison to exogenous regulation of reproductive maturation because in a relatively stable environment, the availability of food resources varies with a regular periodicity. So, animals who are able to adapt their reproductive maturation using periodic environmental cues, such as light signals, would have the possibility to partly prepare themselves to reproduction even before food becames abundant again. In animals without an internal body fluid that is able to carry hormones, such a coordination was probably mainly performed by the nervous system and by the secretion of neuropeptidic hormones. The existence of an internal signaling system based on steroid molecules may have facilitated the integration of both external parameters, such as light periodicity, and internal parameters, such as the metabolic state of the animal. As always, such an increased coordination had surely also its drawbacks, the main being that perturbation in the metabolic state of the animal may impact its reproductive status.

Chapter 7

Remaining problems... or observations without any functional interpretation

7.1 Vertebrates and side-chain cleaved steroids

The CYP11A, also known as side-chain cleavage enzyme, is unambiguously vertebrate-specific (Fig. 26), having arisen from a duplication of an ancestral CYP11 gene leading to CYP11A and CYP11B, that is involved in glucocorticoid synthesis. To date, a lack of phylogenetic resolution prevents us from establishing which CYP is the most closely related to these two vertebrate enzymes in other animals, even in chordates. But due to the fact that both duplicates have different activities, there is no reason to imagine that the activity of this CYP should be more CYP11A-like than CYP11B-like. Interestingly, a recent study succeeded in reporting in vitro steroidogenic biochemical activities for most of the members of the sex steroid synthesis pathway, except from the side-chain cleavage activity, for which the only indirect evidence is a report of mRNA expression in amphioxus ovaries (Mizuta et al., 2008). However, this so-called ≪CYP11A≫ is not even the best candidate for an ortholog of ancestral vertebrate CYP11, but rather is a distant paralog, named CYP374A2, which is clearly orthologous to a sea urchin CYP (Markov et al., 2009). This means that, up to now, the estrogen synthesis pathway remains vertebrate-specific, even if it is possible that amphioxus synthesizes other steroids, for example aromatized steroids with a side chain, which could cross-react with the antibodies used to search for estrogen in amphioxus (Mizuta and Kubokawa, 2007). So for the moment, estrogens cannot be viewed as a physiological ligand for amphioxus SR because there is no sufficient evidence for estrogen in this species. Indeed, even if the amphioxus SR was shown to be activated by estradiol, it was at very high concentrations (Bridgham et al., 2008), which makes it possible that the real physiological activation mechanism for this receptor is through phosphorylation, interactions with coactivators or binding of another ligand. This case is reminiscent of the relationship between RXR and 9-cis retinoic acid in vertebrates: 9-cis RA is an excellent ligand of vertebrate RXR but has not been found in vertebrate extracts and does not regulate RXR activity *in vivo*.

In addition, the result of our cladistic study (Part IV) confirm that side-chain cleavage is the reaction that makes vertebrate sex and adrenal steroids - that are NR3 ligands - different from steroids of other metazoans or other vertebrate steroids that are ligands for the NR1I/J group. The adaptative advantage of such a situation remains totally obscure for the moment. Maybe it is only a by-product of the whole-genome duplications that have given the rise to many new enzymatic activities.

The lack of vertebrate-type steroids outside vertebrates leaves open the question of the origin of steroid binding in the NR3 subfamily. It is possible, but remains untested,

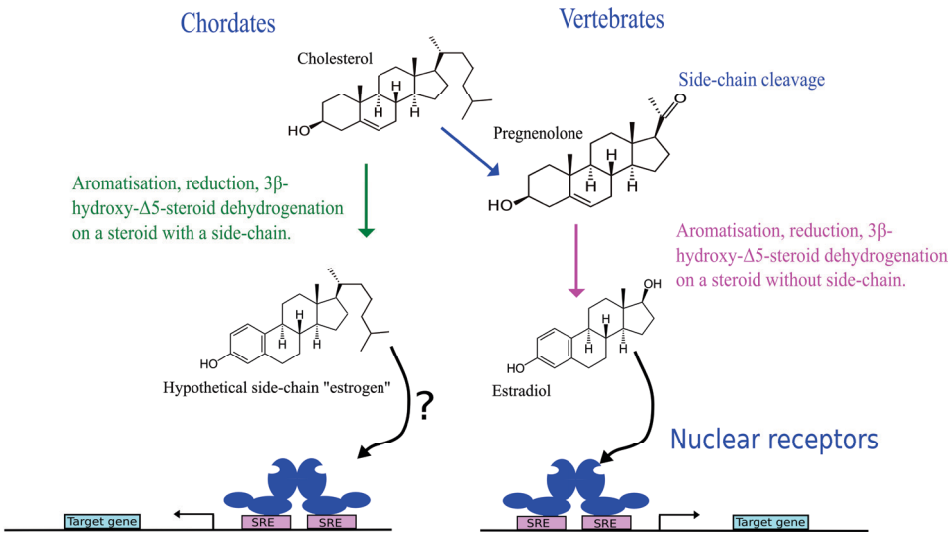


Figure 26: Appearance of side-chain cleaved steroids specifically in vertebrates. On the left side, the hypothesize ancestral situation, where a steroid with a lateral chain was hydroxylated and aromatized. On the right side, the acquisition of a side-chain cleavage activity has rendered possible the recruitement of the previousely existing hydroxylase and aromatase activities to a new substrate without a side-chain.

that the ability to bind steroid without a side-chain in annelids and chordates is the by-product of the physiologically relevant ability to bind another steroid with a lateral chain, or maybe a steroid without a side-chain that has not the structure of vertebrate estrogens. But functional studies on annelid receptors and characterization of the physiologically relevant steroids in these animals will be needed to answer properly this question.

7.2 The cnidarian puzzle

Another point that remains highly ambiguous is the status of steroid signaling in cnidarians. Because they lack an internal circulating body fluid, cnidarians cannot have a vertebrate-like hormonal system, but this does not mean that there is no steroid-mediated intercellular signaling in those animals. Indeed, even in vertebrates, some steroids do not act through the canonical hormonal pathway. Classically, hormones are defined as internal circulating molecules, and so vertebrate bile acids are somehow excluded from this definition (Norris, 2007). However, this view is changing with the acknowledgment of their functional role not only as facilitators of lipid digestion, but also as signaling molecules (Fig. 27), that allow the coupling of the nutritional state with various parameters, from digestive physiology to behavioural traits such as the regulation of appetite (Thomas et al., 2008).

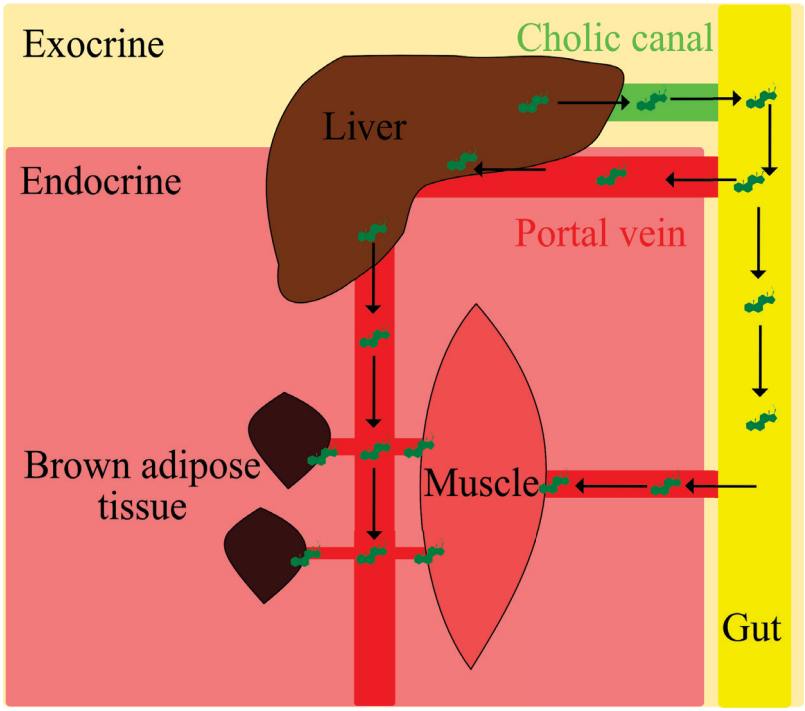


Figure 27: Bile acid signaling begins with an enterocrine part, when bile salts are excreted in the gut. Here there help in lipid absorption. A part of them is reabsorbed with the dietary lipids and go again in the liver, where they are degraded and where they downregulate their own synthesis, or in other organs, such as brown adipose tissue or muscle, where they upregulate thyroid hormone activation, thus promoting energy expenditure. Modified from Thomas et al., 2008.

This example shows that there is no objective reason to strictly limit long-distance intercellular communication to the internal milieu. We thus hypothesize, that in cnidarians, the gut is still a major vehicle for signaling molecules, and that the steroids identified in these animals, if they really have an hormonal role, could be transported in this way. The gut lumen is sometimes viewed as the continuation of the external milieu, because both are communicating through the oral openings. This is why secretions of molecules in the gut are considered as exocrine secretions. But in fact, the real chemical composition of some parts of the digestive tract is very different from the external word, and is mainly controlled by secretions and selective absorption processes in the digestive apparatus of the animal. And because the gut goes throughout the body, it can be a perfect carrier of signaling molecules (Fig. 28) between distant body parts.

From a molecular viewpoint, it is worth to mention that NR7 and NR8 (that is orthologous to the bilaterian NR1 and NR4) are found specifically in anthozoan cnidarians, that are the group where most steroids are reported. So they may be first-rate candidates for NR-mediated steroid signaling in such animals. However, it is also possible that in cnidarians, steroid signaling goes through non-genomic pathways.

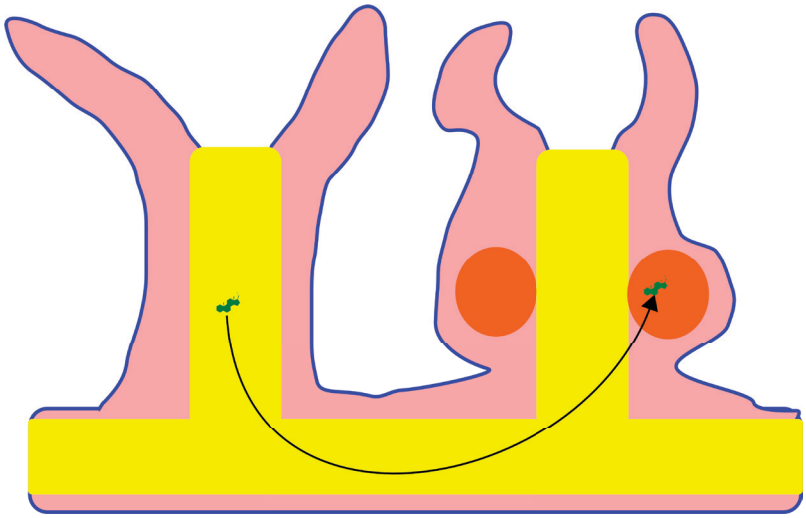


Figure 28: A possible way of steroid enterocrine signaling in corals. The steroids, in green, could be produced from dietary sterols in the gut (in yellow) and go to the gonad (in orange) with other nutrients. The pink part is the mesoglea, which contains cells but probably no circulating body fluid.

Part VI

Conclusion

In this study, we tried to understand when and why the nuclear receptor-mediated steroid signaling has appeared. We are not able to give a definitive answer to this question but we hope we succeeded in making this question more precise.

We propose that nuclear receptor-mediated steroid signaling is a by-product of extracellular digestion in metazoans, reusing signaling pathways that were primitively involved in the regulation of the reproductive maturation according to nutritional conditions.

This has some important implications on the way we have to look at endocrinology. Classically, as we pointed out in the introduction, hormonal signaling is viewed as a bottom-up cascade from the nervous centers to the target cells, and steroids occupy an intermediary position in the middle of this cascade. But our quick journey through metazoan signaling physiology reminds that a significant part of the environmental imput is integrated not through the nervous system, but through the digestive system (Fig. 29). This is consistent with the recent proposal that the liver may be an integrative center at the same level as the brain (Della Torre et al., 2011).

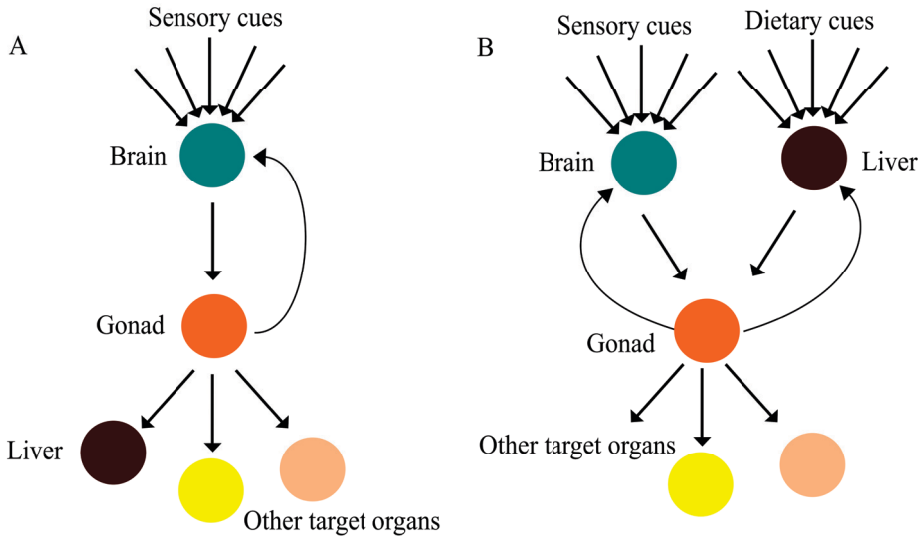


Figure 29: For a reappraisal of the integrative role of the digestive system. (A) A classical textbook hierarchical regulatory model, where the signal goes upside-down -even if retroactions are acknowledged - from the nervous centers to the target organs, among which is the liver. (B) A "pluralistic view", where not only the brain, but also the liver, and possibly the gonad too, are all partial and complementary integration centers.

Probably due to the special status of the brain as a "noble organ", being the supreme integrator of body signals, the nervous system has benefitted from many studies from evolutionary biologists, with many ramifications even out of the strict biological field. Let us hope that our work will contribute in stimulating next generations of scientists to dedicate the same attention to the digestive apparatus and to the functional connections between the various body integrators.

Ironically, this may be the occasion to reappreciate Carl Vogt's provocative proposition that *"thought is for the brain almost the same thing as is bile for the liver or urine for the kidneys."* (Vogt, 1847). This sentence was rapidly criticized by other physiologists, stressing that thought is not a fluid, contrary to bile or urine. But Vogt's aim was to point out that even the though processes have a material basis. This is now well accepted, but paradoxally, the brain has not yet been desacralized in regard to other integrators.

Part VII

Bibliography

Bibliography

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V. C., Caillaud, M.-C., Coutinho, P. M., Dasilva, C., Luca, F. D., Deau, F., Esquibet, M., Flutre, T., Goldstone, J. V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T. R., Markov, G. V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T. J., Blaxter, M., Bleve-Zacheo, T., Davis, E. L., Ewbank, J. J., Favery, B., Grenier, E., Henrissat, B., Jones, J. T., Laudet, V., Maule, A. G., Quesneville, H., Rosso, M.-N., Schiex, T., Smant, G., Weissenbach, J. and Wincker, P., 2008. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol*, 26(8):909–915.

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B. and de Rosa, R., 2000. The new animal phylogeny: reliability and implications. *Proc Natl Acad Sci U S A*, 97(9):4453–4456.

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. and Lake, J. A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632):489–493.

Amero, S. A., Kretsinger, R. H., Moncrief, N. D., Yamamoto, K. R. and Pearson, W. R., 1992. The origin of nuclear receptor proteins: a single precursor distinct from other transcription factors. *Mol Endocrinol*, 6(1):3–7.

Antebi, A., Culotti, J. G. and Hedgecock, E. M., 1998. *daf-12* regulates developmental age and the dauer alternative in *Caenorhabditis elegans*. *Development*, 125(7):1191–1205.

Baker, M. E., 2001. Evolution of 17beta-hydroxysteroid dehydrogenases and their role in androgen, estrogen and retinoid action. *Mol Cell Endocrinol*, 171(1-2):211–215.

Baker, M. E., 2002. Albumin, steroid hormones and the origin of vertebrates. *J Endocrinol*, 175(1):121–127.

Baker, M. E., 2003. Evolution of adrenal and sex steroid action in vertebrates: a ligand-based mechanism for complexity. *Bioessays*, 25(4):396–400.

Baker, M. E., 2005. Xenobiotics and the evolution of multi-cellular animals: Emergence and diversification of ligand-activated transcription factors. *Integr Comp Biol*, 45:172–178.

Baker, M. E., 2008. *Trichoplax*, the simplest known animal, contains an estrogen-related receptor but no estrogen receptor: Implications for estrogen receptor evolution. *Biochem Biophys Res Commun*, 375(4):623–627.

Bannister, R., Beresford, N., May, D., Routledge, E. J., Jobling, S. and Rand-Weaver, M., 2007. Novel estrogen receptor-related transcripts in *Marisa cornuarietis*; a freshwater snail with reported sensitivity to estrogenic chemicals. *Environ Sci Technol*, 41(7):2643–2650.

Benoit, G., Malewicz, M. and Perlmann, T., 2004. Digging deep into the pockets of orphan nuclear receptors: insights from structural studies. *Trends Cell Biol*, 14(7):369–376.

Bento, G., Ogawa, A. and Sommer, R. J., 2010. Co-option of the hormone-signalling module dafachronic acid-daf-12 in nematode evolution. *Nature*, 466(7305):494–497.

Bernard, C., 1865. *Introduction à l'étude de la médecine expérimentale*.

Bertrand, S., Brunet, F. G., Escriva, H., Parmentier, G., Laudet, V. and Robinson-Rechavi, M., 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. *Mol Biol Evol*, 21(10):1923–1937.

Billas, I. M., Moulinier, L., Rochel, N. and Moras, D., 2001. Crystal structure of the ligand-binding domain of the ultraspiracle protein usp, the ortholog of retinoid x receptors in insects. *J Biol Chem*, 276(10):7465–7474.

Bonneton, F. and Laudet, V., 2011. *Comprehensive Molecular Insect Science*, chapter Evolution of nuclear hormone receptors in insects, page in press. Elsevier.

Bookout, A. L., Jeong, Y., Downes, M., Yu, R. T., Evans, R. M. and Mangelsdorf, D. J., 2006. Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. *Cell*, 126(4):789–799.

Bourguet, W., Vivat, V., Wurtz, J. M., Chambon, P., Gronemeyer, H. and Moras, D., 2000. Crystal structure of a heterodimeric complex of rar and rxr ligand-binding domains. *Mol Cell*, 5(2):289–298.

Brelivet, Y., Kammerer, S., Rochel, N., Poch, O. and Moras, D., 2004. Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level. *EMBO Rep*, 5(4):423–429.

Bridgham, J. T., Brown, J. E., Rodríguez-Marí, A., Catchen, J. M. and Thornton, J. W., 2008. Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet*, 4(9):e1000191.

Bridgham, J. T., Carroll, S. M. and Thornton, J. W., 2006. Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 312(5770):97–101.

Bridgham, J. T., Eick, G. N., Larroux, C., Deshpande, K., Harms, M. J., Gauthier, M. E. A., Ortlund, E. A., Degnan, B. M. and Thornton, J. W., 2010. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol*, 8(10).

Brown, C. M., Reisfeld, B. and Mayeno, A. N., 2008. Cytochromes P450: a structure-based summary of biotransformations using representative substrates. *Drug Metab Rev*, 40(1):1–100.

Burris, T. P., 2008. Nuclear hormone receptors for heme: REV-ERBalpha and REV-ERBbeta are ligand-regulated components of the mammalian clock. *Mol Endocrinol*, 22(7):1509–1520.

Bury, N. R. and Sturm, A., 2007. Evolution of the corticosteroid receptor signalling pathway in fish. *Gen Comp Endocrinol*, 153(1-3):47–56.

Butterfield, N. J., 2011. Animals and the invention of the phanerozoic earth system. *Trends Ecol Evol*, 26(2):81–87.

Calléja, C., Messaddeq, N., Chapellier, B., Yang, H., Krezel, W., Li, M., Metzger, D., Mascrez, B., Ohta, K., Kagechika, H., Endo, Y., Mark, M., Ghyselinck, N. B. and Chambon, P., 2006. Genetic and pharmacological evidence that a retinoic acid cannot be the rxr-activating ligand in mouse epidermis keratinocytes. *Genes Dev*, 20(11):1525–1538.

Carr, M., Leadbeater, B. S. C., Hassan, R., Nelson, M. and Baldauf, S. L., 2008. Molecular phylogeny of choanoflagellates, the sister group to metazoa. *Proc Natl Acad Sci U S A*, 105(43):16641–16646.

Champlin, D. T. and Truman, J. W., 2000. Ecdysteroid coordinates optic lobe neurogenesis via a nitric oxide signaling pathway. *Development*, 127(16):3543–3551.

Cheung, E. and Kraus, W. L., 2010. Genomic analyses of hormone signaling and gene regulation. *Annu Rev Physiol*, 72:191–218.

Chevreul, M. E., 1815. Recherches chimiques sur plusieurs corps gras, et particulièrement sur leurs combinations avec les alcalis. cinquième mémoire. des corps qu'on a appeles adipocire, c'est-a-dire, de la substance cristallisee des calculs biliaires humains, du spermaceti et de la substance grasse des cadavres (lu le 19 septembre 1814). *Ann. Chim.*, 95:5–50.

Clayton, G. M., Peak-Chew, S. Y., Evans, R. M. and Schwabe, J. W., 2001. The structure of the ultraspiracle ligand-binding domain reveals a nuclear receptor locked in an inactive conformation. *Proc Natl Acad Sci U S A*, 98(4):1549–1554.

Coley, P. D., Bryant, J. P. and Chapin, F. S., 1985. Resource availability and plant antiherbivore defense. *Science*, 230(4728):895–899.

Conway Morris, S., 2000. The cambrian "explosion": slow-fuse or megatonnage? *Proc Natl Acad Sci U S A*, 97(9):4426–4429.

Darwin, C. R., 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London: John Murray.

de Rosny, E., de Groot, A., Jullian-Binard, C., Borel, F., Suarez, C., Pape, L. L., Fontecilla-Camps, J. and Jouve, H., 2008. DHR51, the *Drosophila melanogaster* homologue of the human photoreceptor cell-specific nuclear receptor, is a thiolate heme-binding protein. *Biochemistry*.

Della Torre, S., Rando, G., Meda, C., Stell, A., Chambon, P., Krust, A., Ibarra, C., Magni, P., Ciana, P. and Maggi, A., 2011. Amino acid-dependent activation of liver Estrogen Receptor Alpha integrates metabolic and reproductive functions via IGF-1. *Cell Metab*, 13(2):205–214.

Denver, R. J., 2009. Stress hormones mediate environment-genotype interactions during amphibian development. *General and Comparative Endocrinology*, 164:20–31.

Durou, C. and Mouneyrac, C., 2007. Linking steroid hormone levels to sexual maturity index and energy reserves in *Nereis diversicolor* from clean and polluted estuaries. *Gen Comp Endocrinol*, 150(1):106–113.

Eick, G. N. and Thornton, J. W., 2011. Evolution of steroid receptors from an estrogen-sensitive ancestral receptor. *Mol Cell Endocrinol*, 334(1-2):31–38.

Ekins, S., Reschly, E. J., Hagey, L. R. and Krasowski, M. D., 2008. Evolution of pharmacologic specificity in the pregnane X receptor. *BMC Evol Biol*, 8:103.

Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H. and Laudet, V., 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet*, 2(7):e102.

Escriva, H., Delaunay, F. and Laudet, V., 2000. Ligand binding and nuclear receptor evolution. *Bioessays*, 22(8):717–727.

Escriva, H., Safi, R., Hänni, C., Langlois, M. C., Saumitou-Laprade, P., Stehelin, D., Capron, A., Pierce, R. and Laudet, V., 1997. Ligand binding was acquired during evolution of nuclear receptors. *Proc Natl Acad Sci U S A*, 94(13):6803–6808.

Evans, R. M., 1988. The steroid and thyroid hormone receptor superfamily. *Science*, 240(4854):889–895.

Fahy, E., Subramaniam, S., Brown, H. A., Glass, C. K., Merrill, A. H., Murphy, R. C., Raetz, C. R. H., Russell, D. W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M. S., White, S. H., Witztum, J. L. and Dennis, E. A., 2005. A comprehensive classification system for lipids. *J Lipid Res*, 46(5):839–861.

Fairclough, S. R., Dayel, M. J. and King, N., 2010. Multicellular development in a choanoflagellate. *Curr Biol*, 20(20):R875–R876.

Faus, H. and Haendler, B., 2006. Post-translational modifications of steroid receptors. *Biomed Pharmacother*, 60(9):520–528.

Feltgen, K., 1993. *Le cholesterol : 1758-1913. Essai historique sur l'interet qu'il a suscité en medecine depuis sa decouverte au milieu du XVIIIe siècle jusqu'a l'aube du XXeme siecle*. Ph.D. thesis, Faculté mixte de médecine et de pharmacie de Rouen.

Folkertsma, S., van Noort, P. I., Brandt, R. F. J., Bettler, E., Vriend, G. and de Vlieg, J., 2005. The nuclear receptor ligand-binding domain: a family-based structure analysis. *Curr Med Chem*, 12(9):1001–1016.

Gauhar, Z., Sun, L. V., Hua, S., Mason, C. E., Fuchs, F., Li, T.-R., Boutros, M. and White, K. P., 2009. Genomic mapping of binding regions for the ecdysone receptor protein complex. *Genome Res*, 19(6):1006–1013.

Giguère, V., Yang, N., Segui, P. and Evans, R. M., 1988. Identification of a new class of steroid hormone receptors. *Nature*, 331(6151):91–94.

Gotoh, S., Ohgari, Y., Nakamura, T., Osumi, T. and Taketani, S., 2008. Heme-binding to the nuclear receptor retinoid X receptor alpha (rxralpha) leads to the inhibition of the transcriptional activity. *Gene*, 423(2):207–214.

Gupta, R. P., Patrick, K. and Bell, N. H., 2007. Mutational analysis of cyp27a1: assessment of 27-hydroxylation of cholesterol and 25-hydroxylation of vitamin d. *Metabolism*, 56(9):1248–1255.

Hahn, M. E., 2002. Aryl hydrocarbon receptors: diversity and evolution. *Chem Biol Interact*, 141(1-2):131–160.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M. and Lake, J. A., 1995. Evidence from 18s ribosomal dna that the lophophorates are protostome animals. *Science*, 267(5204):1641–1643.

Hammel, J. U., Herzen, J., Beckmann, F. and Nickel, M., 2009. Sponge budding is a spatiotemporal morphological patterning process: Insights from synchrotron radiation-based x-ray microtomography into the asexual reproduction of *Tethya wilhelma*. *Front Zool*, 6:19.

Hampshire, F. and Horn, D., 1966. Structure of crustecdysone, a crustacean moulting hormone. *J Chem Soc Chem Commun*, pages 37–38.

He, J., Cheng, Q. and Xie, W., 2010. Minireview: Nuclear receptor-controlled steroid hormone synthesis and metabolism. *Mol Endocrinol*, 24(1):11–21.

Henley, D. V. and Korach, K. S., 2006. Endocrine-disrupting chemicals use distinct mechanisms of action to modulate endocrine system function. *Endocrinology*, 147(6 Suppl):S25–S32.

Hennig, W., 1950. *Grundzüger einer Theorie der Phylogenetischen Systematik*. Deutscher Zentralverlag, Berlin.

Hennig, W., 1966. *Phylogenetic Systematic*. University of Illinois Press, Urbana and Chicago, IL.

Holick, M. F., 2003. Vitamin d: A millenium perspective. *J Cell Biochem*, 88(2):296–307.

Horner, M. A., Pardee, K., Liu, S., King-Jones, K., Lajoie, G., Edwards, A., Krause, H. M. and Thummel, C. S., 2009. The *Drosophila* DHR96 nuclear receptor binds cholesterol and regulates cholesterol homeostasis. *Genes Dev*, 23(23):2711–2716.

Howard-Ashby, M., Materna, S. C., Brown, C. T., Chen, L., Cameron, R. A. and Davidson, E. H., 2006. Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. *Dev Biol*, 300(1):90–107.

Huang, P., Chandra, V. and Rastinejad, F., 2010. Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. *Annu Rev Physiol*, 72:247–272.

Huang, X., Warren, J. T. and Gilbert, L. I., 2008. New players in the regulation of ecdysone biosynthesis. *J Genet Genomics*, 35(1):1–10.

Iwema, T., Billas, I. M. L., Beck, Y., Bonneton, F., Nierengarten, H., Chaumot, A., Richards, G., Laudet, V. and Moras, D., 2007. Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. *EMBO J*, 26(16):3770–3782.

Jacob, F. and Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356.

Jandacek, R. J. and Tso, P., 2001. Factors affecting the storage and excretion of toxic lipophilic xenobiotics. *Lipids*, 36(12):1289–1305.

Jenner, R. A., 2000. Evolution of animal body plans: the role of metazoan phylogeny at the interface between pattern and process. *Evol Dev*, 2(4):208–221.

Keay, J., Bridgham, J. T. and Thornton, J. W., 2006. The *Octopus vulgaris* estrogen receptor is a constitutive transcriptional activator: evolutionary and functional implications. *Endocrinology*, 147(8):3861–3869.

Keay, J. and Thornton, J. W., 2009. Hormone-activated estrogen receptors in annelid invertebrates: implications for evolution and endocrine disruption. *Endocrinology*, 150(4):1731–1738.

Khersonsky, O., Roodveldt, C. and Tawfik, D. S., 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol*, 10(5):498–508.

Khorasanizadeh, S. and Rastinejad, F., 2001. Nuclear-receptor interactions on dna-response elements. *Trends Biochem Sci*, 26(6):384–390.

Kim, T.-W. and Wang, Z.-Y., 2010. Brassinosteroid signal transduction from receptor kinases to transcription factors. *Annu Rev Plant Biol*, 61:681–704.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J. B., Morris, A., Nichols, S., Richter, D. J., Salamov, A., Sequencing, J. G. I., Bork, P., Lim, W. A., Manning, G., Miller, W. T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I. V. and Rokhsar, D., 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, 451(7180):783–788.

King-Jones, K., Horner, M. A., Lam, G. and Thummel, C. S., 2006. The DHR96 nuclear receptor regulates xenobiotic responses in *Drosophila*. *Cell Metab*, 4(1):37–48.

King-Jones, K. and Thummel, C. S., 2005. Nuclear receptors–a perspective from drosophila. *Nat Rev Genet*, 6(4):311–323.

Kodner, R. B., Summons, R. E., Pearson, A., King, N. and Knoll, A. H., 2008. Sterols in a unicellular relative of the metazoans. *Proc Natl Acad Sci U S A*, 105(29):9897–9902.

Koelle, M. R., Talbot, W. S., Segraves, W. A., Bender, M. T., Cherbas, P. and Hogness, D. S., 1991. The *Drosophila* EcR gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell*, 67(1):59–77.

Krylova, I. N., Sablin, E. P., Moore, J., Xu, R. X., Waitt, G. M., MacKay, J. A., Juzumiene, D., Bynum, J. M., Madauss, K., Montana, V., Lebedeva, L., Suzawa, M., Williams, J. D., Williams, S. P., Guy, R. K., Thornton, J. W., Fletterick, R. J., Willson, T. M. and Ingraham, H. A., 2005. Structural analyses reveal phosphatidyl inositols as ligands for the NR5 orphan receptors SF-1 and LRH-1. *Cell*, 120(3):343–355.

Lafont, R., Dauphin-Villemant, C., Warren, J. and Rees, H., 2005. Ecdysteroid chemistry and biochemistry. *Comprehensive Molecular Insect Science*, 3:125–195.

Lafont, R. and Mathieu, M., 2007. Steroids in aquatic invertebrates. *Ecotoxicology*, 16(1):109–130.

Lankester, E. R., 1870. On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Ann Mag Nat Hist*, 4:34–43.

Larroux, C., Fahey, B., Liubicich, D., Hinman, V. F., Gauthier, M., Gongora, M., Green, K., Wörheide, G., Leys, S. P. and Degnan, B. M., 2006. Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evol Dev*, 8(2):150–173.

Laudet, V., 1997. Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. *J Mol Endocrinol*, 19(3):207–226.

Laudet, V. and Adelmant, G., 1995. Nuclear receptors. lonesome orphans. *Curr Biol*, 5(2):124–127.

Laudet, V., Hänni, C., Coll, J., Catzeflis, F. and Stéhelin, D., 1992. Evolution of the nuclear receptor gene superfamily. *EMBO J*, 11(3):1003–1013.

Leadbeater, B. S. and Karpov, S. A., 2000. Cyst formation in a freshwater strain of the choanoflagellate *Desmarella moniliformis* kent. *J Eukaryot Microbiol*, 47(5):433–439.

Lindblom, T. H., Pierce, G. J. and Sluder, A. E., 2001. A *C. elegans* orphan nuclear receptor contributes to xenobiotic resistance. *Curr Biol*, 11(11):864–868.

Lokman, P. M., Harris, B., Kusakabe, M., Kime, D. E., Schulz, R. W., Adachi, S. and Young, G., 2002. 11-oxygenated androgens in female teleosts: prevalence, abundance, and life history implications. *Gen Comp Endocrinol*, 129(1):1–12.

Makishima, M., Lu, T. T., Xie, W., Whitfield, G. K., Domoto, H., Evans, R. M., Haussler, M. R. and Mangelsdorf, D. J., 2002. Vitamin d receptor as an intestinal bile acid sensor. *Science*, 296(5571):1313–1316.

Markov, G., Lecointre, G., Demeneix, B. and Laudet, V., 2008a. The ”street light syndrome”, or how protein taxonomy can bias experimental manipulations. *Bioessays*, 30(4):349–357.

Markov, G. V., Paris, M., Bertrand, S. and Laudet, V., 2008b. The evolution of the ligand/receptor couple: A long road from comparative endocrinology to comparative genomics. *Mol Cell Endocrinol*, 293(1-2):5–16.

Markov, G. V., Tavares, R., Dauphin-Villemant, C., Demeneix, B. A., Baker, M. E. and Laudet, V., 2009. Independent elaboration of steroid hormone signaling pathways in metazoans. *Proc Natl Acad Sci U S A*, 106(29):11913–11918.

Mic, F. A., Molotkov, A., Benbrook, D. M. and Duester, G., 2003. Retinoid activation of retinoic acid receptor but not retinoid x receptor is sufficient to rescue lethal defect in retinoic acid synthesis. *Proc Natl Acad Sci U S A*, 100(12):7135–7140.

Miller, A. E. M. and Heyland, A., 2010. Endocrine interactions between plants and animals: Implications of exogenous hormone sources for the evolution of hormone signaling. *Gen Comp Endocrinol*, 166(3):455–461.

Mizuta, T., Asahina, K., Suzuki, M. and Kubokawa, K., 2008. In vitro conversion of sex steroids and expression of sex steroidogenic enzyme genes in amphioxus ovary. *J Exp Zool A Ecol Genet Physiol*, 309(2):83–93.

Mizuta, T. and Kubokawa, K., 2007. Presence of sex steroids and cytochrome P450 genes in amphioxus. *Endocrinology*, 148(8):3554–3565.

Moore, D. D., 1990. Diversity and unity in the nuclear hormone receptors: a terpenoid receptor superfamily. *New Biol*, 2(1):100–105.

Morange, M., 2011. What will result from the interaction between functional and evolutionary biology? *Stud Hist Philos Biol Biomed Sci*, 42(1):69–74.

Motola, D. L., Cummins, C. L., Rottiers, V., Sharma, K. K., Li, T., Li, Y., Suino-Powell, K., Xu, H. E., Auchus, R. J., Antebi, A. and Mangelsdorf, D. J., 2006. Identification of ligands for DAF-12 that govern dauer formation and reproduction in *C. elegans*. *Cell*, 124(6):1209–1223.

Mukaiyama, Y., Uchida, T., Sato, E., Sasaki, A., Sato, Y., Igarashi, J., Kurokawa, H., Sagami, I., Kitagawa, T. and Shimizu, T., 2006. Spectroscopic and DNA-binding characterization of the isolated heme-bound basic helix-loop-helix-PAS-A domain of neuronal PAS protein 2 (NPAS2), a transcription activator protein associated with circadian rhythms. *FEBS J*, 273(11):2528–2539.

Necela, B. M. and Cidlowski, J. A., 2004. Mechanisms of glucocorticoid receptor action in noninflammatory and inflammatory cells. *Proc Am Thorac Soc*, 1(3):239–246.

Norris, D. O., 2007. *Vertebrate endocrinology*. Elsevier Academic Press.

Nuclear Receptors Nomenclature Committee, 1999. A unified nomenclature system for the nuclear receptor superfamily. *Cell*, 97(2):161–163.

Näär, A. M. and Thakur, J. K., 2009. Nuclear receptor-like transcription factors in fungi. *Genes Dev*, 23(4):419–432.

O'Malley, B. W., 1989. Did eucaryotic steroid receptors evolve from intracrine gene regulators? *Endocrinology*, 125(3):1119–1120.

Ordóñez-Morán, P. and Muñoz, A., 2009. Nuclear receptors: genomic and non-genomic effects converge. *Cell Cycle*, 8(11):1675–1680.

Paris, M., Pettersson, K., Schubert, M., Bertrand, S., Pongratz, I., Escriva, H. and Laudet, V., 2008. An amphioxus orthologue of the estrogen receptor that does not bind estradiol: insights into estrogen receptor evolution. *BMC Evol Biol*, 8(1):219.

Parvy, J.-P., Blais, C., Bernard, F., Warren, J. T., Petryk, A., Gilbert, L. I., O'Connor, M. B. and Dauphin-Villemant, C., 2005. A role for betaFTZ-F1 in regulating ecdysteroid titers during post-embryonic development in *Drosophila melanogaster*. *Dev Biol*, 282(1):84–94.

Patel, D. S., Fang, L. L., Svy, D. K., Ruvkun, G. and Li, W., 2008. Genetic identification of hsd-1, a conserved steroidogenic enzyme that directs larval development in caenorhabditis elegans. *Development*, 135(13):2239–2249.

Payne, A. H. and Hales, D. B., 2004. Overview of steroidogenic enzymes in the pathway from cholesterol to active steroid hormones. *Endocr Rev*, 25(6):947–970.

Pearse, V. B. and Voigt, O., 2007. Field biology of placozoans (*Trichoplax*): distribution, diversity, biotic interactions. *Integrative and Comparative Biology*, 47:677–692.

Reitzel, A. M. and Tarrant, A. M., 2009. Nuclear receptor complement of the cnidarian *Nematostella vectensis*: phylogenetic relationships and developmental expression patterns. *BMC Evol Biol*, 9:230.

Reitzel, A. M. and Tarrant, A. M., 2010. Correlated evolution of androgen receptor and aromatase revisited. *Mol Biol Evol*.

Renard, E., Vacelet, J., Gazave, E., Lapébie, P., Borchiellini, C. and Ereskovsky, A. V., 2009. Origin of the neuro-sensory system: new and expected insights from sponges. *Integr Zool*, 4(3):294–308.

Reschly, E. J., Ai, N., Ekins, S., Welsh, W. J., Hagey, L. R., Hofmann, A. F. and Krasowski, M. D., 2008a. Evolution of the bile salt nuclear receptor FXR in vertebrates. *J Lipid Res*, 49(7):1577–1587.

Reschly, E. J., Ai, N., Welsh, W. J., Ekins, S., Hagey, L. R. and Krasowski, M. D., 2008b. Ligand specificity and evolution of liver X receptors. *J Steroid Biochem Mol Biol*, 110(1-2):83–94.

Revankar, C. M., Cimino, D. F., Sklar, L. A., Arterburn, J. B. and Prossnitz, E. R., 2005. A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science*, 307(5715):1625–1630.

Rochette-Egly, C., 2003. Nuclear receptors: integration of multiple signalling pathways through phosphorylation. *Cell Signal*, 15(4):355–366.

Rosenberg, S. M., 2009. Life, death, differentiation, and the multicellularity of bacteria. *PLoS Genet*, 5(3):e1000418.

Rottiers, V., Motola, D. L., Gerisch, B., Cummins, C. L., Nishiwaki, K., Mangelsdorf, D. J. and Antebi, A., 2006. Hormonal control of *C. elegans* dauer formation and life span by a rieske-like oxygenase. *Dev Cell*, 10(4):473–482.

Ruppert, E. E., 2005. Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can J Zool*, 83:8–23.

Russell, D. W., 2009. Fifty years of advances in bile acid synthesis and metabolism. *J Lipid Res*, 50 Suppl:S120–S125.

Schierwater, B., 2005. My favorite animal, *Trichoplax adhaerens*. *Bioessays*, 27(12):1294–1302.

Schubert, M., Brunet, F., Paris, M., Bertrand, S., Benoit, G. and Laudet, V., 2008. Nuclear hormone receptor signaling in amphioxus. *Dev Genes Evol*, 218(11-12):651–665.

Schug, T. T., Berry, D. C., Shaw, N. S., Travis, S. N. and Noy, N., 2007. Opposing effects of retinoic acid on cell growth result from alternate activation of two different nuclear receptors. *Cell*, 129(4):723–733.

Schwann, T., 1839. *Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachstum der Tiere und Pflanzen*.

Simionato, E., Ledent, V., Richards, G., Thomas-Chollier, M., Kerner, P., Coornaert, D., Degnan, B. M. and Vervoort, M., 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. *BMC Evol Biol*, 7:33.

Simpson, S., Tait, J., Wettstein, A., Neher, R., von Euw, J. and Reichstein, T., 1953. Isolierung eines neuen kristallisierten hormons aus nebennerien mit besonders hoher wirksamkeit auf den mineralsoffwechsel. *Experientia*, 9:333–335.

Sladek, F., 2002. Desperately seeking...something. *Mol Cell*, 10(2):219–221.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N. H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M. L., Signorovitch, A. Y., Moreno, M. A., Kamm, K., Grimwood, J., Schmutz, J., Shapiro, H., Grigoriev, I. V., Buss, L. W., Schierwater, B., Dellaporta, S. L. and Rokhsar, D. S., 2008. The *Trichoplax* genome and the nature of placozoans. *Nature*, 454(7207):955–960.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. A., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N. H., Stanke, M., Adamska, M., Darling, A., Degnan, S. M., Oakley, T. H., Plachetzki, D. C., Zhai, Y., Adamski, M., Calcino, A., Cummins, S. F., Goodstein, D. M., Harris, C., Jackson, D. J., Leys, S. P., Shu, S., Woodcroft, B. J., Vervoort, M., Kosik, K. S., Manning, G., Degnan, B. M. and Rokhsar, D. S., 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature*, 466(7307):720–726.

Stout, E. P., Clair, J. J. L., Snell, T. W., Shearer, T. L. and Kubanek, J., 2010. Conservation of progesterone hormone function in invertebrate reproduction. *Proc Natl Acad Sci U S A*, 107(26):11859–11864.

Summons, R. E., Bradley, A. S., Jahnke, L. L. and Waldbauer, J. R., 2006. Steroids, triterpenoids and molecular oxygen. *Philos Trans R Soc Lond B Biol Sci*, 361(1470):951–968.

Svácha, P., 1992. What are and what are not imaginal discs: reevaluation of some basic concepts (insecta, holometabola). *Dev Biol*, 154(1):101–117.

Theodosiou, M., Laudet, V. and Schubert, M., 2010. From carrot to clinic: an overview of the retinoic acid signaling pathway. *Cell Mol Life Sci*, 67(9):1423–1445.

Thomas, C., Auwerx, J. and Schoonjans, K., 2008. Bile acids and the membrane bile acid receptor tgr5–connecting nutrition and metabolism. *Thyroid*, 18(2):167–174.

Thomas, J. H., 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet*, 3(5):e67.

Thomas, P., Alyea, R., Pang, Y., Peyton, C., Dong, J. and Berg, A. H., 2010. Conserved estrogen binding and signaling functions of the G protein-coupled estrogen receptor 1 (GPER) in mammals and fish. *Steroids*, 75(8-9):595–602.

Thornton, J. W., 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci U S A*, 98(10):5671–5676.

Thornton, J. W., Need, E. and Crews, D., 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, 301(5640):1714–1717.

Thummel, C. S., 2001. Molecular mechanisms of developmental timing in c. elegans and drosophila. *Dev Cell*, 1(4):453–465.

Twan, W.-H., Hwang, J.-S. and Chang, C.-F., 2003. Sex steroids in scleractinian coral, *Euphyllia ancora*: implication in mass spawning. *Biol Reprod*, 68(6):2255–2260.

Tzertzinis, G., Egaña, A. L., Palli, S. R., Robinson-Rechavi, M., Gissendanner, C. R., Liu, C., Unnasch, T. R. and Maina, C. V., 2010. Molecular evidence for a functional ecdysone signaling system in *Brugia malayi*. *PLoS Negl Trop Dis*, 4(3):e625.

Umesono, K. and Evans, R. M., 1989. Determinants of target gene specificity for steroid/thyroid hormone receptors. *Cell*, 57(7):1139–1146.

Vogt, C., 1847. *Physiologische Briefe für Gebildete aller Stände*. J. G. Cotta'ischer Verlag, Stuttgart und Tübingen.

Wada, T., Gao, J. and Xie, W., 2009. PXR and CAR in energy metabolism. *Trends Endocrinol Metab*, 20(6):273–279.

Wang, Z., Benoit, G., Liu, J., Prasad, S., Aarnisalo, P., Liu, X., Xu, H., Walker, N. P. C. and Perlmann, T., 2003. Structure and function of nurr1 identifies a class of ligand-independent nuclear receptors. *Nature*, 423(6939):555–560.

Wang, Z., Zhou, X. E., Motola, D. L., Gao, X., Suino-Powell, K., Conneely, A., Ogata, C., Sharma, K. K., Auchus, R. J., Lok, J. B., Hawdon, J. M., Kliewer, S. A., Xu, H. E. and Mangelsdorf, D. J., 2009. Identification of the nuclear receptor daf-12 as a therapeutic target in parasitic nematodes. *Proc Natl Acad Sci U S A*, 106(23):9138–9143.

Whitfield, G. K., Dang, H. T. L., Schluter, S. F., Bernstein, R. M., Bunag, T., Manzon, L. A., Hsieh, G., Dominguez, C. E., Youson, J. H., Haussler, M. R. and Marchalonis, J. J., 2003. Cloning of a functional vitamin D receptor from the lamprey (*Petromyzon marinus*), an ancient vertebrate lacking a calcified skeleton and teeth. *Endocrinology*, 144(6):2704–2716.

Wiens, M., Batel, R., Korzhev, M. and Müller, W. E. G., 2003. Retinoid X receptor and retinoic acid response in the marine sponge *Suberites domuncula*. *J Exp Biol*, 206(Pt 18):3261–3271.

Yuan, X., Ta, T. C., Lin, M., Evans, J. R., Dong, Y., Bolotin, E., Sherman, M. A., Forman, B. M. and Sladek, F. M., 2009. Identification of an endogenous ligand bound to a native orphan nuclear receptor. *PLoS One*, 4(5):e5609.

Zimmerman, A. W. and Veerkamp, J. H., 2002. New insights into the structure and function of fatty acid-binding proteins. *Cell Mol Life Sci*, 59(7):1096–1116.

Part VIII

Appendix

Additional articles

Here is given a brief description of some additional articles in which I participated.

- On page 154, an article published in *Bioessays* analyses the criteria taken into account for protein naming, and pinpoint how a misleading name can lead to biases in functional experiments (Markov et al., 2008).
- On page 163, an article published in *Development, Genes and Evolution* analyses the degree of conservation in the florida lancelet *Branchiostoma floridae* of the genes encoding proteins that are involved in the thyroid hormone metabolizing pathways. The thyroid axis is an endocrine axis that acts in parallel, with some crosstalks, with the steroid endocrine axes. Here the analysis showed that amphioxus has undergone specific gene duplications and is thus not only a simple "basal relative" to vertebrates (Paris et al., 2008).
- On page 177, an article published in *Nature Biotechnology* reports the annotation of the genome of the root-knot nematode *Meloidogyne incognita*. Here I performed the annotation of the nuclear receptor family, in order to caracterise more precisely the tempo and mode of lineage-specific expansion of nuclear receptors from the NR2A family in nematodes. I showed that the lineage-specific expansion of this nuclear receptor group began before the divergence between three of the five nematode bug clade and proceeded independently in each of them (Abad et al., 2008).
- On page 184, an article published in *Molecular and Cellular Endocrinology* reviews the caveats linked to comparative methods in endocrinology and advocated for a shift to comparative genomics. In particular, we discuss in detail the question of vertebrate-type steroids reported in mollusks and annelids (Markov et al., 2008).
- On page 196, a chapter published in the book *Nuclear Receptors: current concepts and future challenges* reviews the evolution of nuclear receptors, with some discussions about the existence of NR-like ligand-activated transcription factors among other living beings than animals (Markov et al., 2010).
- On page 211, an article published in *Molecular and Cellular Endocrinology* reviews more specifically the evolution of the ligand-binding ability of nuclear receptors (Markov et al., 2011). The beginning of this article was used as a backbone for the introduction section of on evolution of ligand-binding ability in this PhD thesis.
- On page 221, and article currently in press for *Molecular Biology and Evolution* I checked and improved the alignements and the phylogenetical sampling of chordate RARs and I reconstructed ancestral sequences in a study about the evolution of phosphorylation sites in vertebrate RARα(Samarut et al., *in press*).
- On page 260, an article that was rejected by *Current Biol* after reviewing, and is currently in revision before resubmission, addresses the issue of the convergent evolution of the *nanog* gene in mammals and the vent genes in amphibians. Nanog is a major regulator of pluripotency in mammalian stem cells and a target gene for some nuclear receptors. Vent is hypothesized to perform a similar role in amphibians. I contributed to the molecular phylogeny of the *nanog* and *vent* gene families and I performed positive selection mesures (Scerbo et al., *in preparation*).

The "street light syndrome", or how protein taxonomy can bias experimental manipulations

Gabriel Markov,[1,2] Guillaume Lecointre,[3] Barbara Demeneix,[1] and Vincent Laudet[2]*

Summary

In the genomics era, bioinformatic analysis, especially in non-model species, facilitates the identification and naming of numerous new proteins, the function of which is then inferred through homology searches. Here, we question certain aspects of these approaches. What are the criteria that permit such a determination? What are their limits? Naming is classifying. We review the different criteria that are used to name a protein and discuss their constraints. We observe that the name given to a protein often introduces a bias for further functional analyses, a bias that is not often taken into account when analysing results. Last but not least, the heterogeneity of criteria used for naming proteins leads to self-inconsistent or contradictory protein classification that is potentially misleading. Finally, we recommend a wider use of phylogenetic criteria in protein naming.

***BioEssays* 30:349–357, 2008.** © **2008 Wiley Periodicals, Inc.**

[1]USM 501, Evolution des Régulations Endocriniennes. Muséum National d'Histoire Naturelle, Paris, France.
[2]Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, Molecular Zoology team, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland Lyon Sud, France.
[3]UMR 7138 CNRS-UPMC-IRD-MNHN-ENS CP26 Département Systématique et Evolution, Muséum National d'Histoire Naturelle, Paris, France.
Funding agencies: We are grateful to Ecole Normale Supérieure de Lyon, Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique and the Ministère de l'Education Nationale, de la Recherche et de la Technologie for financial support. V.L. and B.D. laboratories are supported by the Cascade EU Network of Excellence.
*Correspondence to: Vincent Laudet, UMR 5242 du CNRS, Institut de Génomique Fonctionnelle de Lyon, Equipe de Zoologie Moléculaire. Université de Lyon, INRA IFR 128 BioSciences Lyon-Gerland, Ecole Normale Supérieure de Lyon 46, allée d'Italie 69364 Lyon Cedex 07, France. E-mail: vincent.laudet@ens-lyon.fr
DOI 10.1002/bies.20730
Published online in Wiley InterScience (www.interscience.wiley.com).

Abbreviations: AR, androgen receptor; CRABP, cellular retinoic acid binding protein; DHA, docosahexaenoic acid; ER, estrogen receptor; ERR, estrogen related receptor; HSD17B, 17-β hydroxysteroid dehydrogenase; GR, glucocorticoid receptor; IkB, inhibitor of kappa B; MR, mineralocorticoid receptor; PR, progesterone receptor; RA, retinoic acid; RXR, retinoic X receptor; SDR, short-chain dehydrogenase/reductase; TPO, thyroperoxidase; USP, ultraspiracle.

Introduction

Among the many steps in describing a new protein, one that could be considered as trivial is its naming. The importance of this step seems to have been underestimated, as many examples show that giving a name to a protein is not neutral. Names refer to concepts and, therefore, could have a major influence on further experimental efforts.

In the pre-genomics era, when genes and proteins were isolated in order to understand the molecular basis of a phenotypic feature, the name chosen was often linked with the approach used to isolate the protein (a point discussed later). DNA probes were designed on the basis of known sequences, and used to search for new sequences hypothesised to be related enough to hybridize with the probe.

In the genomics era, the problem took on a new dimension with the increasing number of available nucleotidic sequences and the progress in prediction algorithms, which resulted in more and more new proteins, especially in non-model species, being predicted every day through bioinformatic analysis, with their function duly inferred by homology searches. Protein annotation is a two-step process (reviewed in Refs 1,2). First, there is a structural annotation step, in which the corresponding DNA sequence is checked for the presence of start and stop codons, splicing sites and other features that permit determination of the coding sequence. This step is now quite well automated (reviewed in Ref. 3), even if all prediction programs need to be refined, especially when working with sequences from non-model species, where the gene structure and splicing mechanisms can vary. The second step is the functional annotation of predicted protein products. This process is increasingly carried by automatic tools, which are very helpful for a quick description of large datasets, but are prone to artefacts and, in particular, can lead to propagation of annotation errors. The process has been discussed and reviewed by Valencia,[4] who argued for a reliability score assignment to sequence annotation in order to facilitate a critical appraisal of the information.

Problems and paradigms

Any language needs concepts, which can be defined as classes of objects. A class is a set of assembled objects sharing a special property.[5] Any name of general use is primarily attached to a class, not to the object itself. For instance biologists do not need to name individual molecules; therefore any protein name actually refers to a class of material entities. When a new protein is described, the name refers not only to the molecules that were in the test tube of the describer, but also to all the molecules sharing the same amino acid sequence that could be found in the body of the animal from which it was purified, and even in the body of other animals of the same species. Giving a name to an object is therefore dealing with classifications: it is assigning the particular object to a set containing other objects sharing common properties (i.e. a class or a concept). These properties are always arbitrary, but problems arise when different properties are used to create non-overlapping kinds of concepts dealing with the same objects. Risks and imprecision ensue when using words in the wrong conceptual framework. For example ''algae'' is an ecological concept. It covers all living things having photosynthetic activity in aquatic environments. Using this term in a phylogenetic classification is potentially misleading. Errors result when we wrongly identify the nature of the concepts that we are using. If one selects ''algae'' to compare their DNA sequences while expecting homogeneity of ''algal'' sequences, one would be surprised to find phaeophycean sequences (brown algae) more similar to ciliate sequences than to green algae sequences.[6] Green algae sequences are more similar to those of land plants. Following on from these ideas, the goal of the present paper is to highlight the heterogeneity of properties chosen to create the concepts used for protein naming. This heterogeneity constrains experimental possibilities and creates misunderstandings: names are chosen according to various criteria and they are later wrongly understood as names referring to structure and origin (i.e. names given using phylogeny), as they should in any comparative approach in biology. In this regard, current protein-naming approaches are not based on a self-consistent classification of proteins.

Here we will review the different criteria that are used to name a protein, in order to pinpoint the limits of each of them. Then we will study how these names can influence further experiments, and we will finally discuss how such bias might be corrected.

Different types of names, but a common definition problem

Proteins are given different kinds of names, i.e. their names refer to heterogeneous concepts, depending on the approach used to isolate them. These different types are summarized in Table 1.

Many proteins studied by traditional approaches were given functional names (for the different meanings of the word ''function'' in this paper, see Box 1): when a protein was purified in order to understand the molecular basis of a given biological function, the name often referred to this function. The reference to this function often supposes that a particular organ exists in the animal having this protein. In some cases, this leads to anomalous situations. Many recent papers still refer to prolactin in teleosts (see Ref. 7 for an example), even if this protein could of course not stimulate lactation in species that do not have a mammary gland. In fact, in teleosts, prolactin is involved in osmoregulation and this probably represents the ancestral function of this protein. An interesting example of trying to correct such inaccuracies of nomenclature is the case of the WNT proteins. The first *wnt* gene was identified as a proto-oncogene, activated in response to proviral insertion of a mouse mammary tumour virus, and was named *int-1*. In *Drosophila*, where its mutation led to a wingless phenotype, it was named ''*Wingless*''. Later it was recognized that *int-1* and ''*Wingless*'' were homologous and that they belong to a large family of related glycoproteins. Since a wingless phenotype is nonsense in mammals, in order to simplify the nomenclature,

Table 1. Summary of the different types of names found in protein nomenclature

Notion alluded to in the name	Example
Biological function	Prolactine
Localisation in an organ	TPO (ThyroPerOxidase)
Mutant phenotype	USP (UltraSPiracle)
Ligand binding ability	RXR (Retinoid X Receptor)
Presence of a conserved domain	HOXB1
Position in a gene cluster	HOXB1, HOXC4
Biochemical function	HSD17B (17-beta HydroxySteroid Dehydrogenase)

The cited examples are discussed in text.

Box 1. Different meanings of the word ''function''.

The use of the word ''function'' can be confusing in protein biology, because it refers to different things. For the purposes of clarity, in this study, we will distinguish three kinds of ''function''.

Biochemical activity refers to the kind of biochemical reaction made by the protein, e.g. dehydrogenation, for an enzyme, or to the type of modification undergone by the protein, e.g. ligand binding for a receptor.

Biochemical function is the reaction made in-vivo by the protein, e.g. dehydrogenation of a 17-beta steroid, binding of retinoic acid.

Biological function is the function in which the protein is involved at organism level, e.g. steroid biosynthesis, metamorphosis.

the whole family was renamed Wnt, a amalgam of Wingless and Int (reviewed in Ref. 8). In other words, the statement that homologous proteins were given names referring to two different frameworks (developmental for *Drosophila*, oncogenetic for mouse) lead to the formation of a framework-neutral name, suitable for both proteins.

When the differences between species are not so great, the given name sometimes refers to a supposed homology between two different species. Defining the peroxidase expressed in the endostyle of *Branchiostoma belcheri* as BbTPO (for *Branchiostoma belcheri* ThyroPerOxidase) suggests that amphioxus endostyle is homologous to vertebrate thyroid. For some authors,[9] the restriction of the expression of BbTTF-1 (thyroid transcription factor-1) and BbTPO to the endostyle strongly suggests that the endostyle is homologous to the follicles of the thyroid gland, whereas the traditional hypothesis is that the whole thyroid gland is homologous to the endostyle.[10] Thus, because BbTPO is considered a thyroperoxidase, its expression pattern provided evidence for the homology between two organs. In this case, the use of the same name is misleading, and even dangerous, because it favours a conclusion that has not yet been adequately tested. The name "Endostyle Peroxidase" would be more neutral in this case.

In the case of developmental genes, the name often refers to the associated mutant phenotype. For example, the name of the nuclear receptor USP was coined from the *ultraspiracle* phenotype, referring to the extra set of spiracles observed on larvae harbouring a loss-of-function mutation in this gene,[11] whereas the orthologous gene was named RXR for "Retinoic X-Receptor" in mammals, referring to its ability to bind retinoids.[12] In basal insects (excluding Diptera and Lepidoptera), the orthologous protein sequence was surprisingly more similar to vertebrate RXR than to *Drosophila* USP. Therefore, the homologous protein of these insects was given the name USP-RXR, even in the absence of any ultraspiracle mutant and even if the receptor was recently shown not to bind retinoid in vivo.[13] Purely, in terms of gene function, a third name would have been preferable, but the authors chose simplification and phylogeny as a guide for nomenclature.

Many protein names derive from the presence of a particular conserved domain. In vertebrates, the name of the famous HOX proteins refers to the existence of a conserved DNA-binding homeodomain, which in turn refers to homeosis (i.e. the transformation of one body part into another), observed when those genes are mutated. But this domain also exists in other proteins, like the gap gene products EMS or OTX in mammals, which are not called "HOX" because this name is reserved for proteins whose gene is located within a *hox* cluster.[14] In some cases, this definition is quite arbitrary: two *Evx* genes are located on the 5′-end of the mammalian *HoxA* and *HoxD* clusters, but *Evx* are not included in the *hox* genes category because its *Drosophila* ortholog, *Even-skipped* (*Eve*) is not located within a *hox* cluster.[14] Since the *hox* gene cluster in *Drosophila* appears quite derived when compared to those of other metazoans, this exclusion of *Evx* from the bona fide *hox* cluster may be considered as arbitrary.

The situation is much more complicated with protein names referring to biochemical functions. For example, the name 17-β hydroxysteroid dehydrogenase (HSD17B) is used for many proteins that are supposed to dehydrogenate 17-β steroids, an important step in steroid hormone biosynthesis. But, in fact, this name describes proteins with very different activities (reviewed in Ref. 15). Another problem is that all the proteins named as HSD17B are not members of the same protein family: the HSD17B5 is a member of the aldoketoreductase (AKR) protein family while the rest of the known HSD17B belong to the short-chain dehydrogenase/reductase (SDR) protein family. Even within the SDR family, the HSD17B-activity seems to have arisen several times independently[15] (see Fig. 1).

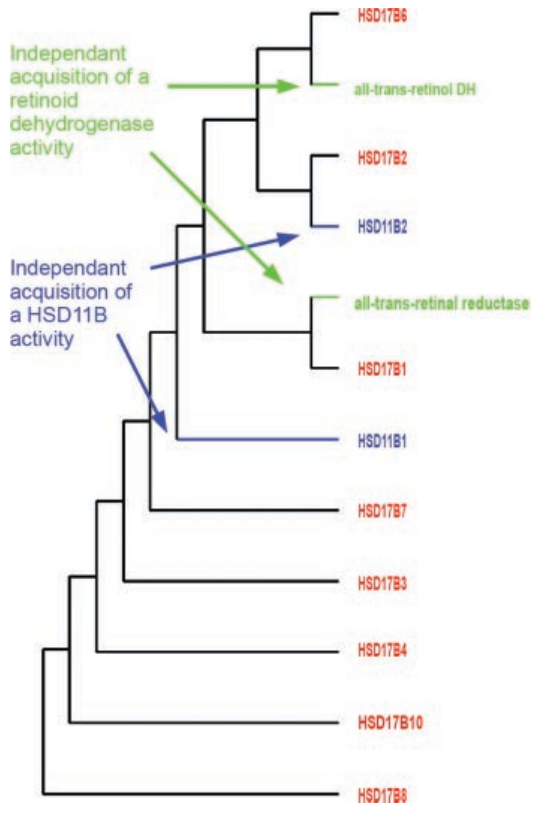


Figure 1. Convergent acquisition of biochemical functions within the SDR family in human. 11-beta-HSD activity and retinoid dehydrogenase activity seems to have appeared twice independantly. Because of the many other members of the family, not show, in that tree, even the HSD17B activity probably arose many times. Modified after Ref. 15.

These proteins often show multi-substrate activities, some of them being able to dehydrogenate either 17-β steroids or retinoids, with the HSD17B-activity sometimes only being established *in vitro* (reviewed in Ref. 16). The result is that the "HSD17B" family is paraphyletic, and even contains members with no 17-β hydroxysteroid dehydrogenase activity, leading to a very puzzling situation, since all these enzymes share the same identification code in the enzyme database (EC 1.1.1.51). This may also indicate that, in certain cases, the experimental efforts addressing the substrate specificity of the protein may have been misguided—or seen in a too narrow a fashion - by the gene name, as we discuss below. This provides a typical example of overlapping frameworks: the concept "HSD17B" was first used to describe a biochemical activity observed in vivo. But when the proteins presumed to be responsible for this activity were isolated, the word "HSD17B" was reused to name them, and was later used to name proteins showing an in vitro HSD17B-activity, whereas there was no evidence for their in vivo activity. So the same name "HSD17B" is used to describe two different concepts that partially overlap. The set "HSD17B" has a biochemical functional meaning. However, it is composed of entities that do not exhibit the biochemical functions in vivo because another framework, the presence of an in vitro HSD17B-activity, has led to their collation as a set. Such problems sometimes appear even in the title of the characterisation paper. For instance, the title "*Expression cloning and characterization of human 17 beta-hydroxysteroid dehydrogenase type 2, a microsomal enzyme possessing 20 alpha-hydroxysteroid dehydrogenase activity*."[17] clearly indicates that assigning the name of a protein with referring to a biochemical activity lead to some overlaps, even at the biochemical level.

Furthermore, even when the biological function is conserved between two orthologous proteins, they can have radically different biochemical functions. The nuclear receptor USP-RXR is a transcription factor that is activated by the transient binding of small fatty acids in deuterostomes and molluscs. It has lost its ligand-binding pocket in insects, becoming an orphan receptor in most of the arthropods. But in Mecopterida, the crown insect group containing Diptera and Lepidoptera, the gain of a large ligand-binding pocket allows the binding of structural ligand.[18] In this case, fatty acids are constantly present in the ligand-binding pocket.[19] So even if the biological function of transcription factor, acting as a dimere with another nuclear receptor, is conserved among bilaterians, the biochemical function, here the ligand-binding abilities, are very different between Mecopterida, other insects and other bilaterians (Fig. 2). This difference cannot be over-emphasised, because the ability to bind a ligand transiently allows fine-tuning of gene expression, and this will depend on the availability of ligand.

To sum up, protein nomenclature is very heterogeneous, often depending on historical circumstances and organism-linked specificities. For multigenic families, this often leads to confusing situations, so unified nomenclature systems have been established, specific for each gene family, to rationalize the system: *CYP* genes,[20] *hox* genes,[14] voltage-gated ion channels,[21] and nuclear receptors[22] provide classical examples. An independent nomenclature system has also been developed for all enzymes (reviewed in Ref. 23), official nomenclature committees such as NC-IUPHAR exist for receptors used in pharmacology,[24] and general nomenclature principles have been defined on a whole-genome scale for man.[25] These nomenclatures have their limitations, especially those based upon one specific organism (human or fly), because the different naming systems do not facilitate cross-species comparisons, and sometimes increase confusion. For example, Nelson recently warned about this situation giving the example of the annotation of rat *Cyp2* genes, where the nomenclature was fully revised by a nomenclature committee to match orthologous mouse genes, but the rat genes had already been given official names that refer to human *CYP* genes that are not their orthologs.[26] Another field that has received little attention is the correction of identified errors. It has been shown that a great number of papers are not retracted due to the lack of post-publication curation, especially for journals with low Impact Factor and limited access.[27] Thus, it would be important to facilitate post-publication correction in name fields in Genbank and other frequently visited databases. This implies that, when a protein name is seen to be erroneous, other curators should be able to submit modifications. As for accession number, former names should be conserved to permit an easy retrieval of sequences

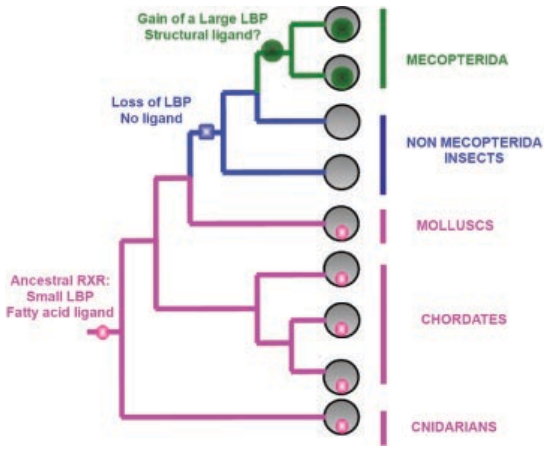


Figure 2. Function diversification in homologous proteins. The USP-RXR receptor binds small fatty-acid ligands in cnidarians, molluscs and chordates. It has lost its ligand-binding pocket (LBP) in insects and then gained a larger LBP in Mecopterida. Modified after Ref. 18.

referenced in old papers. Given the fact that such a work is time consuming, this curation activity should be taken into account both from a funding viewpoint and in the evaluation of researchers' activities. Securing an adequate funding for this often ignored, but critical, activity of curation will certainly have an important impact for the whole biology community since it could avoid controversies and experimental dead ends.

When protein names lead to experimental biases

One could argue that some names have only an historical signification and that their etymology has little importance for further studies. But selecting a name for a protein sometimes - and more often than expected—leads to experimental biases, simply because our experiments depend on the concepts we have in mind and sometimes these concepts are wrongly interpreted.

The estrogen-related receptor (ERR), was found using a low-stringency screen with a DNA probe corresponding to the gene region coding for the DNA-binding domain of human Estrogen Receptor alpha.[28] It was named ERR because of the high percent identity between its conserved domains (DNA- and ligand-binding domains) and the corresponding domains of estrogen receptor (ER). Given this proximity to ERs, its ability to bind estrogens was tested, and it was not possible to demonstrate binding with any major class of steroids.[28] It was later found that ERR-alpha can bind the endocrine disruptors toxaphene and chlordane[29] and that ERR beta and gamma are inhibited by 4-hydroxytamoxifen[30] even if the physiological relevance of these findings is still discussed.[31] But it has never been reported whether or not ERR could bind androgens, gluco- or mineralocorticoids, although new phylogenies suggest that ERR is not more closely related to ER than to other steroid receptors such as the glucocorticoid receptor (GR), mineralocorticoid receptor (MR), androgene receptor (AR) and progesterone receptor (PR).[32,33] The same bias appeared in studies on the biological functions of ERRs. Many features of ERRs (its DNA-binding site, its protein–protein interaction abilities, its implication in physiological pathways etc) were tested in the light of this apparent close relationship with ERs (see Ref. 34 for an example) and very little attention was paid to its possible links with steroid receptors even if we now know that ERRs are not more closely related to ERs than to other steroid receptors (Fig. 3). This bias clearly appears through a database search in PubMed abstract with the keywords "estrogen-receptor related" and other steroid receptor names. As of December 2006, only one abstract mentions together AR and ERR, two for PR and ERR, four for GR and ERR, none for MR and ERR, whereas 29 abstracts discussed relationships between ER and ERR, many of them with eloquent titles: "*Transcriptional targets shared by estrogen receptor- related receptors (ERRs) and estrogen receptor (ER) alpha, but not by ERbeta*",[34] "*The mouse estrogen receptor-related orphan receptor alpha 1: molecular cloning and estrogen responsiveness*",[35] "*Estrogen receptor-related receptors in the killifish Fundulus heteroclitus: diversity, expression, and estrogen responsiveness*".[36] This provides a striking example of a bias in the experimental effort, created by the use of a name derived from poorly resolved phylogeny.

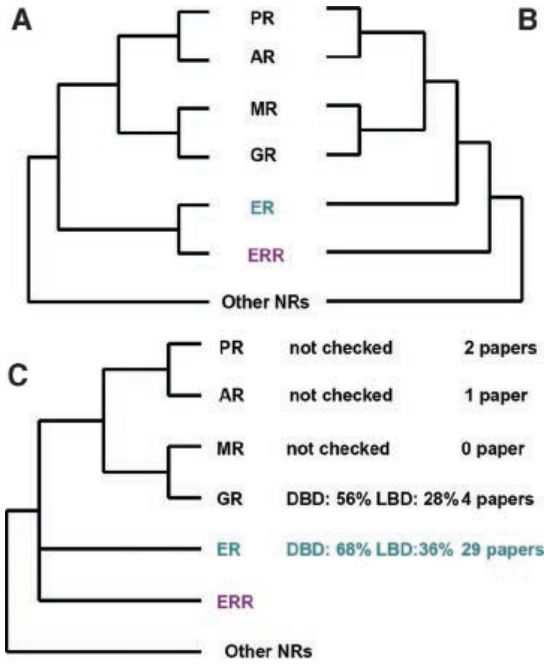


Figure 3. Protein naming and experimental effort. **A:** Phylogeny from Ref. 22, where ERR is closely related to ER, and thus explains why ERR was more often compared to ER than to other steroid receptors. **B:** Phylogeny from Ref. 32, showing that ER is not directly related to ERR, even if the last common ancestor of ER and other steroid receptors may have bound estrogens. **C:** Current consensus view of the phylogeny of ERR and other steroid receptors, showing that ERR is no more closely related to ER than to another SR. When provided by Giguere (Ref. 28), the percentages of amino-acid identity between the ligand-binding domain (LBD) and DNA-binding domain (DBD) of ERR and other receptors are shown. We also indicated the number of PubMed abstracts where comparisons are made between the ERR properties and those of other steroid receptors. There is a clear bias in favour of ER.

Such a bias also occurred in the initial studies on the nuclear receptor RXR. The name "retinoid X receptor" originally referred to its ability to bind vitamin A metabolites.[12] It was therefore supposed that RXR would be involved in a new retinoic acid (RA)-response pathway. Only recently, it was found that retinoids are apparently not *bona fide* natural ligand for RXR and that, in mouse brain, a fatty acid, the

docosahexaenoic acid (DHA) seems to be an endogenous RXR ligand.[37,38] But most of the papers studying RXR focus on its supposed involvement in RA-response pathway, whereas its involvement in fatty acid metabolism and signalling is much less studied, as shown in Table 2.

Another quite spectacular example is the case of the Cellular Retinoic Acid Binding Protein (CRABP) which are proteins implicated in retinoid signalling and related to FABP (Fatty Acid Binding Proteins). Retinoids are well known in vertebrates and their developmental role is well documented. But no defined retinoids have been isolated in arthropods and the existence of this signalling pathway in these organisms awaits clarification. The discovery of a protein that was supposed to be orthologous to vertebrate CRABP in the moth *Manduca sexta* was thus of interest. This first protein was cloned using a cDNA probe from a partial amino acid sequence of prothoracicotropic hormone that was similar to vertebrate retinoid-binding proteins. The newly isolated protein was annotated as CRABP in *Manduca sexta* on the basis of comparisons from percentage identities, without any real phylogenetic analysis.[39] The authors also proposed three-dimensional structures generated by homology-model building that showed the presence of an RA-binding pocket in the *Manduca* "CRABP". This protein was later used for the phylogenetic analysis of a newly cloned putative CRABP in the shrimp *Metapeneaeus ensis*.[40] The binding properties of the newly identified protein were also checked and it was concluded that the putative CRABP of *Metapenaeus* binds both RA and retinol, but not fatty acids (in fact the only fatty acid tested was parinaric acid). In 2005, a more exhaustive phylogenetic analysis of the family was performed, indicating that both sequences are members of the Fatty Acid Binding Protein (FABP) subfamily, a different subfamily, even if related to CRABP subfamily[41] (Fig. 4), showing that the genes coding for these insect proteins are not orthologous to the vertebrate CRABP. The binding abilities of *Manduca sexta* "CRABP" were also checked, showing that the *Manduca* "CRABP" has no significant affinity for RA and retinol, whereas it efficiently binds oleic acid and elaidic acid, and that the RA binding reported for *Metapenaeus ensis* seems due to experimental artefact.[41] This example shows that an excess of confidence in the result of an initial screen, together with a too-rapid phylogenetic analysis, can lead to a distorted experimental follow up. Moreover, the existence of the retinoid signalling pathway in arthropods remains elusive.

Ironically, within the same family, it should also be mentioned that a protein was annotated as CRABP in the amphioxus, *Branchiostoma floridae*[42] on the basis of a phylogeny using sequences from *Manduca* and *Metapenaeus* that initially annotated CRABP as an outgroup (Fig. 4). A more-detailed phylogenetic analysis (our unpublished data) suggests that this protein is probably neither a CRABP nor a CRBP but rather an IFAB (Intestinal Fatty Acid Binding Protein). This should be taken into account in further studies of its binding properties.

This example clearly shows that useful scientific papers can contain nomenclature errors. Such is the case from recent work on CRABP: the fact that the protein was not properly annotated makes not the crystal structure less interesting, even if the interpretation is different. It should be emphasized that the problem of protein naming is not only a problem of bad science, but that the use of functional naming is dangerous in itself. Other authors have already raised similar reservations and warnings on this subject (see for example Ref. 43 on bacterial *RecA*).

Table 2. Experimental effort about RXR

Key word searched	Number of abstracts registered in PubMed on January 3rd, 2007.
RXR Fatty acid(s)	197 (183)
RXR Glucocorticoids	25
RXR Retinoid(s)	1010 (1998)
RXR Steroids	362
RXR Thyroid hormones	176
RXR Vitamin D	315

The number of abstracts mentioning RXR with one of its putative ligand is reported. Even if recent experimental data seems to indicate that RXR is involved in fatty acid response, the number of abstracts mentioning RXR and Fatty acids together is quite low, compare to other metabolites, whereas the number of abstracts mentioning RXR and Retinoids is about ten times higher.

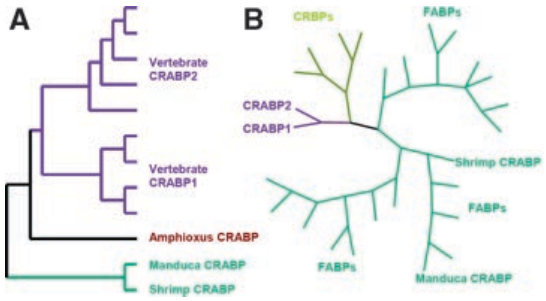


Figure 4. Influence of protein naming on phylogenetic sampling. **A:** Owing to the fact that some arthropod proteins were annotated as CRABP, Jackman et al. (Ref. 42) used them as an outgroup to check the phylogenetic position of a newly isolated amphioxus protein. **B:** Thus, Folli et al. (Ref. 41) showed that those proteins actually belong to the FABP subfamily, so these two proteins are not sufficient to indicate the position of the amphioxus sequence within the whole CRABP/CRBP/FABP family.

In conclusion, we need to emphasize that protein names with a functional significance are not neutral. They influence experimenters, and lead to what could be named the *street light syndrome:* as the joke goes, people tend to search for lost keys under the light of a lamp post, because it is more easy to search there than the actual place where the key was lost! Similarly, biochemists who study a known protein in a new species tend to check only if the protein has the functional properties that they suppose it should have on the basis of what is known about its function in other species, and not on the basis of protein sequence phylogenetic relationships.

One could wonder if such problems are also encountered in large-scale analyses, such as DNA arrays or EST analysis, where scientists have to rely on database annotations. Particularly in these cases, the name is generally not taken into account to infer protein functions. Orthology relationships are inferred from automated phylogeny pipelines, or filtered blast search results, and functional inferences are based upon Gene Ontology,[44] where indications about the reliability level of given information are indicated (for an example, see Ref. 45).

How can we avoid bias?

Even though functional names are insufficient to describe the real complexity of protein features, and can lead to experimental bias, manual correction of those errors represents a bottleneck, given the huge amount of new data. Moreover, powerful tools, such as Gene Ontology,[44] were developed specifically to provide functional information about a given protein, taking into account the reliability level of those data. So in future, it would be better to avoid functional naming of protein, in order to clearly distinguish between the name, which should be a constant tag used to refer to the studied object, and the description, which could evolve with the growing knowledge, and should provide the detailed features of the object. Such a distinction was already made between *signifiers* (i.e. names referring to an object) and *descriptors* (i.e. names referring to the known object properties), and some solutions have been proposed to avoid this confusion.[46] Basically, the proposal was to promote the use of gene names that do no refer to any describing features, either because they are only remotely connected to a mutation phenotype (for example *sonic hedgehog*), or because they come from languages other than English, and so have little signification for the majority of the research community (for example *fushi tarazu*). Generalising such a system could avoid many experimental biases but, however, will be of limited use for managing knowledge about big protein superfamilies. We want to emphasize that naming through a descriptor may not be problematic if the description method is universal and can be applied to all proteins. We propose that phylogeny, which is based on evolution, is an excellent tool to name genes with both accuracy and flexibility.

Indeed, as in systematics, object naming in molecular biology or in biochemistry has to manage structural knowledge. However, this is not enough. Molecular biology and biochemistry are sciences dealing with biological entities and phenomena; i.e. entities that vary through time and among populations and that have evolutionary histories. Therefore naming them according to phylogeny takes into account both structure and history, as clades are sets based on shared derived features. Biology has one general theory, evolution, and it seems appropriate and useful for all its sub-disciplines to take it into account.

Exhaustive phylogenetic analyses may also help to avoid many biases, and datasets should take into account not only the sister groups of the protein of interest, but also more distant families.

Too often, proteins are still annotated with BLAST,[47] which uses distance comparisons between sequences to make similarity scores. This implies that the newly identified protein is considered homologous to the most-similar protein available. The problem is that the first match is not necessarily the orthologous protein. The global similarity may not even indicate an orthology relationship. In multigenic families, one protein could be homologous to two different paralogous proteins in another species.[48] Fast-evolving proteins can also artefactually display great similarity simply because of random multiple substitutions at the same sites,[49,50] or because of similar composition biases.[51] Thus, only a careful phylogenetic analysis, taking into account these risks, can provide reliable information about the position of a new protein within a family. When trying to group different entities under the same name on the basis of their similarity, it is important to distinguish whether this similarity is a result of common ancestry of convergent evolution. Proteins are evolving entities, and undergo modifications of their features as a function of time, according to the diversification of the organisms to which they belong. Thus, an important concept is their evolutionary history, that is their relationships based on descent with modification and the inferred transformation events that they underwent.

Using phylogenetic taxonomy, the fact that some names do not fit the actual characteristics of all group members, due to functional shifts—at a biochemical or biological level—in different organisms, will not be too problematic (see Ref. 52 for an example how to detect these functional shifts). For example, the fact that snakes are tetrapods, even if they secondarily lost their legs, will probably not puzzle any zoologist, because ‘‘tetrapod’’ is an evolutionary concept. The word ‘‘tetrapod’’ is not used in a descriptive, fixist or essentialist meaning, but refers to the character states of the last common ancestor of this vertebrate group. Thus, observing that snakes have no limbs even if they are included within tetrapods

(because they do have other tetrapod features) gives the information that they secondarily lost their limbs. In the same manner, a phylogenetic classification of proteins may help to organise knowledge and to propose evolutionary hypotheses in the field.

Conclusion

We have reviewed many examples showing that protein names are not neutral and can lead to experimental biases. Common names are problematic because protein properties used in the past for creating concepts are heterogeneous, and therefore vary among species. Only a single conceptual framework for names can provide a self-consistent classification that the biochemists, geneticists and molecular biologists need. Further statistical studies could be useful to evaluate the global importance of this phenomenon, but it should be clear that protein names should no longer be based on heterogeneous concepts that narrow the experimental research field. As in systematics, concepts of monophyly should drive the attribution of names, because proteins are evolving entities.

To facilitate such applications we propose a number of suggestions, summarized in Box 2. Clearly, they are only starting points, not definitive methods, and a broad reflection and effort on this nomenclature problem are required to resolve it fully.

Note added in proof:

The ''streetlight syndrome'' should in fact be named the ''moonlight syndrome'', since it is already mentioned in a Middle East story about the mythic 13th century hero Nasreddin Hodja.

One night Nasreddin Hodja lost his ring down in the basement of his house, where it was very dark. Then he went out on the street and started looking for it there, under a splendid moonlight.

A friend passing by stopped and enquired:

- What are you looking for, Hodja? Have you lost something?

- Yes, I've lost my ring down in the basement.

- But Hodja, why don't you look for it down in the basement where you have lost it? asked the friend in surprise.

- Don't be silly, man! I prefer to search where there is some light!

Box 2. Preliminary suggestions to limit the nomenclature problems in protein taxonomy.

1. Official protein names should refer only to their phylogenetic relationships; only proteins from orthologous genes should have the same name; when it is not possible to make a detailed whole-family phylogenetic study, automated tools such as the curated database of phylogenetic trees of animal gene families, TreeFam[53] or the whole phylogeny pipeline http://www.Phylogene.fr should be used.
2. As for Gene Ontology, information about the reliability level of the name (manual curation, phylogeny pipeline, filtered BLAST search) should be mentioned.
3. Organism-based nomenclature should be abandoned.
4. In public databases, the naming field should be updated by database curators.
5. The naming effort should be supported by consequent funding.

Acknowledgments

We thank François Bonneton, Marc Robinson-Rechavi, the editor and three anonymous reviewers for their manuscript reading and useful critical comments, Frédéric Brunet and Michael Schubert for their fruitful discussions.

References

1. Rouze P, Pavy N, Rombauts S. 1999. Genome annotation: which tools do we have for it? Curr. Opin. Plant Biol 2:90–95.
2. Danchin EG, Levasseur A, Rascol VL, Gouret P, Pontarotti P. 2007. The use of evolutionary biology concepts for genome annotation. J Exp Zoolog B Mol Dev Evol 308:26–36.
3. Mathe C, Sagot MF, Schiex T, Rouze P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30:4103–4117.
4. Valencia A. 2005. Automatic annotation of protein function. Curr Opin Struct Biol 15:267–274.
5. Mahner M, Bunge M. 1997. Foundations of Biophilosophy. Berlin, Heidelberg, New York: Springer Verlag. p. 218.
6. Kumar S, Rzhetsky A. 1996. Evolutionary relationships of eukaryotic kingdoms. J Mol Evol 42:183–193.
7. Lee KM, Kaneko T, Aida K. 2006. Prolactin and prolactin receptor expressions in a marine teleost, pufferfish Takifugu rubripes. Gen Comp Endocrinol 146:318–328.
8. Siegfried E, Perrimon N. 1994. Drosophila wingless: a paradigm for the function and mechanism of Wnt signaling. Bioessays 16:395–404.
9. Ogasawara M. 2000. Overlapping expression of amphioxus homologs of the thyroid transcription factor-1 gene and thyroid peroxidase gene in the endostyle: insight into evolution of the thyroid gland. Dev Genes Evol 210:231–242.
10. Kluge B, Renault N, Rohr KB. 2005. Anatomical and molecular reinvestigation of lamprey endostyle development provides new insight into thyroid gland evolution. Dev Genes Evol 215:32–40.
11. Perrimon N, Engstrom L, Mahowald AP. 1985. Developmental genetics of the 2C-D region of the *Drosophila* X chromosome. Genetics 111:23–41.
12. Mangelsdorf DJ, Ong ES, Dyck JA, Evans RM. 1990. Nuclear receptor that identifies a novel retinoic acid response pathway. Nature 345:224–229.
13. Bonneton F, Zelus D, Iwema T, Robinson-Rechavi M, Laudet V. 2003. Rapid divergence of the ecdysone receptor in Diptera and Lepidoptera suggests coevolution between ECR and USP-RXR. Mol Biol Evol 20:541–553.
14. Scott MP. 1993. A rational nomenclature for vertebrate homeobox (HOX) genes. Nucleic Acids Res 21:1687–1688.
15. Baker ME. 2001. Evolution of 17b-hydroxysteroid dehydrogenases and their role in androgen. estrogen and retinoid action. Molecular and Cellular Endocrinology 171:211–215.

16. Peltoketo H, Luu-The V, Simard J, Adamski J. 1999. 17beta-hydroxysteroid dehydrogenase (HSD)/17-ketosteroid reductase (KSR) family; nomenclature and main characteristics of the 17HSD/KSR enzymes. J Mol Endocrinol 23:1–11.
17. Wu L, Einstein M, Geissler WM, Chan HK, Elliston KO, et al. 1993. Expression cloning and characterization of human 17 beta-hydroxysteroid dehydrogenase type 2, a microsomal enzyme possessing 20 alpha-hydroxysteroid dehydrogenase activity. J Biol Chem 268:12964–12969.
18. Iwema T, Billas IML, Beck Y, Bonneton F, Nierengarten H, et al. 2007. Ligand-Independent Functional Conformation of RXR-USP: Insight into Nuclear Receptor-Ligand Evolution. EMBO J 26:3770–3782.
19. Billas IM, Moulinier L, Rochel N, Moras D. 2001. Crystal structure of the ligand-binding domain of the ultraspiracle protein USP, the ortholog of retinoid X receptors in insects. J Biol Chem 276:7465–7474.
20. Nebert DW, Adesnik M, Coon MJ, Estabrook RW, Gonzalez FJ, et al. 1987. The P450 gene superfamily: recommended nomenclature. DNA 6: 1–11.
21. Ertel E, Campbell K, Harpold M, Hofmann F, Mori Y, et al. 2000. Nomenclature of voltage-gated calcium channels. Neuron 25:533–535.
22. Nuclear Receptors Nomenclature Comittee. 1999. A unified nomenclature system for the nuclear receptor superfamily. Cell 97:161–163.
23. Bairoch A. 2000. The ENZYME database in 2000. Nucleic Acids Res 28: 304–305.
24. Spedding M, Foord SM, Hofmann F. 2004. Current status of drug receptor nomenclature: receptor closure? The role of NC-IUPHAR. Expert Opin Investig Drugs 13:461–464.
25. Human Genome Nomenclature Comitee. 2002. Guidelines for Human Gene Nomenclature. Genomics 79:464–470.
26. Nelson DR. 2005. Gene nomenclature by default, or BLASTing to Babel. Hum Genomics 2:196–201.
27. Cokol M, Iossifov I, Rodriguez-Esteban R, Rzhetsky A. 2007. How many scientific papers should be retracted? EMBO Rep 8:422–423.
28. Giguère V, Yang N, Segui P, S Evans RM. 1988. Identification of a new class of steroid hormone receptors. Nature 331:91–94.
29. Yang C, Chen S. 1999. Two organochlorine pesticides, toxaphene and chlordane, are antagonists for estrogen-related receptor -1 orphan receptor. Cancer Res 59:4519–4524.
30. Tremblay GB, Bergeron D, Giguere V. 2001. 4-Hydroxytamoxifen is an isoform-specific inhibitor of orphan estrogen-receptor-related (ERR) nuclear receptors beta and gamma. Endocrinology 142:4572–4575.
31. Horard B, Vanacker JM. 2003. Estrogen receptor-related receptors: orphan receptors desperately seeking a ligand. J Mol Endocrinol 31: 349–357.
32. Thornton JW, Need E, Crews D. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301:1714–1717.
33. Bertrand S, Brunet FG, Escriva E, Parmentier G, Laudet V, et al. 2004. Evolutionary Genomics of Nuclear Receptors: From Twenty-Five Ancestral Genes to Derived Endocrine Systems. Mol Biol Evol 21: 1923–1937.
34. Vanacker JM, Pettersson K, Gustafsson JA, Laudet V. 1999. Transcriptional targets shared by estrogen receptor-related receptors (ERRs) and estrogen receptor (ER) alpha, but not by ERbeta. EMBO J 18:4270–4279.
35. Shigeta H, Zuo W, Yang N, DiAugustine R, Teng CT. 1997. The mouse estrogen receptor-related orphan receptor alpha 1: molecular cloning and estrogen responsiveness. J Mol Endocrinol 19:299–309.
36. Tarrant AM, Greytak SR, Callard GV, Hahn ME. 2006. Estrogen receptor-related receptors in the killifish Fundulus heteroclitus: diversity, expression, and estrogen responsiveness. J Mol Endocrinol 37:105–120.
37. de Urquiza AM, Liu S, Sjoberg M, Zetterstrom RH, Griffiths W, et al. 2000. Docosahexaenoic acid, a ligand for the retinoid X receptor in mouse brain. Science 290:2140–2144.
38. Lengqvist J, Mata DeUrquizaA, Bergman AC, Willson TM, Sjovall J, et al. 2004. Polyunsaturated fatty acids including docosahexaenoic and arachidonic acid bind to the retinoid X receptor alpha ligand-binding domain. Mol Cell Proteomics 3:692–703.
39. Mansfield SG, Cammer S, Alexander SC, Muehleisen DP, Gray RS, et al. 1998. Molecular cloning and characterization of an invertebrate cellular retinoic acid binding protein. Proc Natl Acad Sci USA 95:6825–6830.
40. Gu PL, Gunawardene YI, Chow BC, He JG, Chan SM. 2002. Characterization of a novel cellular retinoic acid/retinol binding protein from shrimp: expression of the recombinant protein for immunohistochemical detection and binding assay. Gene 288:77–84.
41. Folli C, Ramazzina I, Percudani R, Berni R. 2005. Ligand-binding specificity of an invertebrate (Manduca sexta) putative cellular retinoic acid binding protein. Biochim Biophys Acta 1747:229–237.
42. Jackman WR, Mougey JM, Panopoulou GD, Kimmel CB. 2004. *crabp* and *maf* highlight the novelty of the amphioxus club-shaped gland. Acta Zoologica (Stockholm) 85:91–99.
43. Courcelle J, Ganesan AK, Hanawalt PC. 2001. Therefore, what are recombination proteins there for? Bioessays 23:463–470.
44. Lan N, Montelione G, Gerstein M. 2003. Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. Curr Opin Chem Biol 7:44–54.
45. Jaillon O, Aury J, Brunet F, Petit J, Stange-Thomann N, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.
46. Wilkins AS. 2001. Gene names: the approaching end of a century-long dilemma. Bioessays 23:377–378.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.
48. Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19:99–113.
49. Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410.
50. Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1–7.
51. Eyre-Walker A. 1998. Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686–690.
52. Levasseur A, Gouret P, Lesage-Meessen L, Asther M, Asther M, et al. 2006. Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. BMC Evol Biol 6:92.
53. Li H, Coghlan A, Ruan J, Coin L, Heriche J, et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:572–580.

Dev Genes Evol (2008) 218:667–680
DOI 10.1007/s00427-008-0255-7

ORIGINAL ARTICLE

The amphioxus genome enlightens the evolution of the thyroid hormone signaling pathway

Mathilde Paris · Frédéric Brunet · Gabriel V. Markov · Michael Schubert · Vincent Laudet

Received: 11 July 2008 / Accepted: 18 September 2008 / Published online: 7 November 2008

Abstract Thyroid hormones (THs) have pleiotropic effects on vertebrate development, with amphibian metamorphosis as the most spectacular example. However, developmental functions of THs in non-vertebrate chordates are largely hypothetical and even TH endogenous production has been poorly investigated. In order to get better insight into the evolution of the thyroid hormone signaling pathway in chordates, we have taken advantage of the recent release of the amphioxus genome. We found amphioxus homologous sequences to most of the genes encoding proteins involved in thyroid hormone signaling in vertebrates, except the fast-evolving thyroglobulin: sodium iodide symporter, thyroid peroxidase, deiodinases, thyroid hormone receptor, TBG, and CTHBP. As only some genes encoding proteins involved in TH synthesis regulation were retrieved (TRH, TSH receptor, and CRH receptor but not their corresponding receptors and ligands), there may be another mode of upstream regulation of TH synthesis in amphioxus. In accord with the notion that two whole genome duplications took place at the base of the vertebrate tree, one amphioxus gene often corresponded to several vertebrate homologs. However, some amphioxus specific duplications occurred, suggesting that several steps of the TH pathway were independently elaborated in the cephalochordate and vertebrate lineages. The present results therefore indicate that amphioxus is capable of producing THs. As several genes of the TH signaling pathway were also found in the sea urchin genome, we propose that the thyroid hormone signaling pathway is of ancestral origin in chordates, if not in deuterostomes, with specific elaborations in each lineage, including amphioxus.

Keywords *Branchiostoma floridae* · Cephalochordate · Chordate · Development · Evolution · Thyroid hormone · Endostyle

Communicated by J.J. Gibson-Brown

Electronic supplementary material The online version of this article (doi:10.1007/s00427-008-0255-7) contains supplementary material, which is available to authorized users.

M. Paris · F. Brunet · G. V. Markov · M. Schubert · V. Laudet (✉)
Institut de Génomique Fonctionnelle de Lyon, CNRS UMR5242–INRA 1288–ENS–UCBL, IFR128 BioSciences Lyon-Gerland, Ecole Normale Supérieure de Lyon,
46 allée d'Italie,
69364 Lyon Cedex 07, France
e-mail: Vincent.Laudet@ens-lyon.fr

Introduction

In isolated human populations, usually located far from the sea, goiter and mental retardation are efficiently prevented by using iodized table salt. Iodine is mainly used in mammals to synthesize thyroid hormones (THs), so the mental impairments observed among "cretins" indicate the important role that THs play as regulators of development (Flamant and Samarut 2003). Outside mammals, THs can have more drastic effects on development and, for instance, THs are also responsible for the regulation of metamorphosis in amphibians (Shi 2000). In this case again, the main source of TH is endogenous, since inhibition of TH synthesis by chemical means (using goitrogens such as PTU) blocks metamorphosis. The metabolism of TH has been extensively studied in mammals and amphibians, and was found to be very similar. In both cases, the most active form is T_3 (3,3′,5-triiodo-L-thyronine), which can be produced from its precursor T_4 (L-thyroxine). T_4 is an iodinated tyrosine derivative: it is formed by two tyrosines, coupled each with two iodines whereas in T_3 the external

tyrosine lacks an iodine (Fig. 1a). In vertebrates, TH synthesis takes place in the thyroid gland, a gut derivative, which is organized in follicles constituted of single layered epithelial cells called thyrocytes, enclosing a lumen filled with the matrix protein thyroglobulin. The main steps of TH signaling pathway are indicated in Fig. 1b. The first event in TH synthesis involves increasing the concentration of iodine in thyrocytes, through the sodium/iodine symporter (SIS) in their basal membrane. This is an important step, since iodine is present at very low concentrations in food, as well as in blood. In the lumen of the follicles, iodine gets oxidized and transferred to a few specific tyrosine residues on thyroglobulin to form monoiodotyrosine—MIT—if only one iodine is transferred to a tyrosine, or diiodotyrosine—DIT, if two iodines are transferred to a tyrosine. Iodinated tyrosines are further coupled, under the catalysis of the thyroid hormone peroxidase (PERT). The iodinated thyroglobulin is then incorporated back into the thyrocytes and hydrolyzed in lysosomes, allowing the release of T_4 (if two DITs have been coupled), and, to a

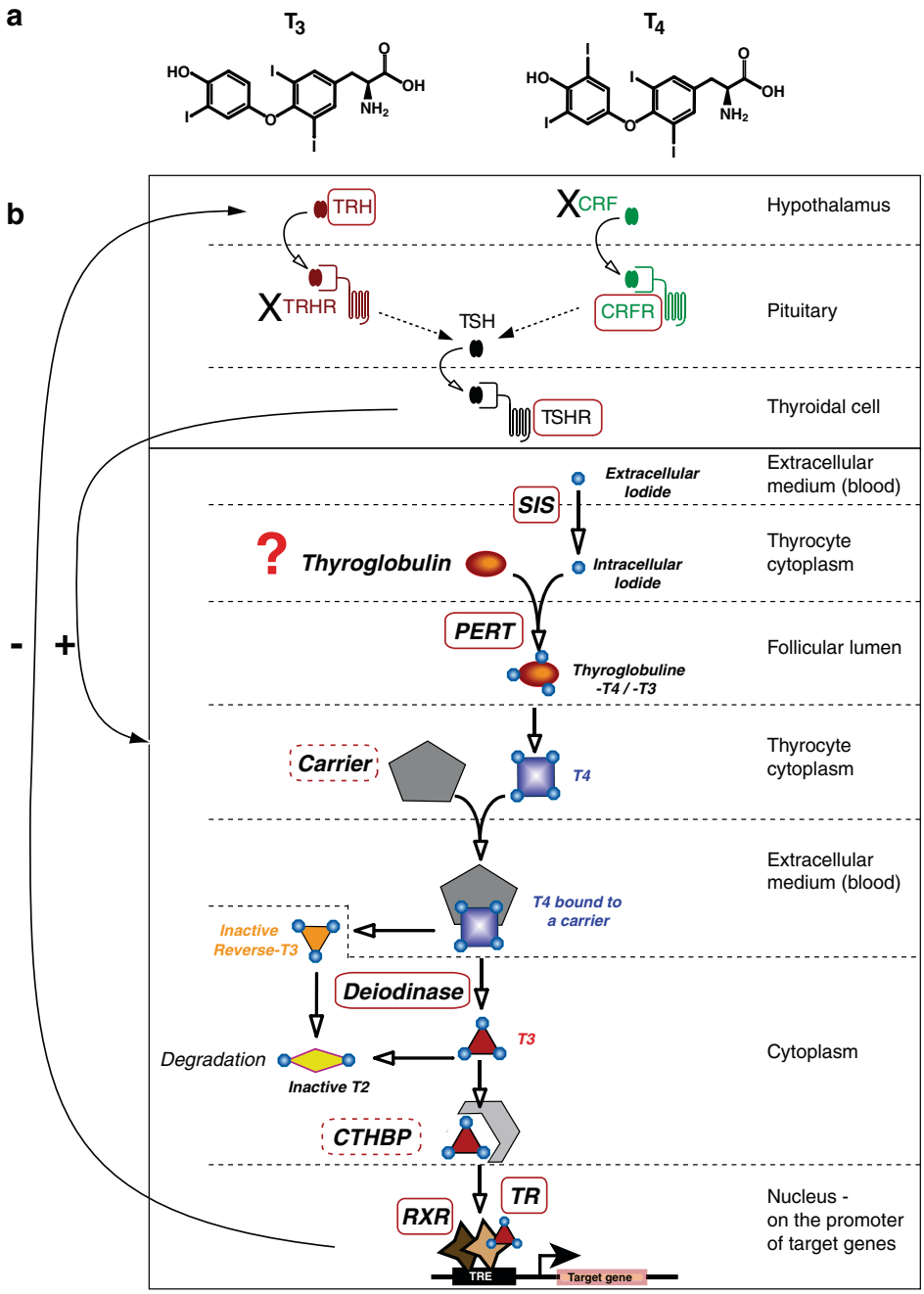


Fig. 1 TH signaling pathway, as it is known in vertebrates. **a** Molecular structure of T_4 and T_3. **b** The different steps of the TH signaling pathway, as it is known in vertebrates, are indicated. The steps of TH production and action have been separated from the hypothalamo-pituitary regulation of TH production (see the *boxes*). The neuroendocrine regulation of TH production is indicated with a "+". The negative feedback loop by TR on this upper regulation is also indicated. The genes encoding members of the TH signaling pathway and studied here are highlighted as follows: (1) the genes for which orthologous sequences were found in the amphioxus genome are *boxed in red* and (2) the genes for which no orthologous sequence was found are indicated with a *cross*. The case of thyroglobulin remains unclear. Given the uncertainty of the carrier (TBG being the only one found) and CTHBP proteins (see the text), the amphioxus sequences are indicated with *dashed lines*

lesser extent, T_3 (if a MIT and a DIT tyrosines have been coupled) (for a review, see Hulbert 2000).

Although THs are produced only in the vertebrate thyroid gland, they are found everywhere in the organism, transported through blood circulation. The hydrophobic benzene rings of TH increase the tendency of THs to partition in plasma membranes (Schreiber and Richardson 1997), so that only a small fraction of T_4 and T_3 is "freely" transported and most of it travels bound to a "carrier" protein, like thyroxine-binding globulin (TBG), transthyretin (TTR), or albumin in mammals. In peripheral tissues, T_4 is deiodinated into the active form T_3 by specific proteins called deiodinases. These enzymes remove iodines from the inner or outer ring of the tyrosine skeleton. There are three different deiodinases in tetrapods (D1, D2, and D3), which have different affinities for the various THs and are responsible for fine-tuning thyroid hormone action jointly with a few alternative TH pathways (like sulfatation and glucoronidation) that produce mostly inactive T_3 derivatives, targeted for fast degradation (Wu et al. 2005). In each target cell, T_3 binds to its receptor, the thyroid hormone receptor (TR), which belongs to the superfamily of nuclear hormone receptors (NRs) (Flamant et al. 2006). As many members of the NR superfamily, TR is a ligand-dependent transcription factor: in the absence of T_3, TR is most often bound to DNA on specific sequences in the promoter of target genes whose expression it inhibits, through the recruitment of transcriptional corepressors. Upon T_3 binding, TR recruits coactivator proteins and activates the expression of target genes that will lead to the biological effects induced by THs (Fig. 1b).

In mammals, whereas deiodinases constitute a peripheral system for controlling TH production, the hypothalamic pituitary thyroid axis allows a central control of TH production by the thyroid. In this axis, the hypothalamus produces TSH-releasing hormone (TRH) that stimulates the secretion by the pituitary of thyroid-stimulating hormone (TSH), which in turn stimulates TH production by TH-producing cells in the thyroid (Fig. 1b). An important aspect of this regulation is the negative feedback loop, through which THs are negatively regulating their own production through the inhibition of TRH and TSH production (Fig. 1b). This TH-dependent axis is a paradigm of integrated exquisite endocrine regulation of hormone production at the level of the organism, because it can integrate external signals (e.g. food, population density), therefore allowing a link between the environment and the endocrine production of THs (Yen 2001). In amphibians, up-regulation of TSH secretion is taken care of by the corticotropin-releasing hormone (CRH), a primary regulator of stress response that can stimulate TSH production. Thus, the production of the CRH peptide by the hypothalamus is dependent on environmental stimuli and integrates stress information that can be transmitted down to the thyroid gland (Denver 1997) (Fig. 1b).

The evolution of this rather complex signaling pathway has been investigated mainly in gnathostomes, where it was found to be well conserved (Hulbert 2000). However, it is still unknown how the capacity to produce THs and their function evolved in the first place. In the basal vertebrate lamprey, THs are produced through a pathway probably homologous to the mammalian one (for instance Manzon et al. 2007) and regulate metamorphosis, as in amphibians, with the difference that it is a drop, and not a peak of TH, that triggers metamorphosis (Manzon et al. 2001). Not much is known about the TH pathways of non-vertebrate chordates (the vertebrate sister group urochordates like the sea squirt *Ciona intestinalis*, and the cephalochordates like amphioxus), except that they possess an organ homologous to the thyroid gland, which is named endostyle, and can produce THs (see Paris and Laudet 2008 for a review). However, biological effects of THs outside vertebrates have been reported several times, and THs have been linked with metamorphosis in urochordates (Patricolo et al. 2001), amphioxus (Paris et al. 2008), and in echinoderms (Heyland and Hodin 2004).

Several lines of evidence, including biochemical studies, demonstrated that there is an active TH metabolism similar to vertebrates in the most basal chordate: amphioxus (reviewed in Eales 1997). First, the endostyle is an amphioxus organ that is widely accepted as being homologous to the follicles of the vertebrate thyroid gland (Ogasawara 2000). Secondly, both T_3 and T_4 have been detected in amphioxus (Covelli et al. 1960). Thirdly, fixation of iodine was reported in the endostyle of amphioxus (Fredriksson et al. 1985) and was shown to be dependent on peroxidase activity. Fourthly, a protein with biochemical properties similar to thyroglobulin has been described in amphioxus (Monaco et al. 1981). Fifthly, deiodinase activity has been indirectly demonstrated by showing that inhibition of T_4 deiodination by chemical means inhibits metamorphosis. Lastly, there is an active TR in amphioxus, that is involved in metamorphosis (Paris et al. 2008). Taken together, these results strongly suggest that there is production of T_3 and T_4 and more generally an active TH signaling pathway in amphioxus. However, only very few members of this signaling pathway have been identified so far in amphioxus. Here we take advantage of the recent release of the amphioxus (*Branchiostoma floridae*) genome (Holland et al. 2008; Putnam et al. 2008) and describe orthologs of the main genes involved in the TH signaling pathway. We propose that in amphioxus THs are produced the same way as in vertebrates, which suggests an ancient origin of the TH signaling pathway within the chordate lineage. Several lines of evidence from

echinoderms suggest an even more ancient origin of TH signaling within deuterostomes.

Materials and methods

Sequence retrieval We used zebrafish or human protein sequences to retrieve Ensembl families (version 45 as of June 2007) of our sequences of interest. Ensembl families are made of genomic sequences as well as Swiss-Prot and TrEMBL data. We subsequently blasted (blastp) the protein sequences of interest (usually a human sequence) against the amphioxus genome (Putnam et al. 2008). Orthologous sequences from other species (e.g. the sea urchin *Strongylocentrotus purpuratus*, the sea anemone *Nematostella vectensis*, and the bee *Apis mellifera*), used as outgroups, were retrieved by blast searches with the amphioxus sequences carried out on the NCBI site (www.ncbi.nlm.nih.gov/). Protein sequences from the lamprey *Petromyzon marinus* were obtained from protein, EST or genomic databases. Protein sequences from the elephant shark *Callorhinchus milii* were obtained from genomic databases (after protein sequence prediction using Genscan). A complete list of retrieved sequences is given Fig. S2. We also did a reverse blast (blastp) onto all Metazoa non-redundant sequences available both at the Swiss-prot group (expasy.org/sprot/) and NCBI (ncbi.nlm.nih.gov/) and constructed phylogenetic trees using selected sites of the sequences. Closely related genes were used as outgroups. Alignments were made using Muscle (Edgar 2004), allelic sequences were removed from the alignments after comparisons of the nucleotide sequences and best conserved sites were selected using GBlock (Castresana 2000) for further phylogenetic analysis.

Phylogenetic analysis For each family, we performed first a neighbor-joining (NJ) analysis, with Poisson law correction and pairwise gap-removal, as implemented in PhyloWin (Galtier et al. 1996) (the relevant data are available upon request to FB), and subsequently we performed a maximum likelihood (ML) phylogenetic analysis using PhyML v2.4.4 (Guindon and Gascuel 2003). For the ML analyses, JTT model with eight rate categories and an estimated gamma shape parameter were used. Robustness was assessed by bootstrap analysis (1,000 repetitions).

Evolutionary rate comparison between vertebrate peroxidases Two urochordate sequences (one—PERT1—from *C. intestinalis* and the other from *Halocynthia roretzi*) were used as outgroups. The sequence Q6NUY7_BRARE, Q640K8_XENLA, and Q9YH34_XENLA were excluded from the analysis because their phylogenetic position does not fall into one of the four peroxidase groups PERT, PERE, PERL, and PERM. The input tree corresponded to the topology described in Fig. 3.

Results

Our strategy was to search for amphioxus sequences from the *B. floridae* genome, that are homologous to vertebrate proteins involved in TH signaling. When possible, sequences from basal vertebrates, urochordates, and sea urchin were also studied for comparative purposes. Our search identified a total of 24 genes in amphioxus (Holland et al. 2008; Putnam et al. 2008; Schubert et al. 2008) plus RXR and TR that were previously identified. For almost all gene families, several orthologs were found, arising from lineage specific duplications. In most cases, the amphioxus genes were not direct orthologs of each vertebrate paralogous gene, but rather branched at the base of a tree of closely related vertebrate genes. The complete list of identified genes, with their corresponding accession numbers assigned by the JGI consortium, is given in Table S1.

SIS

Blast searches allowed us to identify eight predicted genes, for which orthology was tested by a phylogenetic analysis (Fig. 2). We found three groups of closely related vertebrate genes, including SIS, each time with 100% bootstrap (the sodium-iodide symporter SIS, the apical iodide transporter, and the sodium-lactate cotransporter). Two out of the three groups contain genes encoding proteins characterized as sodium-iodine symporters located either on the basolateral membrane (SIS) or on the apical membrane of thyrocytes (apical iodide transporter) in human (Rodriguez et al. 2002) (shown in brackets in Fig. 2). The genes of the third vertebrate group encode proteins identified as sodium-lactate cotransporters (Gopal et al. 2007). Notably, three sea urchin (*S. purpuratus*) genes cluster together at the base of the vertebrate groups (with a bootstrap support of 71%). When going deeper in the tree, eight amphioxus genes cluster together (96% bootstrap support) at the base of this cluster of paralogous genes (with a bootstrap support of 75%). These topologies suggest three independent series of duplication, one at the base of the sea urchin lineage, one in the amphioxus lineage, and the other at the base of the vertebrate lineage, giving rise to the three vertebrate-specific groups. As a sequence from the elephant shark *C. milii* genome was found in all three vertebrate groups (with low bootstrap support, though, Fig. S1), the vertebrate duplications occurred before *C. milii* split, probably during the two rounds of whole genome duplication that occurred at the base of the vertebrate group (Dehal and Boore 2005).

Dev Genes Evol (2008) 218:667–680

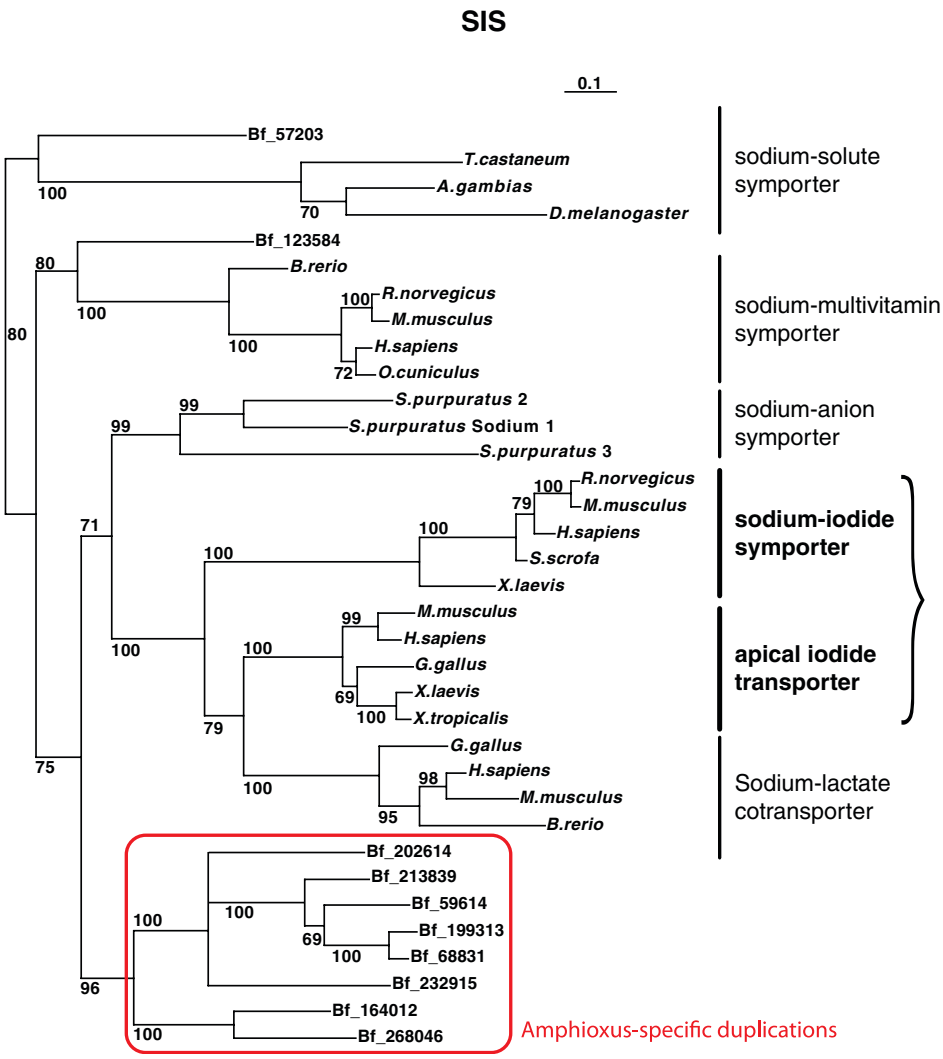


Fig. 2 Phylogenetic tree of SIS and related protein sequences. A maximum likelihood (ML) tree was obtained from analysis of SIS amino acid sequences. Bootstrap percentages obtained after 1,000 replicates are shown. Nodes with bootstrap support below 50% were collapsed. The amphioxus SIS-like sequences have been *boxed in red*. The *scale bar* indicates the number of changes per site. A similar tree including sequences from basal vertebrates is given in Fig. S1

The structures of the amphioxus proteins were then predicted *in silico*, in order to determine if one of them was more likely to be a sodium-iodine symporter. From the TMpred program (http://www.ch.embnet.org/software/TMPRED_form.html), the eight sequences were predicted to harbor 12 to 13 transmembrane domains like the already characterized vertebrate SIS, except the sequence Bf_68831 (Fig. 2), for which 12 supplementary transmembrane domains were predicted in its 200 amino acid longer C-terminal region (data not shown). Moreover, several sites that are known to be important for the general integrity of the function of the transporter and are found to be conserved within sodium symporters, are conserved in the amphioxus sequences (Dohan et al. 2003).

PERT

PERTs are part of a big family of peroxidases that catalyze oxidation of various substrates with hydrogen peroxide (including myeloperoxidase (PERM), eosinophil peroxidase (PERE), thyroid peroxidase (PERT), and lactoperoxidase, (PERL) (Daiyasu and Toh 2000)). Paralogous sequences have been described in urochordates and in the Japanese amphioxus species *Branchiostoma belcheri* (Ogasawara 2000; Ogasawara et al. 1999).

Blast searches allowed us to identify seven predicted genes, for which orthology was tested by phylogenetic analysis (Fig. 3). From our analysis, two groups consisting of three and four amphioxus sequences respectively branch together with high bootstrap support (100%) each. The group of three genes is distantly related to PERTs but still branches within peroxidase-like proteins (and the corresponding proteins may thus display peroxidase activity). The four other amphioxus sequences (red box, Fig. 3), with which the known *B. belcheri* sequence branches (Ogasawara 2000), are more closely related to PERTs and are located at the base of the tree constituted of the vertebrate-specific subgroup and urochordate sequences with a good bootstrap support

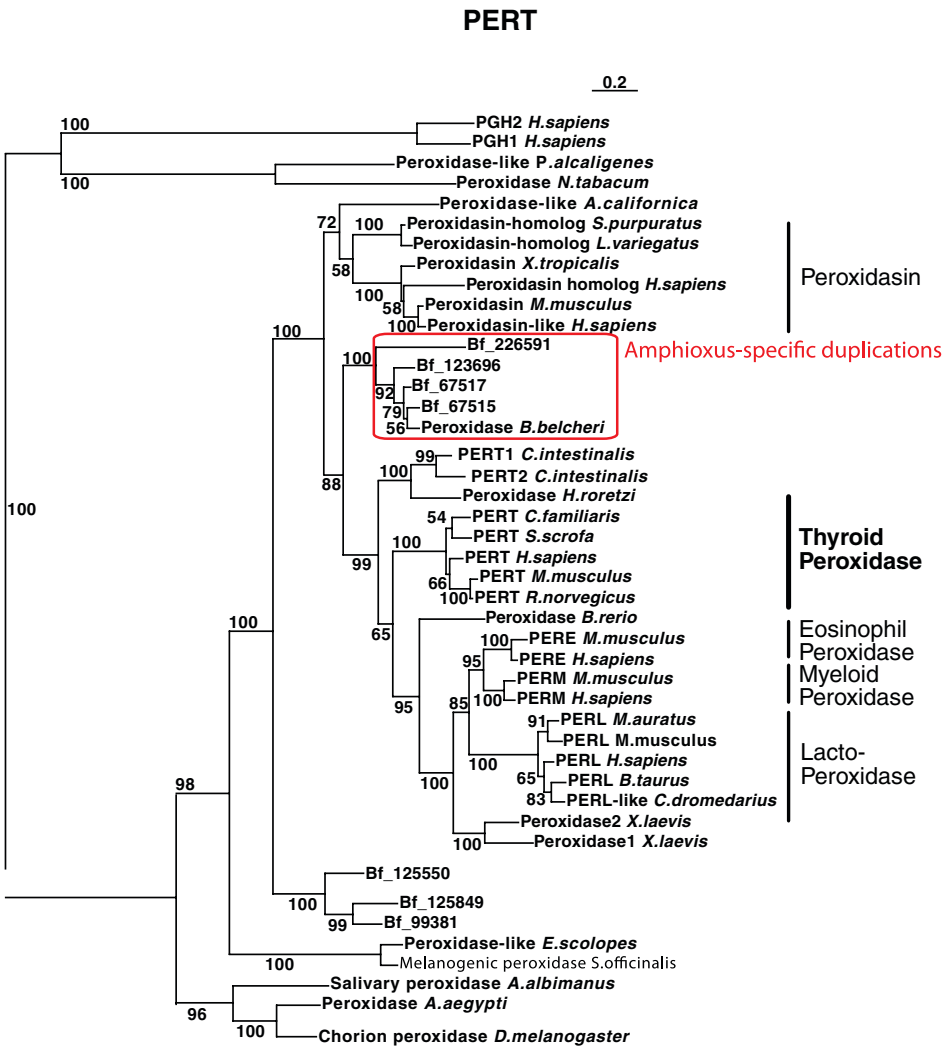


Fig. 3 Phylogenetic tree of PERT and related protein sequences. A maximum likelihood (ML) tree was obtained from analysis of peroxidase amino acid sequences. Bootstrap percentages obtained after 1,000 replicates are shown. The amphioxus PERT-like sequences have been *boxed in red*. PGH: Prostaglandin G/H synthase. The *scale bar* indicates the number of changes per site. A similar tree including sequences from basal vertebrates is given in Fig. S2

(88%). Amphioxus sequences have very short branches, suggesting a low evolutionary rate and a rapid duplication burst. Similarly, PERT genes seem to evolve at a slower rate than the three closely related peroxidases PERM, PERL, and PERE. To test this hypothesis, we compared the evolutionary rates between the different vertebrate peroxidase families, using the relative-rate test on available sequences (Robinson et al. 1998). Results are shown in Table S3, as differences of substitution rate between groups of species. From these analyses, it appears that PERTs have effectively evolved at significantly lower rates than the three other peroxidase groups. This common slow evolutionary rate in amphioxus and vertebrate PERTs may reflect that these proteins have kept an ancestral function (see Discussion).

Thyroglobulin

Thyroglobulin is a large protein (more than 2,700 amino acids in humans) that contains an esterase domain extending to the 500 most C-terminal amino acids, whereas the N-terminal domain houses thyroglobulin type I repeats (TY repeats), in which many PERT-targeted tyrosines have been located. Many hits are retrieved from blasting the human thyroglobulin sequence against the amphioxus genome. However, when they are aligned with esterase domain of thyroglobulins, they branch with esterases and not thyroglobulins in a phylogenetic tree (data not shown) and lack a N-terminal part homologous to the thyroglobulin one. Conversely, TY repeats were recognized in some predicted genes. However, these repeats, although first discovered in thyroglobulins (hence their name), are found in many different proteins (Novinec et al. 2006). Among the amphioxus retrieved sequences, the best hit (Bf_123169) is a very long protein (about 2,400 amino acids). However, nothing else but the TY repeats is homologous to the thyroglobulin (data not shown) and the C-terminal part is not an esterase sequence. Overall, up to now, no unambiguous thyroglobulin sequence was found

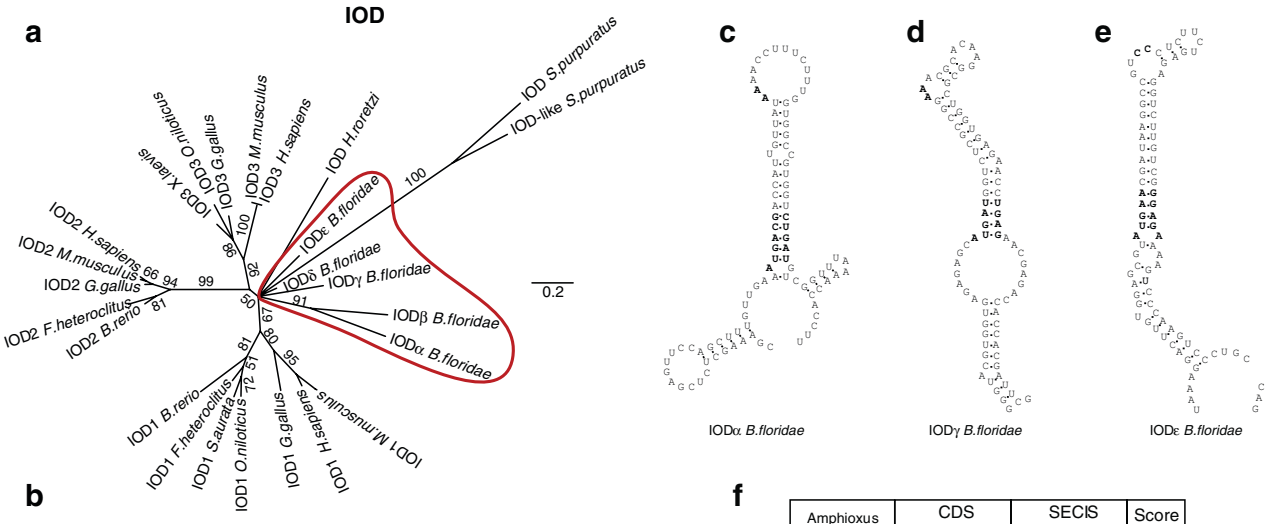
in the amphioxus genome or outside vertebrates, although one rather short sea urchin (*S. purpuratus*) sequence (137 amino acids, accession number 115921343) clusters with vertebrate thyroglobulins.

Transport proteins

Several carriers for THs have been identified in vertebrates, like thyroxine-binding globulin (TBG), transthyretin (TTR), and albumin (with a rather low affinity). Although present in lampreys (Schreiber and Richardson 1997) and sea urchins (*S. purpuratus*) (data not shown), no gene encoding albumin (that is not specific to TH transport) and no TTR gene were found in the amphioxus genome. One gene related to TBG was found but it is not a direct ortholog (Fig. S4). Once in the cell, THs can also bind to cytosolic proteins, like the cytosolic thyroid hormone-binding protein (CTHBP). A gene made of the concatenation of two predicted genes from the genome (Bf_267438 and Bf_110402) branches at the base of a subtree constituted of two vertebrate gene families including CTHBP (Fig. S5). Further studies will be required to assess, which proteins mediate TH transport in amphioxus.

Deiodinases

We retrieved five sequences from the amphioxus genome that cluster with previously described deiodinase sequences (data not shown). However, only one amphioxus sequence (Bf_123596 called IODβ in Fig. 4a) includes a domain known to be important for vertebrate deiodinase function (Bianco et al. 2002) (Fig. 4b). All other predicted sequences were truncated at those positions. Notably, in this activation domain, all vertebrate deiodinases, which are selenopro-



Amphioxus sequence	CDS		SECIS		Score (>15)
	Start	End	Start	End	
IODα *B.floridae*	2437	16226	16544	16639	29.8
IODγ *B.floridae*	676	2481	3058	3153	25.8
IODε *B.floridae*	391	2432	2992	3081	22.07

Fig. 4 Phylogenetic and structural analyses of putative deiodinases in amphioxus. **a** A maximum likelihood (ML) tree was obtained from analysis of deiodinase amino acid sequences. As no outgroup was found for deiodinases, the tree presented is unrooted. Bootstrap percentages obtained after 1,000 replicates are shown. Nodes with bootstrap support below 50% were collapsed. The amphioxus sequences have been *circled in red*. The *scale bar* indicates the number of changes per site. A similar tree including sequences from basal vertebrates is given in Fig. S3a. **b** Amino acid sequence alignment of the active catalytic domain of the deduced amino acid sequences for the human deiodinases as well as the sequences found in invertebrates. The site, where a selenocysteine is located, is shown in *red*. In humans and *H. roretzi*, the translation of the TGA codon into a selenocysteine has been experimentally verified and is shown in *bold* (Berry et al. 1991; Curcio et al. 2001; Baqui et al. 2003). In the amphioxus and sea urchin sequences, the homologous TGA is proposed to be translated into a selenocysteine as well (*italicized*). **c**–**e** SECIS elements in the putative 3′UTR of amphioxus IODα (**c**), IODγ (**d**), and IODε (**e**), as predicted by SECISearch analysis (Kryukov et al. 2003). The characteristic adenosine that precedes the quartet of non-Watson-Crick base pairs, a TGA_GA motif in the quartet and two adenosines or cytosines in the apical loop, are highlighted in *bold* (Kryukov et al. 2003). **f** The position of the CDS in shotgun sequences used for retrieving amphioxus IODs as predicted by Genscan are indicated as well as the positions of the SECIS elements and their scores as given by SECISearch

teins (Bianco et al. 2002), have a selenocysteine, encoded by a TGA codon, which is key to the activity of the protein (Bianco et al. 2002) (hence the name selenoprotein). In the amphioxus sequence Bf_123596, a cysteine encoded by a TGC codon is located at the homologous position to the selenocysteine of vertebrates. This is unlikely to be due to sequencing error as the same codon was found in a sequence obtained from an independent cDNA library (accession number BW699364). This difference casts doubts on the functionality of the Bf_123596 protein as an amphioxus deiodinase.

As the TGA codon encoding the selenoprotein is predicted as a stop codon and not a selenocysteine by gene prediction software, genes encoding selenoproteins would be inappropriately predicted as pseudogenes (with a precocious stop codon) or the protein would be predicted as truncated of the part around the amino acid (with the TGA codon predicted as intronic). In order to test this hypothesis, we performed a tblastn search on the shotgun sequence of the amphioxus genome, using the amino acid sequence of the rat IOD2 as a query to avoid the gene prediction problem. Using this method, we retrieved several genes, some of which included a TGA codon at the position corresponding to the selenocysteine in the query sequence. We were able to attribute all of the new sequences to previously found truncated amphioxus sequences (Fig. 4a, the sequences were renamed IODα to IODε). All newly identified sequences have an activation domain that is very well conserved with the activation domains of vertebrate sequences (Fig. 4b). In particular, all but the sequence Bf_123596 (corresponding to IODβ in Fig. 4) have a TGA codon at the "selenocysteine position". To confirm our hypothesis that at least some of the amphioxus sequences are selenoproteins, we searched for selenocysteine insertion sites (SECIS) known to be required for proper translation of the TGA codon into a selenocysteine (Bianco et al. 2002). SECIS were predicted in the putative 3′UTRs of IODα, IODγ, and IODε using the SECISearch program (Kryukov et al. 2003) (Fig. 4c–f). A histidine residue also important for deiodinase activity (Bianco et al. 2002) is conserved in all sequences. We hence predict that IODα, IODγ, and IODε are selenoproteins, and, taken together, these data strongly suggest that there are several active deiodinases in amphioxus.

A similar analysis was performed with the sea urchin genome (*S. purpuratus*) and one sequence was found containing both the activation domain with a putative senelocysteine and another partial sequence, for which the sequence corresponding to the activation domain was not available (Fig. 4a,b). The corresponding 3′UTR sequence was not available. A single gene has previously been described in the urochordate species *H. roretzi* as being a functional deiodinase (Shepherdley et al. 2004).

The monophyly of the three vertebrate deiodinases was recovered in our phylogenetic analysis (Fig. 4a). However, none of the non-vertebrate sequences could be precisely located in the phylogenetic tree with high bootstrap support. They branched at the base of the three vertebrate deiodinase families (IOD1, IOD2, and IOD3), suggesting that at the origin of deuterostomes, there was only one deiodinase that was duplicated independently in the sea urchin, amphioxus, and vertebrate lineages (Fig. 4a).

Genes involved in TH synthesis regulation

We searched the amphioxus genome for genes orthologous to CRF, TRH, TSH, and their respective receptors. Homologs to vertebrate CRFR, TSHR were found (Figs. S7 and S8). An orthologous sequence to the preproTRH was also found (Fig. S9). Interestingly, it contains only one canonical progenitor sequence (Glu-His-Pro-Gly) (in comparison to about five in mammals and birds, Vandenborne et al. 2005 and references therein), but also 21 sequences with the His replaced with a Ser (Glu-Ser-Pro-Gly). The flanking sequence (Lys-Arg) is conserved. No orthologs to TRH receptor, CRF, and TSH were found in the amphioxus genome. As a gene orthologous to CRH was found in insects and since the corresponding protein is rather small, it is likely that CRF is present in amphioxus, but not in the genome sequence.

Protein sequences involved in TH signaling in basal vertebrates

In order to gain insights into the evolution of the genes encoding proteins implicated in TH signaling and as these genes are mostly known from classical vertebrate models (human, mouse, zebrafish, and *Xenopus*), we searched for them in databases of basal vertebrates (e.g. lamprey and cartilaginous fishes). The sequences of two relevant genomes are publicly available and sufficiently annotated to allow a systematic search: the lamprey *P. marinus* (http://genome.ucsc.edu/cgi-bin/hgGateway?clade=other&org=Lamprey&db=) and the elephant shark *C. milii* (http://esharkgenome.imcb.a-star.edu.sg/). The dataset was completed with EST databases. The gene families we found in amphioxus are also present in cartilaginous fishes: SIS (Fig. S1), PERT (Fig. S2), IOD (Fig. S3), TBG (Fig. S4b), CTHBP (Fig. S5b), CRFR (Fig. S7b), TSHR (Fig. S8b), and TRH (Fig. S9). With the exception of IOD and TRH (which is a short protein), we also found representatives of each gene family in *P. marinus*. In addition, TR and RXR have previously been identified in both *P. marinus* and in the shark *Scyliorhinus canicula* (dogfish) (Escriva et al. 2002; Paris et al. 2008). As in amphioxus, we did not find TG in these genomes.

The predicted protein sequences could not be placed reliably in the context of phylogenetic trees. Given the low coverage of the two genomes (1.4 and 5.9, respectively, for *C. milii* and *P. marinus*) and the small size of DNA fragments in the EST and genome databases (less than 3 kb on average), only partial sequences could be retrieved and probably did not provide enough signal to be well positioned in our phylogenetic trees. Nevertheless, these data confirm the presence of the TH signaling pathway in basal vertebrates.

Discussion

An ancestral origin of TH signaling pathway in chordates

In this analysis, we have shown that the genome of the amphioxus *B. floridae* contains genes coding for proteins homologous to most of the genetic equipment necessary to endogenously produce thyroid hormones. SIS, PERT, deiodinases, cytosolic binding proteins and specific nuclear receptors are present in the genome (Fig. 1b). In addition, biochemical reactions necessary for TH synthesis are apparently similar in amphioxus and vertebrates (see the Introduction for more details). Many genes were also found in cartilaginous fishes, lampreys, and urochordates. Consequently, we propose that the TH signaling cascades of extant chordates are homologous, *i.e.* they evolved from an ancestral cascade that was present in the common ancestor of all chordates. Nevertheless, several key components of the pathway are missing (e.g. thyroglobulin) or are different (e.g. independent duplications) in amphioxus when compared to vertebrates.

In all chordates studied so far, THs were shown to regulate some features of post-embryonic development, the most spectacular example being metamorphosis (in amphibians (Tata 2006), in mammals (Flamant and Samarut 2003), in teleost fishes (Power et al. 2001), in lamprey (Manzon et al. 2001), in urochordates (Patricolo et al. 2001), and also in amphioxus (Paris et al. 2008)). The ancestry of the TH signaling pathway suggests that already in the chordate ancestor, metamorphosis was regulated by THs, which were synthesized endogenously. Interestingly, lamprey metamorphosis is inhibited by THs suggesting a specific elaboration of the TH signaling pathway in lampreys when compared to other chordates (Youson 1997).

Evolution of members of the TH signaling pathway

In most cases, several amphioxus sequences branch at the base of a group of several vertebrate sequences. The topology, where an amphioxus group corresponds to several vertebrate paralogs points to gene duplications at the base of the vertebrate lineage, probably due to the two rounds of whole genome duplication (WGD) events that occurred at the base of the vertebrate clade (Dehal and Boore 2005; Putnam et al. 2008). This pattern was observed repeatedly throughout the analysis of the amphioxus genome (about 25% of chordate gene families follow this scenario (Holland et al. 2008; Putnam et al. 2008), the other gene families having most probably undergone loss (Lynch and Conery 2000)). Accordingly, several paralogous genes were found for basal vertebrates (Figs. S1–S9). However, because of the limited genomic resources at this taxonomic position, only partial sequences were retrieved. Therefore, they could not be reliably placed in our phylogenetic trees.

Gene and genome duplications have been proposed to be substantial sources of new genes (Ohno 1970). After duplication, two identical genes would encode proteins able to perform the same function. This functional redundancy would free one of the two copies that could increase its mutation rate (Force et al. 1999). Several scenarios have been proposed: proteins encoded by both copies may share parts of the original function (subfunctionalization), one of the proteins may fulfill the ancestral function while the other one either degenerates (nonfunctionalization) or gains new functions (neofunctionalization). The study of the evolutionary rate of the peroxidase proteins is in agreement with the last point. Indeed, within the thyroid/myeloid/eosinophil/lymphoid peroxidase (PERT/PERM/PERE/PERL) family, the thyroid peroxidases PERTs, may carry the most ancestral function of the peroxidase quartet. PERTs are evolving significantly more slowly than the three other groups (see Table S3, this low evolutionary rate is illustrated by their short branches in Fig. 3) and the closely related invertebrate chordate sequences (amphioxus and urochordate) also have short branches indicating low mutation rates. Accordingly, thyroid peroxidase activity was reported in lampreys, urochordates, and amphioxus (see Paris and Laudet 2008 for a review). Interestingly, a duplication occurred specifically in the *C. intestinalis* lineage (Fig. 3), and the characterized gene (PERT1) displays only partial functional redundancy with peroxidase activity in the *C. intestinalis* endostyle (Ogasawara et al. 1999), possibly suggesting a partition of PERT function between the two proteins because of a subfunctionalization event (Markov et al. 2008). Further characterization of PERT2 will give further insights into the evolution of thyroid peroxidase activity in *C. intestinalis*. We propose that in the chordate ancestor, there was one peroxidase gene, which was a thyroid peroxidase (or more appropriately called an endostyle peroxidase). During subsequent lineage-specific evolution, this gene duplicated several times independently in the

vertebrate, *C. intestinalis* and in the amphioxus lineage. Among the four amphioxus retrieved sequences, Bf_67515 is the best candidate for encoding an active PERT. Indeed, PERTs are the only transmembrane peroxidases and Bf_67515 is the only amphioxus sequence harboring a putative transmembrane domain (as predicted by TMpred) located at the same position as in vertebrate PERTs (sites 752–776 corresponding to the sites 847 to 871 in the human PERT). Moreover, peroxidase activity has been located in the endostyle of amphioxus larvae in the outer surface of plasma membranes as well as to the inner surface of membranes in cytoplasmic compartments (Fredriksson et al. 1985). Additionally, the orthologous sequence from *B. belcheri* (Fig. 3) is expressed in the thyroid gland homolog, the endostyle during development (Ogasawara 2000). However, it does not contain the transmembrane domain. Whether this corresponds to an alternative splicing or a genomic difference needs further investigation.

Several genes were not retrieved in our analysis whereas biochemical studies suggested a different result. The fact that no gene related to thyroglobulin was found in amphioxus, whereas all the other key members of the signaling pathway are present, is not surprising because thyroglobulin is a long and divergent protein in which only some tyrosines are implied in TH production. A biochemical study revealed the existence of a protein harboring classical thyroglobulin properties in the amphioxus endostyle: a large protein that incorporates iodines through peroxidase activity to produce T_3 and T_4 (Monaco et al. 1981). It is possible that a protein non-homologous with vertebrate thyroglobulins is the source of tyrosine and the amphioxus sequence Bf_123169 is an interesting candidate. The purification of the protein detected by Monaco et al. (1981) and the further cloning of the corresponding gene as well as the cloning of similar proteins in other lineages (lampreys, urochordates, sea urchins) may help to resolve the issue of iodinated tyrosine sources in invertebrates. Similarly, no direct ortholog of TH transporters was found in the amphioxus genome. Only sequences distantly related to TBG were retrieved (no ortholog of TTR or albumin was found). Nonetheless, THs have been shown to be produced in amphioxus (Covelli et al. 1960) and considering the high hydrophobicity of THs, it is most probable that different carriers are involved in TH transport in amphioxus, but their exact nature remains unclear. This illustrates limitations to the approach we chose for studying TH signaling: phylogenetic studies only allow the detection of possible candidates that carry out a biological function known from vertebrates.

The case of deiodinases is interesting with respect to genome annotation. We found several amphioxus sequences that are most probably bona fide deiodinases. However, based on the genome annotation, the predicted sequences lack the activation domain because of a TGA codon very likely wrongly annotated as a stop codon instead of a selenocysteine codon that is always present in vertebrate deiodinase genes (Bianco et al. 2002). By manually annotating the amphioxus genome, we discovered several sequences that are likely to be active deiodinases in amphioxus. We propose that the ancestral deiodinase displayed a T_4-to-T_3 outer-ring deiodination activity because (1) T_4-to-T_3 production probably occurs in amphioxus (Covelli et al. 1960) and (2) the *H. roretzi* deiodinase was shown to have such an activity (Shepherdley et al. 2004). Of course, this hypothesis will require functional evidence with the biochemical characterization of the deiodinases retrieved in this study (especially IODα, IODγ, and IODε) (Fig. 4).

An upper and neuroendocrine regulation of TH production seems to have appeared only recently in the vertebrate lineage. We found only a few genes (TRH, TSHR, and CRFR) in the amphioxus genome that are implicated in a higher regulation of TH production. Orthologs of these genes performing functions not related to TH signaling exist in protostomes questioning the physiological role of the amphioxus sequences we retrieved. Based on our data, and since there is no clear hypothalamic pituitary thyroid axis in amphioxus (Holland et al. 2008), we can conclude that there is probably no homologous higher regulation of TH synthesis in amphioxus comparable to that of vertebrates. It is possible that the ability of a higher regulation by this axis evolved specifically in the vertebrate lineage. In contrast, a TH regulatory system different from the one in vertebrates may exist in amphioxus, since metamorphosis is regulated by THs in amphioxus (Paris et al. 2008), and since environmental conditions probably influence the onset of metamorphosis (crowded animals metamorphose more slowly than animals kept out of close contact from each other, M.P. unpublished observation). A similar trend is observed in anurans, for which stress situations influence TH production through alteration of CRH, TRH, and TSH production (Denver 1997).

Although the upper TH synthesis regulation seems to be divergent in amphioxus, TR, the receptor of TH in the TH signaling pathway, is well conserved within chordates (Schubert et al. 2008). Indeed, the only amphioxus TR (Fig. S6) is responsive to TRIAC, a T_3 derivative, and regulates amphioxus metamorphosis (Paris et al. 2008). Its partner, RXR, also displays functional characteristics that are very well conserved (as a heterodimer partner of several NRs (Schubert et al. 2008)). The functional conservation of TR and RXR not only within chordates, but even in the last deuterostome ancestor, is a plausible hypothesis: the genome of the sea urchin *S. purpuratus* contains only one TR and one RXR (Howard-Ashby et al. 2006), but these two NRs have not been functionally characterized yet.

Specific elaboration of the TH signaling pathway in amphioxus

As discussed above, the amphioxus genome did not undergo the two WGD events that occurred specifically during vertebrate evolution (Dehal and Boore 2005; Putnam et al. 2008). In contrast, specific duplications of genes encoding members of the TH signaling pathway occurred in amphioxus. In the part of the pathway upstream of TR (TH metabolic pathway), we found eight SIS-related genes corresponding to three paralogous vertebrate genes, four PER genes corresponding to three vertebrate paralogs, and five deiodinase genes corresponding to the three deiodinases genes known in vertebrates. This is in sharp contrast to what is observed for the more downstream part of the pathway (*i.e.* the receptor part): there is only one TR (for two vertebrate genes), one RXR (for three vertebrate genes), and only one copy for each of the co-activator/co-repressor paralogous groups (Schubert et al. 2008). This feature suggests that an independent elaboration of the upper part (the TH metabolic pathway) occurred in amphioxus, whereas the lower part (the receptor part) is similar in amphioxus and vertebrates. The reason why such an elaboration occurred remains elusive and will certainly need the precise functional characterization of these duplicates to be fully understood. It is tempting to link it to the central role that TR (and the TR cofactors) play in the TH signaling pathway and which could constrain its evolution, whereas TH availability could evolve more easily (see for instance the plasticity of TH metabolism in amphibians (Boorse and Denver 2002; Callery et al. 2001; Safi et al. 2004)). In each species, TH availability could independently evolve depending on specific selection pressures, which could be reflected in the lineage-specific duplications that are reported here. For instance, the active TH in amphioxus is not T_3, but probably the poorly characterized T_3 derivative TRIAC (Paris et al. 2008). This change may reflect a series of specific alterations in the metabolic pathway controlling TH production. It would thus be very interesting to study, if the specific amphioxus duplicates of deiodinases are differentially involved in the regulation of the iodine content of T_4 and its derivatives. Two main pieces of information are needed to test this model: (1) biochemical characterization of the amphioxus duplicates of SIS, PER, and deiodinases and (2) knowledge of the TH derivatives that are present, and active, in amphioxus. We are currently addressing these issues experimentally in our laboratory. Nonetheless, the amphioxus-specific duplications described here allow us to point out that amphioxus is not "our ancestor", as it is often more or less implied in many gradualist views of chordate evolution, but rather a cousin.

Is the TH signaling pathway conserved in deuterostomes?

TH metabolism and metamorphosis have both been present at the origin of deuterostomes (Paris and Laudet 2008). In chordates, TH signaling seems to be homologous (as mentioned at the beginning of the Discussion). Data on echinoderms suggest an even more ancient origin of TH signaling: at the base of the deuterostome tree. THs are important inducers of metamorphosis in some echinoideas (sea urchins, sea biscuits, and sand dollars), asteroideas (sea stars), and ophiuoideas (brittle stars) (Hodin 2006 and

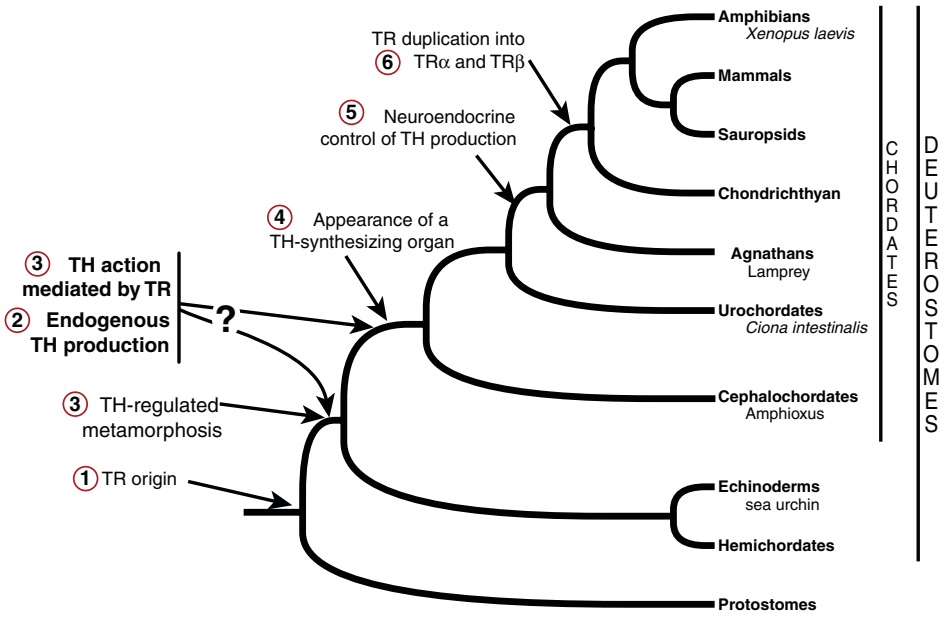


Fig. 5 Evolution of the thyroid signaling pathway. On a simplified bilaterian tree, the main steps of TH signaling pathway elaboration have been indicated. *Numbers* refer to chronological steps of the evolution of the TH signaling pathway in bilaterians and are discussed in the main text. The tree is based on Marletaz et al. (2008). Metamorphosis is to be understood in a broad sense as a TH/TR-regulated post-embryonic developmental phase characterized by ecological, metabolic, and morphological modifications (discussed in Paris et al. 2008)

references therein). Moreover, peroxidase activity-dependent T_4 production was also demonstrated (Heyland et al. 2006). We found in the sea urchin genome many genes orthologous to vertebrate TH signaling members. Consequently, it is very probable that the deuterostome ancestor was capable of endogenous TH production. However, as echinoderms do not have a recognizable endostyle/thyroid, it will be interesting to localize TH production.

There are still many deuterostome groups that have been neglected in terms of developmental and genomic studies. This is, for example, the case for hemichordates: almost nothing is known about TH signaling in this sister group of echinoderms (Fig. 5), although iodine was detected in the pharyngeal area (same body part as the endostyle/thyroid) (Ruppert 2005) supporting the notion of an ancestral origin of TH production in deuterostomes. Other deuterostome taxa that are poorly studied include sea stars, sea cucumber, crinoids as well as divergent urochordates, such as larvaceans and thaliaceans (Brusca and Brusca 2003).

The existence of a TH signaling pathway outside deuterostomes still remains elusive. Indeed, in protostomes or cnidarians, TH metabolism has been poorly studied (Eales 1997), few reports have investigated TH effects on metamorphosis, and, although TRs were cloned in several protostome species, they have not been molecularly characterized (Paris and Laudet 2008). We propose the following scenario regarding the evolution of the TH signaling pathway (Fig. 5): (1) appearance of a TR gene at the base of the bilaterians, whose function remains elusive; (2) the chordate, and likely even the deuterostome, ancestor acquired the ability to endogenously produce THs (TH production has been reported in several deuterostomes and the main members of the TH signaling pathway like SIS, PERT, IODs, TR, and some carrier proteins have been described in vertebrates, amphioxus, and, to a lesser extent, in echinoderms); (3) this process can probably be correlated with the ability to metamorphose under TH/TR regulation; (4) more complete and localized internalization of TH synthesis in a specialized organ with the evolution of a dual-function endostyle in chordates and its subsequent specialization into a thyroid within the vertebrate lineage; (5) in the vertebrate lineage, there is the appearance of a neuroendocrine control of TH synthesis using peptide hormones, such as TRH and TSH, that are not found in the amphioxus genome; (6) further elaboration of the pathway occurred in gnathostomes with the duplication of the ancestral TR into two genes, TRα and TRβ, allowing a further refinement of the pathway (there has also been a lamprey-specific TR duplication (Escriva et al. 2002)). According to this scenario, the downstream part of the pathway, namely the genes regulated by TH and TR, are derived in each species giving rise to the extraordinary diversity of morphological, physiological, and ecological rearrangements observed during metamorphosis (Paris and Laudet 2008).

Conclusions

From our study and previous data, we propose that the amphioxus TH signaling pathway is homologous to the vertebrate TH signaling pathway implying an ancient origin of TH metabolism. However, biochemical investigation on the proteins encoded by the genes described here should be carried out in the future. Outside chordates, much scarcer data are available. In some echinoderms, TH production and biological actions by TH are similar to what has been observed in chordates. Further work on echinoderms will be required to address questions such as where are THs produced and whether TR is involved in TH action in echinoderms. In order to better understand the evolution of the TH signaling pathway and its link to development and especially to metamorphosis, data on a wider range of animals should be obtained. Thus, even if a continuous effort should be maintained to keep improving our understanding of the TH signaling pathway in chordates, the sister group of chordates, the Ambulacraria (regrouping the echinoderms, the hemichordates, and xenoturbella (Marletaz et al. 2008)) and the protostomes are a “thyroidal desert” that may be worth our attention.

Acknowledgements We would like to thank Maria Theodosiou for critical reading of the manuscript. This work was supported by funds from the ANR, CNRS, and the MENRT. This study was further supported by CRESCENDO, a European Union Integrated Project of FP6, and by CASCADE, a Network of Excellence of FP6.

References

Baqui M, Botero D, Gereben B, Curcio C, Harney JW, Salvatore D, Sorimachi K, Larsen PR, Bianco AC (2003) Human type 3 iodothyronine selenodeiodinase is located in the plasma membrane and undergoes rapid internalization to endosomes. J Biol Chem 278:1206–1211

Berry MJ, Banu L, Larsen PR (1991) Type I iodothyronine deiodinase is a selenocysteine-containing enzyme. Nature 349:438–440

Bianco AC, Salvatore D, Gereben B, Berry MJ, Larsen PR (2002) Biochemistry, cellular and molecular biology, and physiological roles of the iodothyronine selenodeiodinases. Endocr Rev 23:38–89

Boorse GC, Denver RJ (2002) Acceleration of Ambystoma tigrinum metamorphosis by corticotropin-releasing hormone. J Exp Zool 293:94–98

Brusca RC, Brusca GJ (2003) Invertebrates. Sinauer Associates, Sunderland

Callery EM, Fang H, Elinson RP (2001) Frogs without polliwogs: evolution of anuran direct development. Bioessays 23:233–241

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Covelli I, Salvatore G, Sena L, Roche J (1960) Sur la formation d'hormones thyroidiennes et de leurs précurseurs par Branchiostoma lanceolatum. C R Soc Biol Paris 154:1165–1169

Curcio C, Baqui MM, Salvatore D, Rihn BH, Mohr S, Harney JW, Larsen PR, Bianco AC (2001) The human type 2 iodothyronine deiodinase is a selenoprotein highly expressed in a mesothelioma cell line. J Biol Chem 276:30183–30187

Daiyasu H, Toh H (2000) Molecular evolution of the myeloperoxidase family. J Mol Evol 51:433–445

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314

Denver RJ (1997) Environmental stress as a developmental cue: corticotropin-releasing hormone is a proximate mediator of adaptive phenotypic plasticity in amphibian metamorphosis. Horm Behav 31:169–179

Dohan O, De la Vieja A, Paroder V, Riedel C, Artani M, Reed M, Ginter CS, Carrasco N (2003) The sodium/iodide symporter (NIS): characterization, regulation, and medical significance. Endocr Rev 24:48–77

Eales JG (1997) Iodine metabolism and thyroid-related functions in organisms lacking thyroid follicles: are thyroid hormones also vitamins? Proc Soc Exp Biol Med 214:302–317

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Escriva H, Manzon L, Youson J, Laudet V (2002) Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol 19:1440–1450

Flamant F, Samarut J (2003) Thyroid hormone receptors: lessons from knockout and knock-in mutant mice. Trends Endocrinol Metab 14:85–90

Flamant F, Baxter JD, Forrest D, Refetoff S, Samuels H, Scanlan TS, Vennstrom B, Samarut J (2006) International Union of Pharmacology. LIX. The pharmacology and classification of the nuclear receptor superfamily: thyroid hormone receptors. Pharmacol Rev 58:705–711

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Fredriksson G, Öfverholm T, Ericson LE (1985) Electron-microscopic studies of iodine-binding and peroxidase activity in the endostyle of the larval amphioxus (Branchiostoma lanceolatum). Cell Tissue Res 241:257–266

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Gopal E, Umapathy NS, Martin PM, Ananth S, Gnana-Prakasam JP, Becker H, Wagner CA, Ganapathy V, Prasad PD (2007) Cloning and functional characterization of human SMCT2 (SLC5A12) and expression pattern of the transporter in kidney. Biochim Biophys Acta 1768:2690–2697

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Heyland A, Hodin J (2004) Heterochronic developmental shift caused by thyroid hormone in larval sand dollars and its implications for phenotypic plasticity and the evolution of nonfeeding development. Evolution Int J Org Evolution 58:524–538

Heyland A, Reitzel AM, Price DA, Moroz LL (2006) Endogenous thyroid hormone synthesis in facultative planktotrophic larvae of the sand dollar Clypeaster rosaceus: implications for the evolutionary loss of larval feeding. Evol Dev 8:568–579

Hodin J (2006) Expanding networks: Signaling components in and a hypothesis for the evolution of metamorphosis. Integr Comp Biol 46:719–742

Holland LZ, Albalat R, Azumi K, Benito-Gutierrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, Ferrier DE, Garcia-Fernandez J, Gibson-Brown JJ, Gissi C, Godzik A, Hallbook F, Hirose D, Hosomichi K, Ikuta T, Inoko H, Kasahara M, Kasamatsu J, Kawashima T, Kimura A, Kobayashi M, Kozmik Z, Kubokawa K, Laudet V, Litman GW, McHardy AC, Meulemans D, Nonaka M, Olinski RP, Pancer Z, Pennacchio LA, Pestarino M, Rast JP, Rigoutsos I, Robinson-Rechavi M, Roch G, Saiga H, Sasakura Y, Satake M, Satou Y, Schubert M, Sherwood N, Shiina T, Takatori N, Tello J, Vopalensky P, Wada S, Xu A, Ye Y, Yoshida K, Yoshizaki F, Yu JK, Zhang Q, Zmasek CM, de Jong PJ, Osoegawa K, Putnam NH, Rokhsar DS, Satoh N, Holland PW (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res 18:1100–1111

Howard-Ashby M, Materna SC, Brown CT, Chen L, Cameron RA, Davidson EH (2006) Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. Dev Biol 300:90–107

Hulbert AJ (2000) Thyroid hormones and their effects: a new perspective. Biol Rev Camb Philos Soc 75:519–631

Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN (2003) Characterization of mammalian selenoproteomes. Science 300:1439–1443

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Manzon RG, Holmes JA, Youson JH (2001) Variable effects of goitrogens in inducing precocious metamorphosis in sea lampreys (Petromyzon marinus). J Exp Zool 289:290–303

Manzon RG, Neuls TM, Manzon LA (2007) Molecular cloning, tissue distribution, and developmental expression of lamprey transthyretins. Gen Comp Endocrinol 151:55–65

Markov GV, Paris M, Bertrand S, Laudet V (2008) The evolution of the ligand/receptor couple: a long road from comparative endocrinology to comparative genomics. Mol Cell Endocrinol 293:5–16

Marletaz F, Gilles A, Caubit X, Perez Y, Dossat C, Samain S, Gyapay G, Wincker P, Le Parco Y (2008) Chaetognath transcriptome reveals ancestral and unique features among bilaterians. Genome Biol 9:R94

Monaco F, Dominici R, Andreoli M, Pirro RD, Roche J (1981) Thyroid hormone formation in thyroglobulin synthesized in the amphioxus (Branchiostoma lanceolatum Pallas). Comp Biochem Physiol B 70:341–343

Novinec M, Kordis D, Turk V, Lenarcic B (2006) Diversity and evolution of the thyroglobulin type-1 domain superfamily. Mol Biol Evol 23:744–755

Ogasawara M (2000) Overlapping expression of amphioxus homologs of the thyroid transcription factor-1 gene and thyroid peroxidase gene in the endostyle: insight into evolution of the thyroid gland. Dev Genes Evol 210:231–242

Ogasawara M, Di Lauro R, Satoh N (1999) Ascidian homologs of mammalian thyroid peroxidase genes are expressed in the thyroid-equivalent region of the endostyle. J Exp Zool 285:158–169

Ohno S (1970) Evolution by gene duplication. Springer, Berlin

Paris M, Escriva H, Schubert M, Brunet F, Brtko J, Ciesielski F, Roecklin D, Vivat-Hannah V, Jamin EL, Cravedi JP, Scanlan TS, Renaud JP, Holland ND, Laudet V (2008) Amphioxus postembryonic development reveals the homology of chordate metamorphosis. Curr Biol 18:825–830

Paris M, Laudet V (2008) The history of developmental stages: metamorphosis in chordates. Genesis, in press

Patricolo E, Cammarata M, D'Agati P (2001) Presence of thyroid hormones in ascidian larvae and their involvement in metamorphosis. J Exp Zool 290:426–430

Power DM, Llewellyn L, Faustino M, Nowell MA, Bjornsson BT, Einarsdottir IE, Canario AV, Sweeney GE (2001) Thyroid hormones in growth and development of fish. Comp Biochem Physiol C Toxicol Pharmacol 130:447–459

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutierrez EL, Dubchak I, Garcia-Fernandez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin IT, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071

Robinson M, Gouy M, Gautier C, Mouchiroud D (1998) Sensitivity of the relative-rate test to taxonomic sampling. Mol Biol Evol 15:1091–1098

Rodriguez AM, Perron B, Lacroix L, Caillou B, Leblanc G, Schlumberger M, Bidart JM, Pourcher T (2002) Identification and characterization of a putative human iodide transporter located at the apical membrane of thyrocytes. J Clin Endocrinol Metab 87:3500–3503

Ruppert EE (2005) Key character uniting hemichordates and chordates: homologies or homoplasies. Can J Zool 83:8–23

Safi R, Bertrand S, Marchand O, Duffraisse M, de Luze A, Vanacker JM, Maraninchi M, Margotat A, Demeneix B, Laudet V (2004) The axolotl (Ambystoma mexicanum), a neotenic amphibian, expresses functional thyroid hormone receptors. Endocrinology 145:760–772

Schreiber G, Richardson SJ (1997) The evolution of gene expression, structure and function of transthyretin. Comp Biochem Physiol B Biochem Mol Biol 116:137–160

Schubert M, Brunet F, Paris M, Bertrand S, Benoit G, Laudet V (2008) Nuclear hormone receptor signaling in amphioxus. Dev Genes Evol, in press

Shepherdley CA, Klootwijk W, Makabe KW, Visser TJ, Kuiper GGJM (2004) An ascidian homolog of vertebrate iodothyronine deiodinases. Endocrinology 145:1255–1268

Shi Y-B (2000) Amphibian metamorphosis: from morphology to molecular biology. Wiley, New York

Tata JR (2006) Amphibian metamorphosis as a model for the developmental actions of thyroid hormone. Mol Cell Endocrinol 246:10–20

Vandenborne K, Roelens SA, Darras VM, Kuhn ER, Van der Geyten S (2005) Cloning and hypothalamic distribution of the chicken thyrotropin-releasing hormone precursor cDNA. J Endocrinol 186:387–396

Wu SY, Green WL, Huang WS, Hays MT, Chopra IJ (2005) Alternate pathways of thyroid hormone metabolism. Thyroid 15:943–958

Yen PM (2001) Physiological and molecular basis of thyroid hormone action. Physiol Rev 81:1097–1142

Youson JH (1997) Is lamprey metamorphosis regulated by thyroid hormones? Amer Zool 37:441–460

ARTICLES

nature biotechnology

Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*

Pierre Abad[1–3], Jérôme Gouzy[4], Jean-Marc Aury[5–7], Philippe Castagnone-Sereno[1–3], Etienne G J Danchin[1–3], Emeline Deleury[1–3], Laetitia Perfus-Barbeoch[1–3], Véronique Anthouard[5–7], François Artiguenave[5–7], Vivian C Blok[8], Marie-Cécile Caillaud[1–3], Pedro M Coutinho[9], Corinne Dasilva[5–7], Francesca De Luca[10], Florence Deau[1–3], Magali Esquibet[11], Timothé Flutre[12], Jared V Goldstone[13], Noureddine Hamamouch[14], Tarek Hewezi[15], Olivier Jaillon[5–7], Claire Jubin[5–7], Paola Leonetti[10], Marc Magliano[1–3], Tom R Maier[15], Gabriel V Markov[16,17], Paul McVeigh[18], Graziano Pesole[19,20], Julie Poulain[5–7], Marc Robinson-Rechavi[21,22], Erika Sallet[23,24], Béatrice Ségurens[5–7], Delphine Steinbach[12], Tom Tytgat[25], Edgardo Ugarte[5–7], Cyril van Ghelder[1–3], Pasqua Veronico[10], Thomas J Baum[15], Mark Blaxter[26], Teresa Bleve-Zacheo[10], Eric L Davis[14], Jonathan J Ewbank[27], Bruno Favery[1–3], Eric Grenier[11], Bernard Henrissat[9], John T Jones[8], Vincent Laudet[16], Aaron G Maule[18], Hadi Quesneville[12], Marie-Noëlle Rosso[1–3], Thomas Schiex[24], Geert Smant[25], Jean Weissenbach[5–7] & Patrick Wincker[5–7]

Plant-parasitic nematodes are major agricultural pests worldwide and novel approaches to control them are sorely needed. We report the draft genome sequence of the root-knot nematode *Meloidogyne incognita*, a biotrophic parasite of many crops, including tomato, cotton and coffee. Most of the assembled sequence of this asexually reproducing nematode, totaling 86 Mb, exists in pairs of homologous but divergent segments. This suggests that ancient allelic regions in *M. incognita* are evolving toward effective haploidy, permitting new mechanisms of adaptation. The number and diversity of plant cell wall–degrading enzymes in *M. incognita* is unprecedented in any animal for which a genome sequence is available, and may derive from multiple horizontal gene transfers from bacterial sources. Our results provide insights into the adaptations required by metazoans to successfully parasitize immunocompetent plants, and open the way for discovering new antiparasitic strategies.

Plant-parasitic nematodes are responsible for global agricultural losses amounting to an estimated $157 billion annually. Although chemical nematicides are the most reliable means of controlling root-knot nematodes, they are increasingly being withdrawn owing to their toxicity to humans and the environment. Novel and specific targets are thus needed to develop new strategies against these pests.

The Southern root-knot nematode *Meloidogyne incognita* is able to infect the roots of almost all cultivated plants, making it perhaps the

[1]INRA, UMR 1301, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [2]CNRS, UMR 6243, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [3]UNSA, UMR 1301, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [4]Laboratoire Interactions Plantes Micro-organismes, UMR441/2594, INRA/CNRS, Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [5]Genoscope (CEA), 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France. [6]CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France. [7]Université d'Evry, F-91057 Evry, France. [8]Plant Pathology Programme, SCRI, Invergowrie, Dundee DD2 5DA, UK. [9]CNRS, UMR 6098 CNRS and Universités d'Aix-Marseille I & II, Case 932, 163 Av. de Luminy, F-13288 Marseille, France. [10]Istituto per la Protezione delle Piante, Consiglio Nazionale delle Ricerche, Via G. Amendola 165/a, 70126 Bari, Italy. [11]INRA, Agrocampus Rennes, Univ. Rennes I, UMR1099 BiO3P, Domaine de la Motte, F-35653 Le Rheu Cedex, France. [12]INRA, UR1164 Unité de Recherche en Génomique et Informatique (URGI), 523 place des terrasses de l'Agora, F-91034 Evry, France. [13]Biology Department, Woods Hole Oceanographic Institution, Co-op Building, MS #16, Woods Hole, Massachusetts 02543, USA. [14]Department of Plant Pathology, North Carolina State University, 840 Method Road, Unit 4, Box 7903 Raleigh, North Carolina 27607,USA. [15]Department of Plant Pathology, Iowa State University, 351 Bessey Hall, Ames, Iowa 50011, USA. [16]Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, Molecular Zoology team, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland, Lyon Sud, 46 allée d'Italie, F-69364 Lyon Cedex 07, France. [17]USM 501, Evolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, 7 rue Cuvier, F-75005 Paris, France. [18]Biomolecular Processes: Parasitology, School of Biological Sciences, Medical Biology Centre, 97 Lisburn Road, Queen's University Belfast, Belfast BT9 7BL, UK. [19]Dipartimento di Biochimica e Biologia Molecolare "E. Quagliariello", University of Bari, Via Orabona 4, 70126 Bari, Italy. [20]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Via G. Amendola, 122/D, 70126 Bari, Italy. [21]Department of Ecology and Evolution, University of Lausanne, UNIL-Sorge, Le Biophore, CH-1015 Lausanne, Switzerland. [22]Swiss Institute of Bioinformatics, quartier sorge, Bâtiment Genopode, CH-1015 Lausanne, Switzerland. [23]Plateforme Bioinformatique du Génopole Toulouse Midi-Pyrénées, GIS Toulouse Genopole, 24 Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [24]Unité de Biométrie et d'Intelligence Artificielle UR875, INRA, Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [25]Laboratory of Nematology, Wageningen University, Binnenhaven 5, 6709PD Wageningen, The Netherlands. [26]Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK. [27]INSERM/CNRS/Université de la Méditerranée, Centre d'Immunologie de Marseille-Luminy, 163 av. de Luminy, Case 906, F-13288, Marseille cedex 09, France. Correspondence should be addressed to P.A. (pierre.abad@sophia.inra.fr).

Received 30 April; accepted 25 June; published online 27 July 2008; doi:10.1038/nbt.1482

ARTICLES

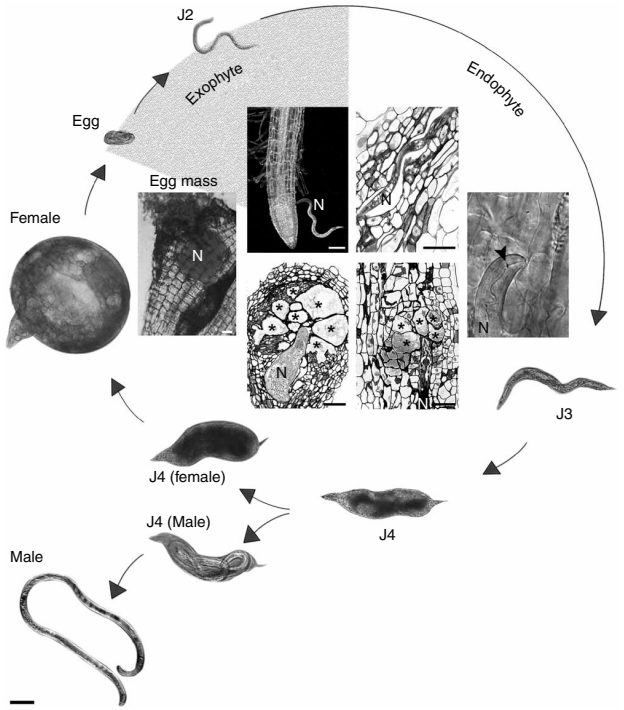


Figure 1 The parasitic life cycle of *Meloidogyne incognita*. Infective second-stage juveniles (J2) penetrate the root and migrate between cells to reach the plant vascular cylinder. The stylet (arrowhead) connected to the esophagus is used to pierce plant cell walls, to release esophageal secretions and to take up nutrients. Each J2 induces the dedifferentiation of five to seven root cells into multinucleate and hypertrophied feeding cells (*). These giant cells supply nutrients to the nematode (N). The nematode becomes sedentary and goes through three molts (J3, J4, adult). Occasionally, males develop and migrate out of the roots. However, it is believed that they play no role in reproduction. The pear-shaped female produces eggs that are released on the root surface. Embryogenesis within the egg is followed by the first molt, generating second-stage juveniles (J2). Scale bars, 50 μm.

most damaging of all crop pathogens[1]. *M. incognita* is an obligatory sedentary parasite that reproduces by mitotic parthenogenesis[2]. Root-knot nematodes have an intimate interaction with their hosts. Within the host root, adult females induce the redifferentiation of root cells into specialized 'giant' cells, upon which they feed continuously (**Fig. 1**). *M. incognita* can infect *Arabidopsis thaliana*, making this nematode a key model system for the understanding of metazoan adaptations to plant parasitism[3,4] (**Supplementary Data**, section 1 online).

The phylum Nematoda comprises >25,000 described species, many of which are parasites of animals or plants[2]. As many as 10 million species may have yet to be described. Although the model free-living nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* have been the subjects of intensive study[5,6], little is known about the other members of this diverse phylum. These two free-living models will likely not illuminate the biology of nematode parasitism (**Supplementary Fig. 1** online), as shown by the substantial differences between their genome sequences and that of the human parasite *Brugia malayi*[7].

The genome sequence of *M. incognita* presented here provides insights into the adaptations required by metazoans to successfully parasitize and counter defenses of immunocompetent plants, and suggests new antiparasitic strategies.

RESULTS

General features of the *M. incognita* genome

The *M. incognita* genome was sequenced using whole-genome shotgun strategy. Assembly with Arachne[8] yielded 2,817 supercontigs, totaling 86 Mb (**Table 1**; **Supplementary Data**, section 2; **Supplementary Fig. 2**; **Supplementary Table 1** online)—almost twice the estimated genome size (47- to 51-Mb haploid genome)[9]. All-against-all comparison of supercontigs revealed that 648 of the longest (covering ~55 Mb) consist of homologous but diverged segment pairs (**Fig. 2**) that might represent former alleles (**Supplementary Data**, section 2; **Supplementary Figs. 3** and **4** online). About 3.35 Mb of the assembly constitutes a third partial copy aligning with these supercontig pairs. Average sequence divergence between the aligned regions is ~8% (**Fig. 3**). A combination of different processes may explain the observed pattern in *M. incognita*, including polyploidy, polysomy, aneuploidy and hybridization[10,11]; all are frequently associated with asexual reproduction. These observations are consistent with a strictly mitotic parthenogenetic reproductive mode, which can permit homologous chromosomes to diverge considerably, as hypothesized for bdelloid rotifers[12] (**Supplementary Data**, section 2.2). No DNA attributable to bacterial endosymbiont genome(s) was identified.

Noncoding DNA repeats and transposable elements represent 36% of the *M. incognita* genome (**Supplementary Data**, section 3; **Supplementary Figs. 5** and **6** and **Supplementary Tables 2** and **3** online). One repeat family with 283 members on 46 contigs encoded the nematode *trans*-spliced leader (SL) exon, SL1, of which 258 members were found associated with a satellite DNA[13] (**Supplementary Fig. 7** online). In nematodes, many mature mRNAs share this 5′ SL exon, and *trans*-splicing is also associated with resolution of polycistronic pre-mRNAs derived from operons. We identified 1,585 candidate

Table 1 General features of the *Meloidogyne incognita* genome in comparison with the genomes of *B. malayi*[7] and *C. elegans*[5]

Features	*M. incognita*	*B. malayi*	*C. elegans*
Overall			
Estimated size of genome (Mb)	47–51[a]	90–95[a]	100[a]
Total size of assembled sequence (Mb)	86	88	100
Number of scaffolds and/or chromosomes (chr.)	2,817	8,180	6 chr.
G + C content (%)	31.4	30.5	35.4
Protein-coding regions			
Number of protein-coding gene models	19,212	11,515	20,072
Protein-coding sequence (% of genome)	25.3	17.8	25.5
Maximum/average protein length (amino acids)	5,970/354	9,420/343	18,562/440
Mean length of intergenic region (bp)	1,402	3,783	2,218
Gene density (genes per Mb)	223	162	228
Operon number	1,585	926	1,118
Percent of genes present in operon	19	18	14

For *B. malayi* a gene count ranging from 14,500 to 17,800 was inferred after inclusion of genes in the unannotated portion of the genome[7]. For *C. elegans* the gene and protein count is according to Wormpep database (WS183 release).
[a]*M. incognita*: flow cytometry[9]; *B. malayi*: flow cytometry and clone-based[7]; *C. elegans* genome has been completely sequenced telomere to telomere (no gaps) and is exactly 100,291,840 bp[45].

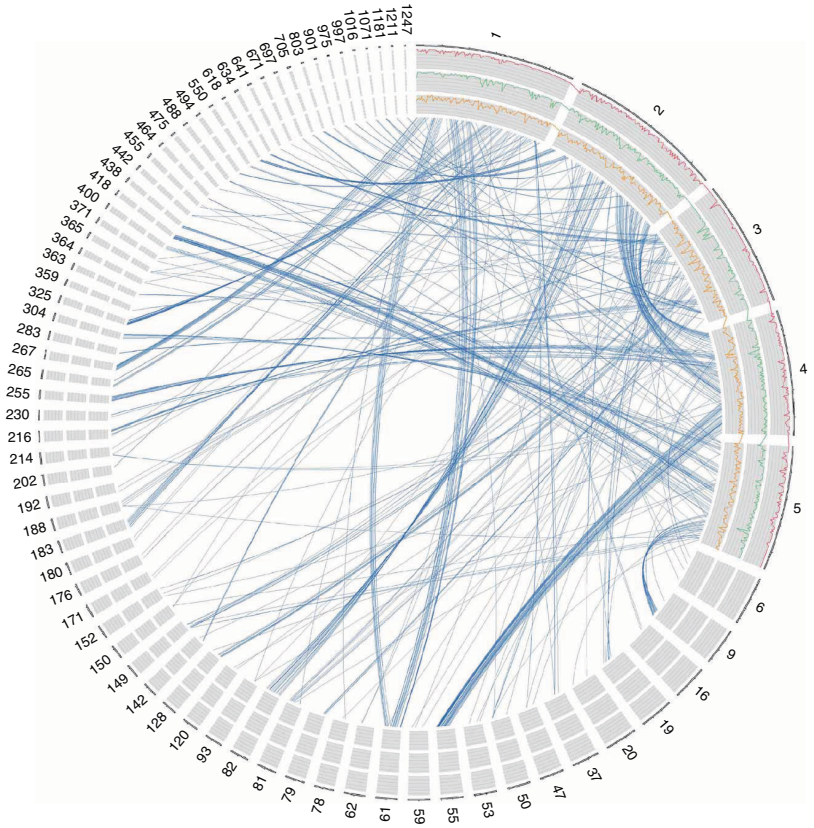


Figure 2 Allelic-like relationships for the five largest supercontigs of the *M. incognita* assembly. The five largest supercontigs are shown with plots of gene density (orange curve), conservation with *C. elegans* at amino acid level (green curve) and EST density (pink curve). Blue lines represent most similar matches at the protein level between each predicted gene on these five supercontigs and 70 matching supercontigs.

M. incognita operons containing a total of 3,966 genes. The two longest operons contained ten genes each and are not allelic copies (**Supplementary Table 4** online). Operons are a dynamic component of nematode genome architecture, as different sets of genes were operonic in *M. incognita*, *C. elegans* and *B. malayi*, and only one operon was found to be strictly conserved between the three nematodes (**Supplementary Data**, section 4; **Supplementary Figs. 8** and **9**; **Supplementary Table 5** online).

The gene content of a plant-parasitic nematode

The genome sequence was annotated using the integrative gene prediction platform EuGene[14], specifically trained for *M. incognita* (**Supplementary Data**, section 5; **Supplementary Table 6** online). We identified 19,212 protein-coding genes (**Table 1**). Due to the high variation between allelic-like copies (**Fig. 3**) potentially allowing functional divergence, all copies were considered to be different genes. Indeed, 69% of protein sequences were <95% identical to any other (**Supplementary Table 7** and **Supplementary Fig. 10** online). The protein-coding genes occupy 25.3% of the sequence at an average density of 223 genes Mb^{-1}, and 36% are supported by expressed sequence tags (ESTs). InterPro protein domains were identified in 55% of proteins and 22% were predicted to be secreted. Comparison of domain occurrence in *M. incognita* with that in *C. elegans* identified an increased abundance of 'pectate lyase', glycoside hydrolase family GH5 and peptidase C48 (SUMO) domains, and fewer chemoreceptor domains. We compared the domain content of the *M. incognita* protein set to those of *C. elegans*, *B. malayi*, *Drosophila melanogaster* and three fungi, of which two are plant pathogens. Thirty-two domains were detected only in *M. incognita*, and two additional domains were only shared between the two plant-pathogenic fungi and *M. incognita*. Functions assigned to the 34 domains specific to plant pathogens encompassed plant cell-wall degradation and chorismate mutase activity (see below). OrthoMCL[15] clustering of the same eight proteomes suggested that 52% of *M. incognita* predicted proteins had no ortholog in the other species. Among them, 1,819 proteins (of which 338 were supported by ESTs) are secreted and lack any known domain (**Supplementary Data**, section 6; **Supplementary Figs. 11** and **12**; **Supplementary Tables 8–10** online). The core complement of proteins in the phylum Nematoda is relatively small: ~23% of the ortholog groups were shared by *M. incognita*, *C. elegans* and *B. malayi* (**Supplementary Fig. 12b**).

Identifying plant parasitism genes

Nematode proteins produced in and secreted from specialized gland cells into the host are likely to be important effectors of plant parasitism[4,16]. We identified gene products that might be involved in parasitic interaction, particularly those that might modify plant cell walls.

M. incognita has an unprecedented set of 61 plant cell wall–degrading, carbohydrate-active enzymes (CAZymes). Although a few such individual CAZymes had been identified previously in some plant-parasitic nematodes and in two insect species[4,16,17], they are absent from all other metazoans studied to date (**Table 2**; **Supplementary Data**, section 7.1; **Supplementary Tables 11–14** online). We identified 21 cellulases and six xylanases from family GH5, two polygalacturonases from family GH28 and 30 pectate lyases from family PL3. We also identified CAZymes not previously reported from metazoans, including two additional plant cell wall–degrading arabinases (family GH43) and two invertases (family GH32). Invertases catalyze the conversion of sucrose (an abundant disaccharide in plants) into glucose and fructose, which can be used by *M. incognita* as a carbon source. We also identified a total of 20 candidate expansins in *M. incognita*, which may disrupt noncovalent bonds in plant cell walls, making the components more accessible to plant cell wall–degrading enzymes[18]. This suite of plant cell wall–degrading CAZymes, expansins and associated invertases was probably acquired by horizontal gene transfer (HGT), as the most similar proteins (outside plant-parasitic nematodes) were bacterial homologs (**Supplementary Table 12**). *M. incognita* also has four secreted chorismate mutases[19], which most closely resemble bacterial enzymes. Chorismate mutase is a key enzyme in biosynthesis of aromatic amino acids and related products, and *M. incognita* may subvert host tyrosine-dependant lignification or defense responses.

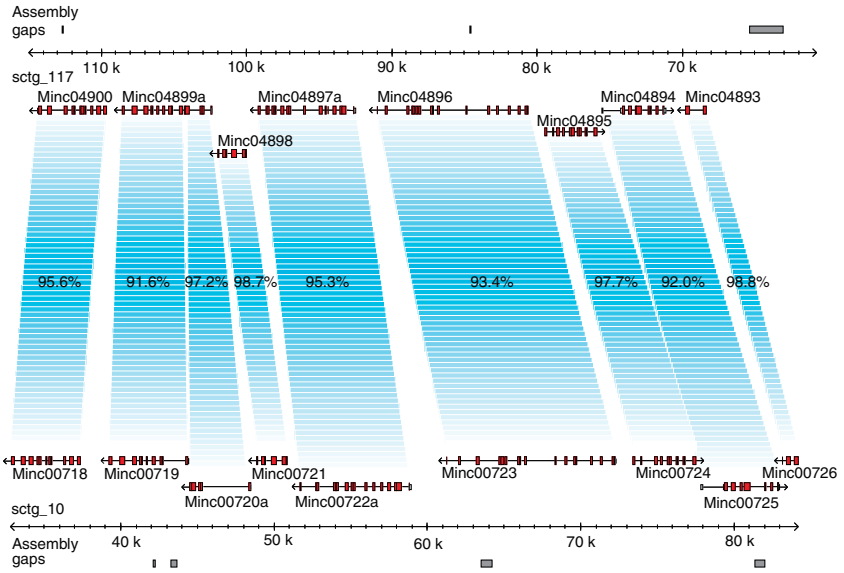


Figure 3 Example of two allelic-like regions in the *Meloidogyne incognita* assembly. Exons are represented by red boxes and are linked together to form genes (arrows indicate the direction of transcription). Gray boxes show assembly gaps. Highly diverged allelic genes are linked together using blue boxes. Gene order is well conserved between the two allelic-like regions, with only minor differences in predicted gene structure. Percentages of sequence identity at the protein level between the two allelic-like regions are indicated.

Overall, these genes suggest a critical role of HGT events in the evolution of plant parasitism within root-knot nematodes.

Apart from genes restricted to *M. incognita*, we also identified gene families showing substantial expansion compared to *C. elegans*. Among the most notable idiosyncrasies in *M. incognita*, we identified more than 20 cysteine proteases of the C48 SUMO (small ubiquitin-like modifier) deconjugating enzyme family—four times the number in *C. elegans* (**Supplementary Data**, section 7.2; **Supplementary Table 15** online). As some phytopathogenic bacterial virulence factors are SUMO proteases[20], the proteolysis of sumoylated host substrates may be a general strategy used by pathogens to manipulate host plant signal transduction. The *M. incognita* genome also encodes nine serine proteases from the S16 sub-family (Lon proteases), whereas only three are identified in *C. elegans*. These proteases regulate type III protein secretion in phytopathogenic bacteria[21] and may have analogous roles in *M. incognita*.

We identified orthologs to other known candidate plant-parasitic nematode parasitism genes in the genome of *M. incognita*. As most of these gene families are also present in animal-parasitic nematodes and *C. elegans*, *M. incognita* members putatively involved in parasitism were probably recruited from ancestral nematode families (**Supplementary Data**, section 7.3; **Supplementary Table 16** online). Twenty-seven previously described *M. incognita*–restricted pioneer genes expressed in esophageal glands[22] were retrieved in the genome. Eleven additional copies were identified; all remain *Meloidogyne spp.* specific (**Supplementary Data**, section 7.4; **Supplementary Table 17** online). These secreted proteins of as-yet-unknown function are likely targets for novel intervention strategies, and warrant deeper investigation.

Protection against environmental stresses

One aspect of plant defense responses is the production of cytotoxic oxygen radicals. However, *M. incognita* has fewer genes encoding superoxide dismutases and glutathione peroxidases than *C. elegans* (**Supplementary Data**, section 7.5; **Supplementary Table 18** online). More striking still was the reduction in glutathione *S*-transferases (GSTs) and cytochromes P450 (CYPs), enzymes involved in xenobiotic metabolism and protection against peroxidative damage. Whereas *C. elegans* has 44 GSTs, including representatives from the Omega, Sigma and Zeta classes[23], *M. incognita* possesses only 5 GSTs, all from the Sigma class. Sigma class GSTs are involved in protection against oxidants rather than xenobiotics. A comparable reduction in *gst* genes was observed in *B. malayi*[7]. Similarly, whereas *C. elegans* has 80 different *cyp* genes from 16 families[24], only 27 full or partial *cyp* genes, from 8 families, were identified in *M. incognita*. CYP35 and other families of xenobiotic-metabolizing P450s are absent from *M. incognita* (**Supplementary Data**, section 7.5; **Supplementary Table 18**).

We identified *M. incognita* orthologs of all genes of the innate immunity signaling pathways of *C. elegans*[25] except *trf-1*, which is part of the Toll pathway (**Supplementary Data**, section 7.5; **Supplementary Table 19** online). However, immune effectors such as lysozymes, C-type lectins and chitinases were much less abundant in *M. incognita* than in *C. elegans*. As previously observed in *B. malayi*[7], entire classes of immune effectors known from *C. elegans* were absent from *M. incognita*, including antibacterial genes such as *abf* and *spp*[26] and antifungal genes of several classes (*nlp*, *cnc*, *fip*, *fipr*)[25] (**Supplementary Data**, section 7.5; **Supplementary Table 19**). As plant parasites embedded in root tissues are protected from a variety of biotic and abiotic stresses, we speculate that the reduction and specialization of chemical and immune defense genes is a result of life in this privileged environment.

C. elegans has a broad range of unusual fucosylated N-glycan structures compared to other metazoans[27]. *M. incognita* has almost twice as many candidate fucosyltransferases as *C. elegans* (**Supplementary Data**, section 7.1; **Supplementary Table 14**). As suggested for animal-parasitic nematodes, multi-fucosylated structures on the surface of the nematode cuticle could help *M. incognita* to evade recognition[27].

Table 2 ***Meloidogyne incognita*** **enzymes with predicted plant cell wall–degrading activities, compared with those in** ***C. elegans*** **and** ***D. melanogaster***

Substrate	Cellulose	Xylan	Arabinan	Pectin		Other	
Family	GH5 (cel)	GH5 (xyl)	GH43	GH28	PL3	EXPN	Total
M. incognita	21	6	2	2	30	20	81
C. elegans	0	0	0	0	0	0	0
D. melanogaster	0	0	0	0	0	0	0

Number of genes encoding enzymes with candidate activity on different substrate is listed in the three selected species. GH, glycoside hydrolases; PL, polysaccharide lyases; EXPN, expansin-like proteins, following the CAZy nomenclature (http://www.cazy.org/). A total of nine and two cellulose-binding modules of family CBM2 (bacterial type) were found appended to candidate expansins and cellulases, respectively.

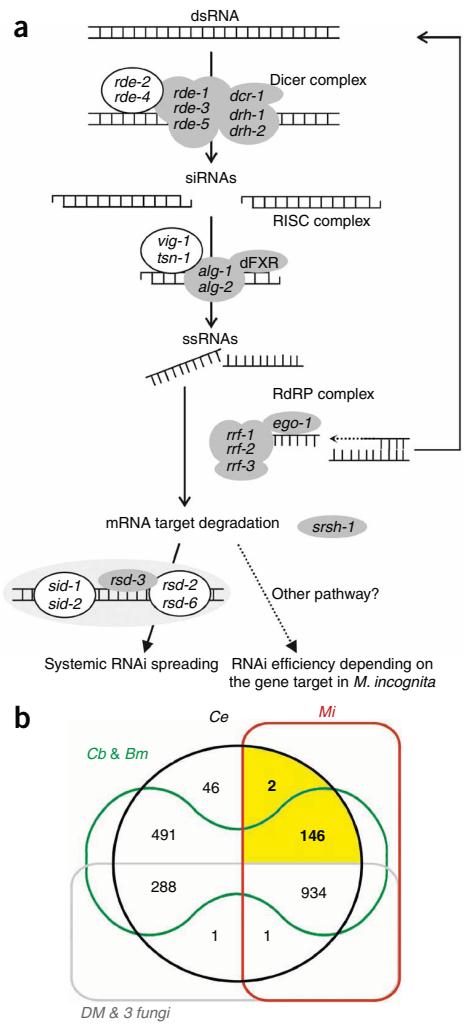


Figure 4 RNAi pathway and lethal targets. (**a**) Comparison of the RNAi pathway genes of *C. elegans* and *M. incognita*. A gray background indicates that at least one homologous gene was found in *M. incognita,* and a white background indicates that no homologous gene was found in *M. incognita.* (**b**) Distribution of orthologs to *C. elegans* lethal RNAi genes (Ce, black) between *M. incognita* (Mi, red), *C. briggsae* and *B. malayi* (Cb & Bm, green), *D. melanogaster* and three fungi, *N. crassa*, *G. zea* and *M. grisea* (Dm & 3 fungi, gray) using OrthoMCL. A yellow background indicates 148 nematode-only gene clusters.

Core biological processes

Nuclear receptors, kinases, G-protein coupled receptors (GPCRs) and neuropeptides encompass some of the gene products most extensively involved in core physiological, developmental and regulatory processes.

C. elegans has a surprisingly large number of nuclear receptors, but curiously lacks orthologs of many nuclear receptor types conserved in other animals[28]. Some of these conserved nuclear receptors are present in *B. malayi*[7]. Among the 92 predicted nuclear receptors in *M. incognita*, we identified orthologs of several known nematode nuclear receptors, although many of the nuclear receptors present in *B. malayi* and absent in *C. elegans* were also absent in *M. incognita* (**Supplementary Data**, section 7.6; **Supplementary Table 20** online). Many *C. elegans* nuclear receptors are classified as supplementary nuclear receptors (SupNRs), likely derived from a hepatocyte nuclear factor-4-like ancestor[29]. Orthologs of SupNRs were found in *M. incognita*, including a 41-member, *M. incognita*-specific expansion. Fourteen SupNRs are one-to-one orthologs between *B. malayi*, *M. incognita* and *C. elegans*, or conserved only between *M. incognita* and *C. elegans*, with secondary losses in *B. malayi* (**Supplementary Data**, section 7.6; **Supplementary Fig. 13** online). Thus the expansion of SupNRs started before the *Brugia-Meloidogyne-Caenorhabditis* split and has proceeded independently in *C. elegans* and *M. incognita*.

M. incognita has 499 predicted kinases compared to 411 in *C. elegans*[30] and 215 in *B. malayi*[7]. The kinases were grouped into 232 OrthoMCL clusters, 24 of which contained only nematode members, suggesting that they have nematode-specific functions. Four kinase families contained only *M. incognita* and *B. malayi* members, suggesting potential roles for these genes in parasitism. Finally, 66 kinase families, containing 122 genes, appear to be *M. incognita*-specific (**Supplementary Data**, section 7.7; **Supplementary Table 21** online). Seven percent (1,280) of all *C. elegans* genes are predicted to encode GPCRs that play crucial roles in chemosensation. These *C. elegans* genes have been divided into three serpentine receptor superfamilies and five solo families[31]. *M. incognita* has only 108 GPCR genes and these derive from two of the three serpentine receptor superfamilies and one of the solo families. These *M. incognita* chemosensory genes are commonly found as duplicates clustered on the genome, as observed in *C. elegans* (**Supplementary Data**, section 7.8; **Supplementary Fig. 14**; **Supplementary Table 22** online).

Neuropeptide diversity is remarkably high in nematodes, given the structural simplicity of their nervous systems. *C. elegans* has 28 Phe-Met-Arg-Phe-amide-like peptide (*flp*) and 35 neuropeptide-like protein (*nlp*) genes encoding ~200 distinct neuropeptides[32]. The identified neuropeptide complement of *M. incognita* is smaller: 19 *flp* genes and 21 *nlp* genes. However, two *flp* genes, *Mi-flp-30* and *Mi-flp-31*, encode neuropeptides that have not been identified in *C. elegans*, suggesting that they could fulfill functions specific to a phytoparasitic lifestyle (**Supplementary Data**, section 7.9; **Supplementary Table 23** online).

The XX-XO sex determination pathway in *C. elegans* is intimately linked to the dosage compensation pathway[33]. *M. incognita* reproduces exclusively by mitotic parthenogenesis, and males do not contribute genetically to production of offspring[11]. *M. incognita* also displays an environmental influence on sex determination: under less favorable environmental conditions far more males are produced. These males can arise due to sex reversal[34] and intersexual forms can be produced. *M. incognita* homologs of at least one member of each step of the *C. elegans* sex determination cascade were identified, including *sdc-1* from the dosage compensation pathway, *tra-1*, *tra-3* and *fem-2* from the sex determination pathway itself, and also downstream genes such as *mag-1* (which represses male-promoting genes) and *mab-23* (which controls male differentiation and behavior). In addition, a large family (~35 genes) of *M. incognita* secreted proteins, similar to the C2H2 zinc finger motif–containing *tra-1* from *C. elegans*, was identified (**Supplementary Data**, section 7.10; **Supplementary Table 24** online). It is therefore possible that *M. incognita* uses a similar genetic system for sex determination, but with the male pathway also modulated in response to environmental cues.

Taken together, these comparative analyses of genes, underpinning important traits, highlight the huge biodiversity in the phylum Nematoda. Idiosyncrasies identified in *M. incognita* may account for

its parasitic lifestyle and lead to the development of new control strategies directed against plant-parasitic nematodes.

RNA interference and lethal phenotypes

RNA interference (RNAi) is a promising technology for the functional analysis of parasitic nematode genes. RNAi can be induced in *M. incognita* by feeding, with variable silencing efficiencies depending on the gene target[35,36]. *M. incognita* has many genes of the *C. elegans* RNAi pathway, including components of the amplification complex (*ego-1*, *rrf-1*, *rrf-2* and *rrf-3*). However, we found no homologs of *sid-1*, *sid-2*, *rsd-2* and *rsd-6*, which are genes involved in systemic RNAi and double-stranded RNA spreading to surrounding cells (**Fig. 4**, **Supplementary Data**, section 7.11; **Supplementary Table 25** online). These genes are also absent from *B. malayi*[7] and *Haemonchus contortus*[37], suggesting that systematic RNAi may spread through the action of novel or poorly conserved factors. We retrieved 2,958 *C. elegans* genes having a lethal RNAi phenotype and searched for orthologs in *M. incognita*. Among the 1,083 OrthoMCL families identified, 148 (containing 344 *M. incognita* genes) appear to be nematode specific (**Supplementary Data**, section 7.12). Because of their lethal RNAi phenotype and distinctive sequence properties, these genes provide an attractive set of new antiparasite drug targets.

DISCUSSION

The genome of *M. incognita* has many traits that render it particularly attractive for studying the fundamentals of plant parasitism in the Nematoda. One remarkable feature is that most of the genome is composed of pairs of homologous segments that may denote former diverged alleles. This suggests that *M. incognita* is evolving without sex toward effective haploidy through the Meselson effect[38–40]. As the *M. incognita* genome is the first one sequenced and assembled for a strictly parthenogenetic species, we expect that its comparison with sexual nematode genomes will shed light on mechanisms leading to its peculiar structure. Functional divergence between ancient alleles of genes involved in the host-parasite interface could explain the extremely wide host range and geographic distribution of this polyphagous nematode. Analysis of the gene content of *M. incognita* revealed a suite of plant cell wall–degrading enzymes, which has no equivalent in any animal studied to date. The striking similarity of these enzymes to bacterial homologs suggests that these genes were acquired by multiple HGT events. Just as many instances of bacterial HGT involve sets of genes implicated in adaptations to new hosts or food sources, the candidate HGT events in *M. incognita* involve genes with potential roles in interactions with hosts. The alternative hypothesis—that these genes were acquired vertically from a common ancestor of bacteria and nematodes and lost in most eukaryote lineages—appears less parsimonious. Other singularities encompass *M. incognita*-restricted secreted proteins or lineage-specific expansions and/or reductions that may play roles in host-parasite interaction.

Transcriptional profiling, proteomic analysis and high throughput RNAi strategies are in progress and will lead to a deeper understanding of the processes by which a nematode causes plant disease. Combining such knowledge with functional genomic data from the model host plant *A. thaliana* should provide new insights into the intimate molecular dialog governing plant-nematode interactions and allow the further development of target-specific strategies to limit crop damage. Through the use of comparative genomics, the availability of free-living, animal- and plant-parasitic nematode genomes should provide new insights into parasitism and niche adaptation.

METHODS

Strain and DNA extraction. We used the *M. incognita* strain 'Morelos' from the root-knot nematode collection held at INRA (Institut National de la Recherche Agronomique) Sophia Antipolis, France. Nematode eggs were collected in a sterile manner from tomato roots and checked for the presence of plant material contaminants. DNA was extracted as described in **Supplementary Methods**, section 8.1 online.

Genome sequencing and assembly. We obtained paired-end sequences from plasmid and BAC libraries with the Sanger dideoxynucleotide technology on ABI3730xl DNA analyzers. The 1,000,873 individual reads were assembled in 2,817 supercontigs using Arachne[8] (**Supplementary Methods**, section 8.2; **Supplementary Table 26** online).

Genome structure, operons and noncoding elements. The assembled genome was searched for repetitive and non-coding elements. Scaffolds were aligned to determine pairs and triplets of allelic-like regions. Gene positions along scaffolds were used to predict clusters of genes forming putative operons (**Supplementary Methods**, section 8.3–8.7).

Prediction of protein coding genes. Gene predictions were performed using EuGene[14], optimized for *M. incognita* models and tested on a data set of 230 nonredundant, full-length cDNAs. Translation starts and splice sites were predicted by SpliceMachine[41]. Available *M. incognita* ESTs were aligned on the genome using GenomeThreader[42]. Similarities to *C. elegans* and other species' protein, genome and EST sequences were identified using BLAST[43]. Repetitive sequences were masked using RepeatMasker (http://repeatmasker.org/, **Supplementary Methods**, section 8.8; **Supplementary Fig. 15** online).

Automatic functional annotation. Protein domains were searched with InterproScan[44]. We also submitted proteins from seven additional species to the same InterproScan search. We included three other nematodes (*C. elegans*, *C. briggsae* and *B. malayi*), the fruitfly (*D. melanogaster*) and three fungi (*Magnaporthe grisea, Gibberella zea* and *Neurospora crassa*). To identify clusters of orthologous genes between *M. incognita* and the seven additional species, we used OrthoMCL[15] (**Supplementary Methods**, section 8.9).

Expert functional annotation. The collection of predicted protein coding genes was manually annotated by a consortium of laboratories. Each laboratory focused on a particular process or gene family relevant to the different aspects of *M. incognita* biology. Patterns of presence and/or absence and expansion and/or reduction in comparison to *C. elegans*, and other species were examined. The quality of predicted genes was manually checked and a functional annotation was proposed accordingly (**Supplementary Methods**, sections 8.10–8.20). A genome browser and additional information on the project are available from http://meloidogyne.toulouse.inra.fr/.

Accession codes. The 9,538 contigs resulting from the *Meloidogyne incognita* genome assembly and annotation were deposited in the EMBL/Genbank/DDBJ databases under accession numbers CABB01000001–CABB01009538.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

SCRI laboratory (V.C.B. and J.T.J.) received funding from the Scottish Government. This work benefited from links funded via COST Action 872. G.V.M. and V.L. are supported by ARC, CNRS, EMBO, MENRT and Region Rhone-Alpes. G.V.M., M.R.-R. and V.L. are also funded by the EU Cascade Network of Excellence and the integrated project Crescendo. M.-C.C. is supported by MENRT. We thank Philippe Lecomte for critical reading of the manuscript and all our collaborators from the "Plant-Nematode interaction" team of INRA Sophia Antipolis for technical help and support.

AUTHOR CONTRIBUTIONS

P.A. and J.G. contributed equally as first authors. J.-M.A., P.C.-S., E.G.J.D., E.D. and L.P.-B. contributed equally as second authors. T.J.B., M.B., T.B-Z., E.L.D., J.J.E., B.F., E.G., B.H., J.T.J., V.L., A.G.M., H.Q., M.-N.R., T.S., G.S., J.W. and P.W. contributed equally as senior authors. P.A., M.B., P.C.-S. and E.G.J.D. wrote the manuscript with input from J.T.J. and A.G.M. For biological material,

contributions were as follows. F.D., M.M. and L.P.-B. for strain growth, control and selection and DNA extraction. P.A., M.-C.C., F.D., E.D., B.F., M.-N.R. and L.P.-B. for cDNA libraries and EST data. For genome sequencing and assembly, contributions were as follows. B.S., E.U., J.P., V.A. for sequencing. C.J. for assembly. C.D. for cDNA clustering and library analyses. J.-M.A., O.J., C.J., F.A. for bioinformatics of allelism characterization. J.W. and P.W. supervision and coordination of the sequencing. For genome structure and organization, contributions were as follows. P.C.-S., T.F., H.Q. and D.S. for repetitive and transposable elements. J.G., E.S. for rRNAs, tRNAs, miRNAs. M.B. for operonic structures. M.-N.R., E.S. and C.V.G. for splice leaders (SL). For *in-silico* global genome analysis, contributions were as follows. E.D., J.G. and T.S. for gene predictions, automatic functional annotation, databases and bioinformatics. E.D. and B.F. for global protein set comparative analysis. Proteome expert annotation was as follows: P.M.C., E.G.J.D. and B.H., for Carbohydrate-Active enZymes. P.C.-S. and E.G. for proteases. M.-C.C., E.L.D., M.E., B.F., E.G.J.D., E.D., E.G., J.T.J., N.H., L.P.-B., G.S. and T.T. for candidate nematode parasitism and pioneer genes. P.A., T.B.-Z., E.G.J.D., E.D., J.J.E., J.V.G., G.P. and M.-N.R. for protection against plant defenses and immune system. V.L., G.V.M. and M.R.-R. for nuclear receptors. T.J.B., T.H. and T.R.M. for the kinome. E.G.J.D. and L.P.-B. for GPCRs. T.B.-Z., F.D.L., P.L. and P.V. for collagen. A.G.M. and P.M.V. for neuropeptides. J.T.J. for sex determination. V.C.B., E.G.J.D. and L.P.-B. for RNAi pathway and lethal RNAi phenotypes.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Trudgill, D.L. & Blok, V.C. Apomictic, polyphagous root-knot nematodes: exceptionally successful and damaging biotrophic root pathogens. *Annu. Rev. Phytopathol.* **39**, 53–77 (2001).
2. Blaxter, M.L. Nematoda: genes, genomes and the evolution of parasitism. *Adv. Parasitol.* **54**, 101–195 (2003).
3. Caillaud, M.C. *et al.* MAP65–3 Microtubule-associated protein is essential for nematode-induced giant cell ontogenesis in *Arabidopsis*. *Plant Cell* **20**, 423–437 (2008).
4. Caillaud, M.C. *et al.* Root-knot nematodes manipulate plant cell functions during a compatible interaction. *J. Plant Physiol.* **165**, 104–113 (2008).
5. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
6. Stein, L.D. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).
7. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
8. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
9. Leroy, S., Duperray, C. & Morand, S. Flow cytometry for parasite nematode genome size measurement. *Mol. Biochem. Parasitol.* **128**, 91–93 (2003).
10. Triantaphyllou, A.C. in *An Advance Treatise on Meloidogyne* vol. 1 (eds. Sasser, J.N. & Carter, C.C.) 113–126, (North Carolina State University Graphics, Raleigh, USA, 1985).
11. Castagnone-Sereno, P. Genetic variability and adaptive evolution in parthenogenetic root-knot nematodes. *Heredity* **96**, 282–289 (2006).
12. Mark Welch, D.B., Cummings, M.P., Hillis, D.M. & Meselson, M. Divergent gene copies in the asexual class Bdelloidea (Rotifera) separated before the bdelloid radiation or within bdelloid families. *Proc. Natl. Acad. Sci. USA* **101**, 1622–1625 (2004).
13. Piotte, C., Castagnone-Sereno, P., Bongiovanni, M., Dalmasso, A. & Abad, P. Cloning and characterization of two satellite DNAs in the low-C-value genome of the nematode *Meloidogyne* spp. *Gene* **138**, 175–180 (1994).
14. Foissac, S. & Schiex, T. Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* **6**, 25 (2005).
15. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
16. Davis, E.L., Hussey, R.S. & Baum, T.J. Getting to the roots of parasitism by nematodes. *Trends Parasitol.* **20**, 134–141 (2004).
17. Wei, Y.D. *et al.* Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, *Apriona germari*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **145**, 220–229 (2006).
18. Qin, L. *et al.* Plant degradation: a nematode expansin acting on plants. *Nature* **427**, 30 (2004).
19. Lambert, K.N., Allen, K.D. & Sussex, I.M. Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic nematode *Meloidogyne javanica*. *Mol. Plant Microbe Interact.* **12**, 328–336 (1999).
20. Hotson, A. & Mudgett, M.B. Cysteine proteases in phytopathogenic bacteria: identification of plant targets and activation of innate immunity. *Curr. Opin. Plant Biol.* **7**, 384–390 (2004).
21. Tang, X., Xiao, Y. & Zhou, J.M. Regulation of the type III secretion system in phytopathogenic bacteria. *Mol. Plant Microbe Interact.* **19**, 1159–1166 (2006).
22. Huang, G. *et al.* A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita*. *Mol. Plant Microbe Interact.* **16**, 376–381 (2003).
23. Lindblom, T.H. & Dodd, A.K. Xenobiotic detoxification in the nematode *Caenorhabditis elegans*. *J. Exp. Zoolog. A Comp. Exp. Biol.* **305**, 720–730 (2006).
24. Menzel, R., Bogaert, T. & Achazi, R. A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. *Arch. Biochem. Biophys.* **395**, 158–168 (2001).
25. Ewbank, J.J. Signaling in the immune response. *WormBook* doi/10.1895/wormbook.1.83.1, <http://www.wormbook.org/> (2006).
26. Alegado, R.A. & Tan, M.W. Resistance to antimicrobial peptides contributes to persistence of Salmonella typhimurium in the *C. elegans* intestine. *Cell Microbiol.* **10**, 1259–1273(2008).
27. Paschinger, K., Gutternigg, M., Rendic, D. & Wilson, I.B. The N-glycosylation pattern of *Caenorhabditis elegans*. *Carbohydr. Res.* **343**, 2041–2049 (2007).
28. Bertrand, S. *et al.* Evolutionary genomics of nuclear receptors: from 25 ancestral genes to derived endocrine systems. *Mol. Biol. Evol.* **21**, 1923–1937 (2004).
29. Robinson-Rechavi, M., Maina, C.V., Gissendanner, C.R., Laudet, V. & Sluder, A. Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J. Mol. Evol.* **60**, 577–586 (2005).
30. Plowman, G.D., Sudarsanam, S., Bingham, J., Whyte, D. & Hunter, T. The protein kinases of *Caenorhabditis elegans*: a model for signal transduction in multicellular organisms. *Proc. Natl. Acad. Sci. USA* **96**, 13603–13610 (1999).
31. Robertson, H.M. & Thomas, J.H. The putative chemoreceptor families of *C. elegans*. *WormBook* doi/10.1895/wormbook.1.66.1, <http://www.wormbook.org/> (2006).
32. Marks, N.J. & Maule, A.G. in *Neuropeptide Systems as Targets for Parasite and Pest Control* (eds. Geary, T.G. & Maule, A.G.) (Landes Bioscience/Eurekah.com, Austin, TX, USA, 2008).
33. Zarkower, D. Somatic sex determination. *WormBook* doi/10.1895/wormbook.1.84.1, <http://www.wormbook.org/> (2006).
34. Papadopoulou, J. & Triantaphyllou, A.C. Sex-determinantion in *Meloidogyne incognita* and anatomical evidence of sexual reversal. *J. Nematol.* **14**, 549–566 (1982).
35. Rosso, M.N., Dubrana, M.P., Cimbolini, N., Jaubert, S. & Abad, P. Application of RNA interference to root-knot nematode genes encoding esophageal gland proteins. *Mol. Plant Microbe Interact.* **18**, 615–620 (2005).
36. Huang, G., Allen, R., Davis, E.L., Baum, T.J. & Hussey, R.S. Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene. *Proc. Natl. Acad. Sci. USA* **103**, 14302–14306 (2006).
37. Zawadzki, J.L., Presidente, P.J., Meeusen, E.N. & De Veer, M.J. RNAi in Haemonchus contortus: a potential method for target validation. *Trends Parasitol.* **22**, 495–499 (2006).
38. Birky, C.W. Jr. Bdelloid rotifers revisited. *Proc. Natl. Acad. Sci. USA* **101**, 2651–2652 (2004).
39. Mark Welch, D. & Meselson, M. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**, 1211–1215 (2000).
40. Mark Welch, D.B., Mark Welch, J.L. & Meselson, M. Evidence for degenerate tetraploidy in bdelloid rotifers. *Proc. Natl. Acad. Sci. USA* **105**, 5145–5149 (2008).
41. Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332–1338 (2005).
42. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
43. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
44. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
45. Hillier, L.W. *et al.* Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res.* **15**, 1651–1660 (2005).

Molecular and Cellular Endocrinology 293 (2008) 5–16

Contents lists available at ScienceDirect

Molecular and Cellular Endocrinology

journal homepage: www.elsevier.com/locate/mce

Review

The evolution of the ligand/receptor couple: A long road from comparative endocrinology to comparative genomics

Gabriel V. Markov[a,b], Mathilde Paris[a], Stéphanie Bertrand[c], Vincent Laudet[a,*]

[a] *Molecular Zoology Team, Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland Lyon Sud, France*
[b] *USM 501/UMR CNRS 5166-Evolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, Paris, France*
[c] *Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, edifici annex, 1a planta, 08028 Barcelona, Spain*

ARTICLE INFO

Article history:
Received 11 February 2008
Received in revised form 14 May 2008
Accepted 11 June 2008

Keywords:
Nuclear receptors
Comparative endocrinology
Gene duplication
Phylogeny
Orthology

ABSTRACT

Comparative endocrinology considers the evolution of bioregulatory systems and the anatomical structures and molecules that constitute the neuroendocrine and endocrine systems. One aim of comparative endocrinology is to trace the origins of the main endocrine systems. The understanding of the evolution of the ligand/receptor couple is central to this objective. One classical approach to tackle this question is the characterization of receptors and ligands in various types of non-model organisms using as a starting point the knowledge accumulated on classical models such as mammals (mainly human and mouse) and arthropods (with *Drosophila* among other insects). In this review we discuss the potential caveats associated to this two-by-two comparison between a classical model and non-model organisms. We suggest that the use of an evolutionary approach involving comparisons of several organisms in a coherent framework permits reconstruction of the most probable scenarios. The use of the vast amount of genomic data now available, coupled to functional experiments, offers unprecedented possibilities to trace back the origins of the main ligand/receptor couples.

Contents

Comparative endocrinology, a very active branch of endocrinology, is mainly interested in the origin and diversification of hormonal systems in living organisms. Given the medically oriented knowledge that is a salient feature of modern endocrinology research, one basic focus of comparative endocrinologists is to trace back the origin of the human major endocrine systems and to understand the main events that have prompted the diversification

* Corresponding author at: Molecular Zoology Team, Institut de Génomique Fonctionnelle de Lyon, UMR 5242 du CNRS, INRA, IFR128 BioSciences Lyon-Gerland, Université de Lyon, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France. Tel.: +33 4 72 72 81 90; fax: +33 4 72 72 80 80.
E-mail address: Vincent.Laudet@ens-lyon.fr (V. Laudet).

0303-7207/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.mce.2008.06.011

of these systems (Norris, 2007). This is not an easy task, since this in fact tackles one of the major questions in comparative sciences, that is the origin of complex systems, here the cell–cell communication systems that are acting at the level of the whole organism. Comparative endocrinology, like any other field of life science with an evolutionary component relies on the comparison itself to infer the existence of a given process, anatomical structure, or molecule, back in time. Indeed, since it is usually impossible to have direct information on ancestor species (with the, still anecdotical, but promising exception of ancient DNA research, see Lalueza-Fox et al., 2007) it is the observation that a given feature is conserved in two distant species that allows to conclude that this feature was indeed present in the common ancestor of these two species. Thus a major aspect of research projects in comparative endocrinology is the characterization of endocrine systems in non-model organisms, using model organisms (mainly human but also mouse or arthropods such as *Drosophila*) as a starting point.

The purpose of this short review is double. First, we will emphasize that despite its apparent conceptual simplicity the comparative approach in endocrinology is paved with methodological difficulties. This clearly suggests that if artefactual inferences are to be avoided the traditional comparative approach should be replaced by multi-disciplinary evolutionary and functional studies. Second, we will detail how the impressive amount of data generated by recent genomic analysis offers unprecedented possibilities to carry out this type of research, thus placing comparative endocrinology in front of a major shift in its methods and approaches. Of course such a short review can only provide rapid glances on this burgeoning field. We will thus illustrate this paper by several examples taken from the recent literature without being exhaustive.

1. The ligand/receptor couple

The evolution of the ligand/receptor couple is a question that attracts considerable debate and theoretical discussion of complex experimental approaches. In fact hormones and receptors are central in the understanding of endocrine systems and, their origin as well as their parallel variation through co-evolution is a major evolutionary question. Indeed, divergence of proteins in different species requires ligand and receptor(s) coevolution to improve binding affinity and/or specificity. Coevolution is thus a ubiquitous process that is responsible for the parallel adaptative evolution of hormone/receptors couples in the broadest sense.

On this aspect of coevolution of ligand/receptor pairs the field is sharply cut into two parts given the chemical nature of the ligand. All the ligands that are peptides or proteins, i.e. that are encoded by genes provide conceptually relatively simple cases of ligand/receptor coevolution, with continuous adaptation across time. Protein–protein interaction in general is, from the coevolutionary point of view, not basically different from the interaction between a given receptor and its ligand (Waddell et al., 2007). In such cases, it is believed, and it has been demonstrated in several specific cases, that the genes encoding the ligand and the receptors are undergoing parallel evolution. One of the first examples of such coevolution is the one of the receptors for LH and FSH, which suggests that indeed the specificity of each ligand/receptor pair is maintained in divergent species (Moyle et al., 1994). More recently the case of prolactin receptors in mammals showed how episodes of adaptative evolution have modified the genes encoding the receptor and the ligand (Li et al., 2005, and references therein). In most mammals the prolactin gene evolved very slowly but this near-stasis was interrupted by bursts of rapid changes during the evolution of several mammals orders such as artiodactyls, primates or rodents. Since prolactin has to bind its receptor to fulfill its function, it was anticipated that the gene encoding the prolactin receptor should be subjected to selective pressure in the same mammals. This has been shown to be the case and the correlation between the evolutionary rates of the ligand and the receptor is effectively indicative of such coevolution. Similar examples including G protein-coupled receptors (GPCRs) such as the receptors for PRXamides (Park and Palczewski, 2005) or the secretin (PACAP and VIP) and their receptors (Cardoso et al., 2007) support this theory of ligand–receptor coevolution. Thus, conceptually, the existence of evolutionary couples is relatively well understood and provides a coherent framework for functional evolution studies (Dean and Thornton, 2007).

This situation contrasts with the second, that of receptors for which the ligand is not a peptide or a protein but rather a small molecule. In such cases the ligand is not a gene product but is derived from a biochemical pathway that starts from an inactive precursor, sometimes derived from an external source such as food, which is transformed into the active molecule (see Simões-Costa et al., 2008 for a recent illustration on retinoic acid metabolism). This is the case for some GPCRs but also for many nuclear receptors (NRs), for which the ligands are the products of complex biochemical pathways. In most cases these pathways contain a rate-limiting step, producing the active compound. This critical step is most often the one that is physiologically regulated. In addition it contains a catabolic part that is responsible for the degradation of the ligand and that is also subjected to precise regulation (see You, 2004; Bélanger et al., 1998, for a review on steroids). In these cases of ligand/receptor pairs a simple coevolution mechanism obviously cannot operate. In the case of NRs, several models such as ligand exploitation (Thornton, 2001) or refinement of ligand-binding specificity by mutations (Escriva et al., 2006) have recently been proposed to explain how changes of specificity can take place during evolution. Nevertheless even if the situation for these ligand/receptor couples is more complex, the existence of coevolution is still possible. Indeed, a recent report on cannabinoid receptors suggests that the evolution of cannabinoid receptors is correlated with the evolution of diacylglycerol lipase, an enzyme implicated in the metabolism of anandamide and 2-arachidonyl glycerol (2-AG) the two endogenous ligands of cannabinoid receptors (McPartland et al., 2007). It remains to be investigated if such example can be generalized to other receptor systems.

It is clear nevertheless that any discussion on the evolution of a ligand/receptor couple should be confronted with experimental data, which consists in the characterization of ligands and receptors in various species, distantly related to the most classical models. It has also to be emphasized that receptors and ligands structures are often more conserved than their physiological function which depends on the expression patterns of receptors. This often blurs the recognition of orthology between related ligand/receptor pairs.

2. The phylogenetic framework: metazoan evolution

Any analysis of the evolution of a given ligand/receptor couple should take into account the phylogeny of the organisms from which these various pairs are coming from. Therefore, it is important at that level to rapidly present here the framework on which these comparative approaches are developed. Since most of these studies are done at the scale of animals we will limit this rapid presentation to the case of metazoans.

Fig. 1 depicts a simplified version of the currently endorsed evolutionary tree for metazoans. It is striking that even if the distribution of model organisms is very much biased towards two of the main clades of metazoans (namely ecdysozoans with *Drosophila* and *Caenorhabditis elegans* and deuterostomes with human, mouse,

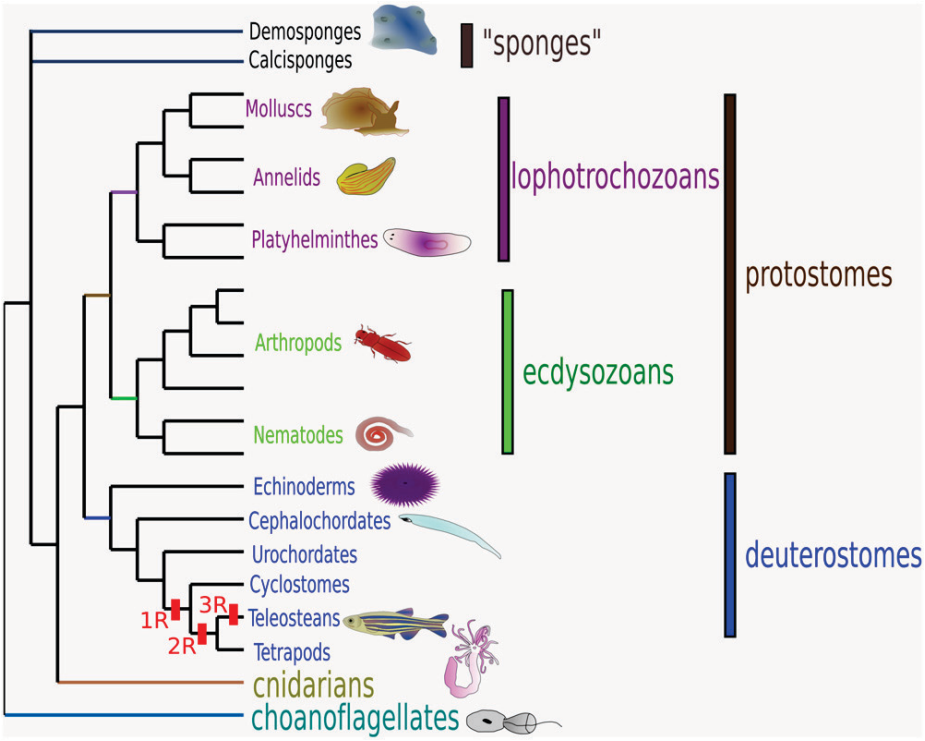


Fig. 1. An updated view on metazoan phylogeny. 1R, 2R, and 3R indicate the whole genome duplication events that occurred in the vertebrate lineage.

chicken, xenopus and zebrafish, among others) genomic data are now available for all the main metazoan clades. For example, looking at cnidarians and sponges, at the base of the metazoan tree, the genome of the sea anemone *Nematostella vectensis* (Putnam et al., 2007) is now available whereas the one of the sponge – *Amphimedon queenslandica* – is available as trace archives. These genomes are bringing crucial information that now allow one to make inferences on early metazoan gene families (Simionato et al., 2007).

From the phylogenetic tree depicted in Fig. 1, the bilaterians, that are the metazoans with three embryonic layers (ectoderm, endoderm and mesoderm) and with a clear antero-posterior axis, are divided into three main clades: on one hand, the lophotrochozoans and the ecdysozoans that together form the protostomes and, on the other hand, the deuterostomes. The monophyly of these three clades is relatively well accepted as is their branching order even if some analyses based on large genomic data sets tend to propose alternative schemes (for example the Coelomata hypothesis that groups arthropods and deuterostomes, excluding nematodes; see for example Rogozin et al., 2007).

The topology of the metaozan tree may have a major impact on our inferences regarding the origin and diversification of endocrine genes. This influence is well illustrated by a recent analysis of NR genes distribution and phylogeny in metazoans in which are compared the scenarios implied by the Ecdysozoan or Coelomata hypothesis on the evolutionary history of nuclear receptors (Bertrand et al., 2004). The use of alternative topologies of the metazoan tree (as the Coelomata hypothesis used for the analysis of the Forkhead family; Carlsson and Mahlapuu, 2002) may affect the conclusions drawn relative to the ancestry and evolution of specific genes. At the present stage of our phylogenetic knowledge, this should be regarded with caution.

Another important observation, developed below, is that the number of genes present in the genomes of several classical model organisms tends to be extremely variable because of large scale events such as whole genome duplication, lineage specific expansion of specific genes or, alternatively gene loss (see Panopoulou and Poustka, 2005 for a review on vertebrates). It is now widely recognized that extremely important models such as *Drosophila*, *C. elegans*, or the urochordate *Ciona intestinalis* have experienced extensive gene loss (see Bertrand et al., 2004 for references). This may have important implications in terms of endocrine gene evolution. For example, the estrogen receptor has been found in vertebrates but not in invertebrate chordates such as *Ciona*, *Drosophila* or *C. elegans*. This has led to the proposal that this receptor was a key innovation of vertebrates (Laudet, 1997; Escriva et al., 2000). In fact the observation that an estrogen receptor orthologous gene is present in several mollusks (Thornton et al., 2003; Keay et al., 2006) as well as in cephalochordates (Paris et al., 2008a) shows that it is in fact much more ancient than expected and that the gene was lost independently in ecdysozoans and urochordates (Bertrand et al., 2004; Escriva et al., 2004). A very similar situation was found for the thyroid hormone receptor (Bertrand et al., 2004; Wu et al., 2006, 2007). It is now widely accepted that no conclusion can be reached on the presence or absence of a given gene in the common ancestor of all bilaterians if data from the three main lineages of metazoans (that is lophotrochozoans, ecdysozoans and deuterostomes) are not available. Given that deuterostomes plus insects and nematodes contain the most dominant model organisms this conclusion should be strongly re-emphasized: no safe conclusion can be drawn on the ancestry of a given gene family without data from lophotrochozoans and/or cnidarians.

3. Two by two comparisons of receptors, proceed at your own risk!

A classical approach in comparative endocrinology is to identify in various organisms, for example the cephalochordate amphioxus, an orthologue of a given receptor known in a classical model organ-

isms like human. This type of analysis may be important to clearly show if this receptor, and by extension, the corresponding signaling pathway is effectively present in the organism of interest. We will highlight below that, even if of course the conclusions reached from this type of analysis are interesting, this two-by-two comparative approach has several caveats that renders it quite risky. Thus, the conclusions and evolutionary models reached by these traditional comparative approaches, based on the assumption that one can extend to their zoological groups the knowledge accumulated in human or *Drosophila*, should be put into a larger perspective. It is only through their independent confirmation based on large-scale evolutionary analysis that these conclusions will be firmly assessed.

3.1. Gene duplication

An important result generated over the last 10 years of genomic analysis is that genes, and even genomes, duplications once believed to be relatively rare events are in fact a major evolutionary mechanism that has been instrumental in shaping the current biodiversity (Ohno, 1970, and see also Volff, 2005 for a recent review in teleosts). Gene duplication is an important mechanism of gene diversification that can be observed in nearly all organisms. Of course tandem duplication of individual genes, the simplest case of gene duplication, occurs quite often. But two more global processes that have broad implications for the functional anatomy of genomes should be emphasized. The first is the whole genome duplication, the importance of which was first highlighted in the 1970s by Susumu Ohno (Ohno, 1970) and later revealed by the study of invertebrate chordates and early vertebrates (Garcia-Fernández and Holland, 1994) as well as fishes (Wittbrodt et al., 1998). It is now well established that in several cases, such as at the base of the vertebrate tree (Dehal and Boore, 2005) or early on during actinopterygian fish evolution (Jaillon et al., 2004) whole genome duplication took place and impacted strongly on the appearance and diversification of complex features, including the endocrine systems (Holland et al., 2008). The fate of duplicated genes (non-, sub- or neo-functionalization, see Force et al., 1999 that is, respectively loss of one copy, sharing of the ancestral function between the two duplicates or acquisition of a new function by one of the copy), and its link with the origin of evolutionary novelties, is currently a major and fruitful research area and the study of endocrine genes should be placed in this context. A very convincing example of such an analysis in the context of actinopterygian fish whole genome duplication has been recently published for the MyoD family (Macqueen and Johnston, 2008).

The other extreme case of gene duplication is lineage-specific duplication events that can sometimes be extensive. This is the mechanism at the origin of the 270 nuclear receptor genes that are present in the genome of the nematode *C. elegans* (Sluder and Maina, 2001). It has been shown that most of these genes correspond in fact to orthologues of a unique gene encoding the orphan receptor HNF-4, which was massively duplicated in *C. elegans* and related species (Robinson-Rechavi et al., 2005). The functional significance of this burst of duplication is still under investigation, but it is clear that such an event should have consequences on the physiology of the animal and may completely modify the evolutionary scenarios constructed at the level of an individual gene family. GPCRs (Cardoso et al., 2006), but also receptors with tyrosine kinase activities (Rikke et al., 2000), the heterotrimeric G protein a-subunit (O'Halloran et al., 2006) and the *Hedgehog*-related genes (*Hog*) (Aspöck et al., 1999) also provide cases of lineage-specific expansion in nematodes.

The presence of these duplication events has a major consequence for the comparative endocrinologist, as it reinforces the

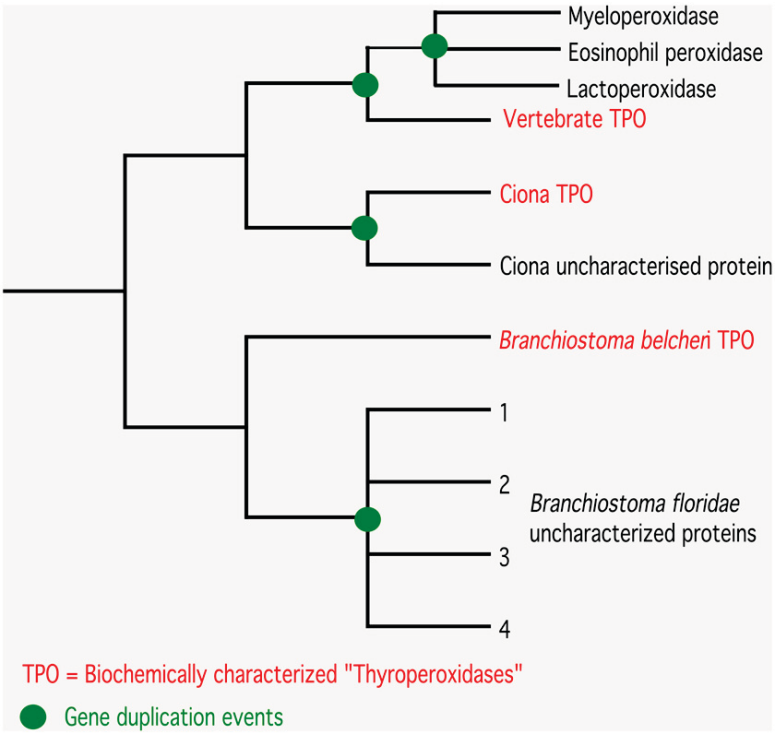


Fig. 2. Phylogenetic relationships in the chordate thyroperoxidase family. Duplication events are indicated by green spots. Note that myeloperoxidase, lactoperoxidase and eosinophil peroxidases are the mammalian members of a gene family where duplications occurred at various levels in the vertebrate lineage. General tree topology is based on Heyland, 2006, and completed with data from *Ciona* complete genome (http://www.treefam.org/cgi-bin/TFinfo.pl?ac=TF314416) and from Holland et al., 2008, for *Branchiostoma floridae* sequences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

need for an accurate phylogenetic study of any protein of interest in order to avoid simplified, and often erroneous interpretations. Fig. 2 shows the example of the thyroperoxidase (TPO) gene, which is implicated in the synthesis of vertebrate thyroid hormones. The vertebrate TPO is a member of a multigenic family, being the orthologue of a large group of peroxidases that experienced a complex series of duplication events within the vertebrate lineages. In mammals we find not only thyroid peroxidase but also lactoperoxidase, myelo-peroxidase and eosinophil peroxidase, all these genes coming from a unique ancestor gene that underwent duplications at various levels within the vertebrate lineages (Heyland et al., 2006). Using primers corresponding to a various set of metazoan peroxidases, a gene encoding a peroxidase was cloned in the urochordates *C. intestinalis* and *Halocynthia roretzi*. Since in vertebrates four closely related types of peroxidases are found, one cannot conclude that this *Ciona* gene is indeed a TPO (see our recent review on the importance of gene nomenclature, Markov et al., 2008). It is only with additional data, here through the analysis of the expression pattern and the comparison to the previously reported biochemical activity that the conclusion that the gene encodes a TPO can be reached. Interestingly, a more extensive genomic analysis (Fig. 2) shows that this protein is the product of only one of the two duplicated genes in the *Ciona* genome. The original "*Ciona* TPO" was shown to be expressed in a domain that does not overlap that of its classical regulator TTF1 in vertebrates, and that its expression domain was restricted to the endostyle zone 7, whereas previous histochemical studies reported a peroxidase activity also in other parts of the endostyle (zones 8 and 9) (Ogasawara et al., 1999). The presence of the second gene suggests that a more complete study is needed in order to test if it too has also a TPO-activity, if it is expressed in zones 8 and 9 of the endostyle and if the two genes have overlapping or complementary expression patterns and thus to reconstruct the detailed history of these genes. Since the duplication event that gave rise to the two *Ciona* genes is independent of the duplications that occurred in vertebrates, only functional analysis can determine whether the *Ciona* genes have activities related to thyroid, myelo-, lacto- or eosinophile peroxidases. The situation is even more complex in cephalochordates. Another gene encoding a TPO was characterised in the Chinese amphioxus *Branchiostoma belcheri* (Ogasawara, 2000) but the analysis of the *Branchiostoma floridae* genomes (Holland et al., 2008) shows that there are four orthologues of this gene, once again arising from an independent series of duplications from the ones that occurred in *Ciona* or vertebrates. Thus, starting from a unique gene with an unknown activity, the peroxidase gene family has been independently elaborated three times in vertebrates, urochordates and cephalochordates. Since proteins that undergo duplication are prone to subfunctionalisation or neofunctionalisation events (Force et al., 1999), this should be taken into account when comparing the functions of two proteins in two different organisms, because this can have major effects at the physiological level. This example illustrates that one should take into account the full set of genes and their complex history in order to infer their ancestral functions.

3.2. Gene loss

Gene loss is an often neglected aspect that, due to the current interest in whole genome sequences, has recently been shown to be a frequent and important evolutionary mechanism that contributed significantly to the emergence of divergent animal lineages (Danchin et al., 2006). The analysis of the presence of NRs genes in complete genome sequences of metazoans shows that the NR complement of different animal models is extremely variable and that gene loss was effectively frequent, as discussed above in the case of *Drosophila*, nematodes and urochordates (Bertrand et al., 2004). The case of nematodes is particularly puzzling since the events of gene loss are hidden by the massive lineage specific expansion of the HNF4 gene discussed above (Robinson-Rechavi et al., 2005) thus illustrating the complexity of the individual gene family history (Fig. 3).

But gene loss is not only revealed by comparison of distant evolutionary organisms. One example, taken from the evolution of NRs illustrates the importance of taking gene loss into account. When we compared, through careful phylogenetical analysis, the NR complement present in mammals and teleost fishes we were surprised to find cases that could not be explained by the classical "more genes in fish" scenario (Crollius and Weissenbach, 2005; Bertrand et al., 2004). This is well exemplified by the case of the Rev-erb genes, orphan nuclear receptors. In all known mammals, two paralogous Rev-erb genes, called Rev-erbα and Rev-erbβ, corresponding to a unique orthologue in *Drosophila* (called E75) and amphioxus are known (Laudet, 1997). Given the actinopterygian fish whole genome duplication event discussed above, we were expecting to find a maximum of four Rev-erb genes in zebrafish, namely two Rev-erbα and two Rev-erbβ. During our systematic analysis of NR genes expression patterns in zebrafish we were thus surprised to find in fact five Rev-erb genes in zebrafish whereas four are effectively present in pufferfishes (tetraodon and fugu) (Bertrand et al., 2007).

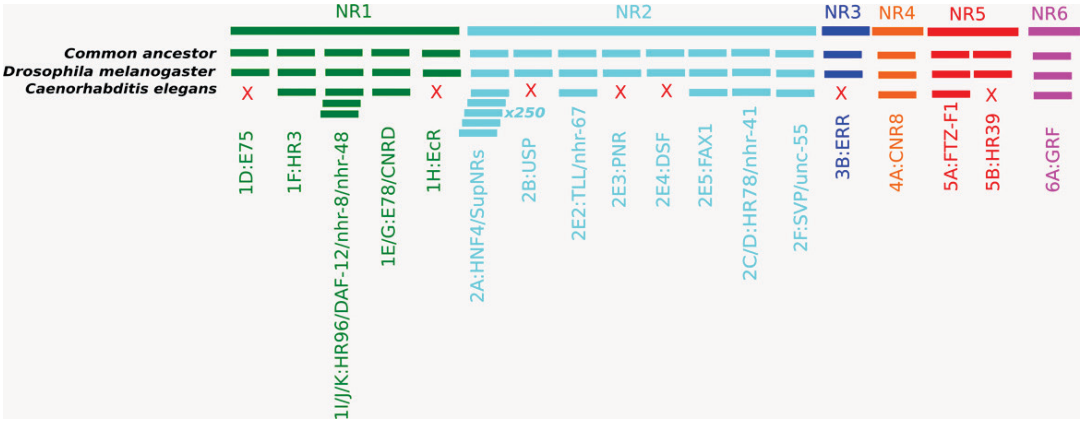


Fig. 3. Gene duplications and losses in *Caenorhabditis elegans*: the example of the nuclear receptor superfamily. The figure shows the loss of seven genes and duplications of two of them in *C. elegans*, in comparison with the *Drosophila* gene set. Data are from Bertrand et al. (2004). Note that the gene set of the "common ancestor" presented here is minimal, and coherent with the comparison of only these two species. A more complete analysis would show that some other genes were also lost independently in *Drosophila* and *C. elegans*.

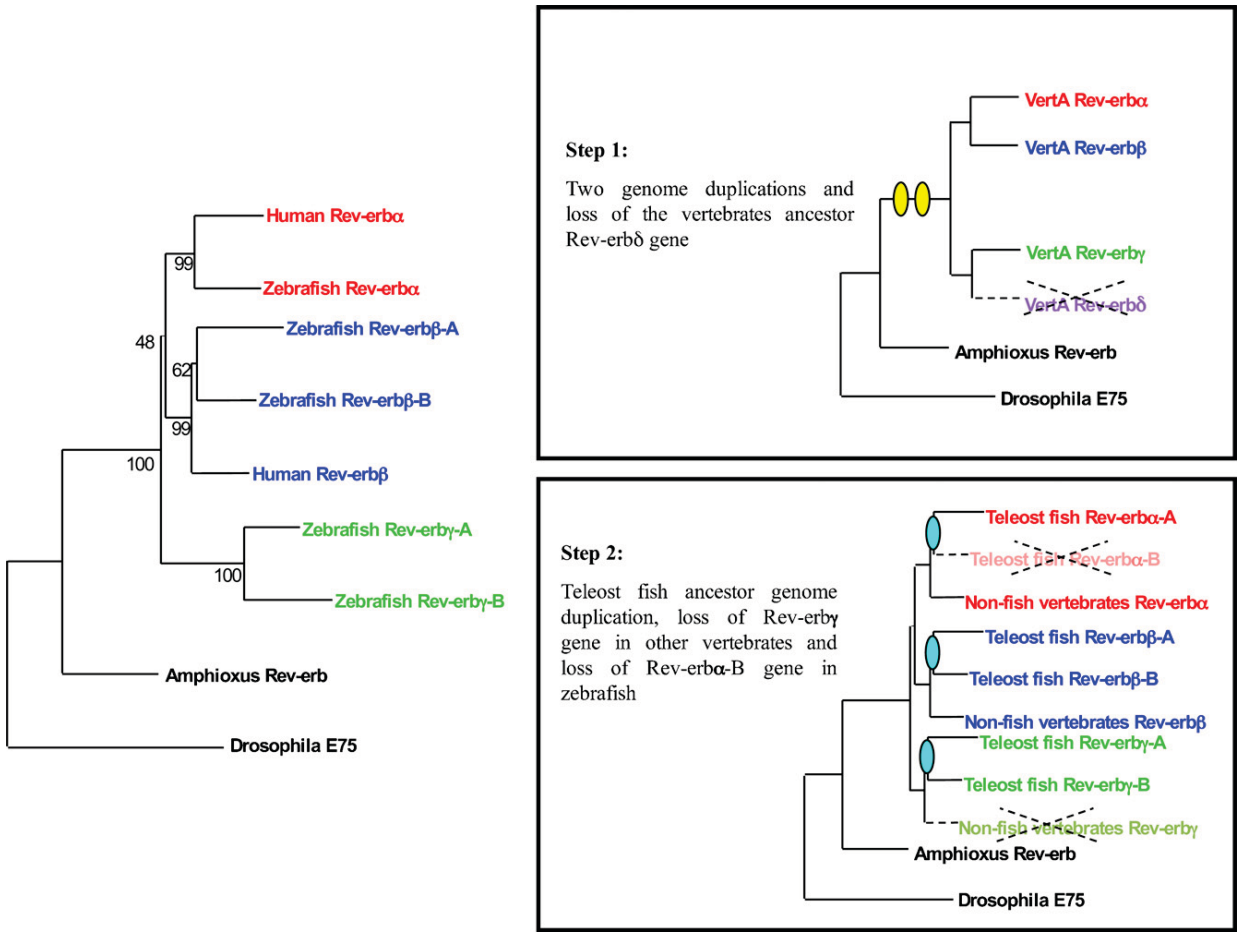


Fig. 4. Phylogenetical tree of the Rev-erb subfamily in zebrafish and human and the most probable evolutionary scenario. Branch-lengths are arbitrary. The two whole genome duplications that took place at the origin of vertebrates and the whole genome duplication at the origin of teleost fish are schematised by yellow and light-blue spots, respectively. VertA means vertebrate ancestor. The most parsimonious scenario explaining the actual evolutionary relationships between human and zebrafish Rev-erb genes can be separated in two steps. The first one corresponds to the loss of one Rev-erb paralogue (here named Rev-erbδ) in the ancestor of vertebrates after the two whole genome duplications. The second step is the loss of Rev-erbγ paralogue in the ancestor of the vertebrates that diverged after fish divergence, and the loss of one Rev-erbα duplicate (here named Rev-erbα-B) in zebrafish after the whole genome duplication that took place at the base of teleost fish lineage. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

By phylogenetical analysis (Fig. 4) we were able to deduce that in fact three paralogous Rev-erb genes (Rev-erbα, β and a third gene called γ) were present in the ancestor of vertebrates. In tetrapods this third gene, Rev-erbγ was lost. In fishes, the three ancestral Rev-erb genes were duplicated giving rise to six genes. The differential loss of these fish-specific paralogues explains the different numbers of Rev-erb genes in various fish species: zebrafish lost one Rev-erbα copy whereas pufferfish lost the two Rev-erbα genes. The analysis of the expression patterns strongly suggests that if these genes share a similar function each of them has specific implications in distinct processes, probably linked to circadian rhythm regulation (Bertrand et al., 2007; Kakizawa et al., 2007). This case is certainly not an isolated one and taking only the example of NR genes several other cases of specific gene loss in mammals were found (e.g. ERRs, COUP-TFs, SF-1) whereas, in contrast, one case of fish specific loss (CAR) was found (Bertrand et al., 2007). It is now known that these gene losses occurred in other gene families in mammals but also in other zoological groups (Danchin et al., 2006; Wyder et al., 2007).

All these examples illustrate that the evolutionary history of any given gene cannot be reconstructed without a careful phylogenetic analysis based on the use of adequate methods and that the importance of mechanisms such as gene duplication and gene loss should be adequately tested before concluding on the presence or absence of a given signaling pathway. Classical comparative work based on the two-by-two comparison of human and another animal will never match the quality of information reached by an in depth phylogenetical analysis.

3.3. Evolutionary shifts

Of course genes are not only duplicated or lost. Many genes are conserved as unambiguous orthologues in a wide variety of organisms, but even in such cases one should be careful when studying gene evolution since evolutionary shifts can have dramatic impacts in terms of gene function, as well as at the level of phylogeny. Long branch attraction is a well-known artefact in molecular phylogeny that can artificially group rapidly evolving clades together in a position clearly not compatible with the known phylogeny of species (Delsuc et al., 2005).

RXR evolution in insects provides an illustration of this phenomenon (Fig. 5). The known orthologue of RXR in *Drosophila* is the USP gene that plays an important role as a common heterodimeric partner for several NRs, including the ecdysone receptor, EcR. When we did phylogenetic studies of USP in several insects we saw a

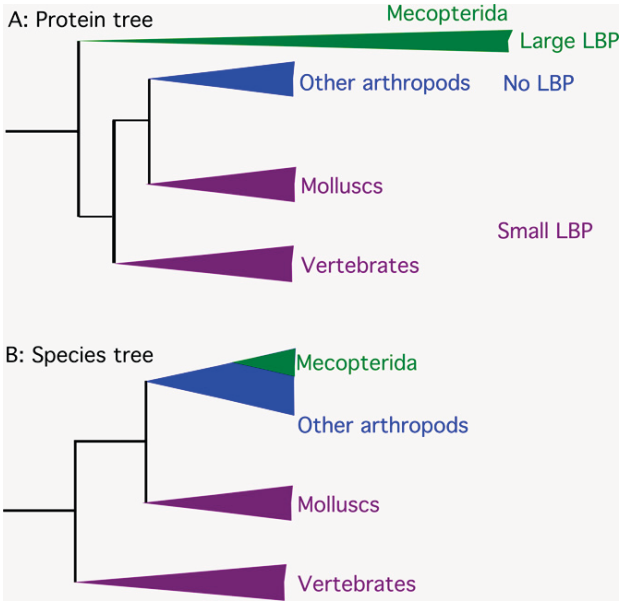


Fig. 5. Long branches attraction and functional shifts: the example of USP-RXR. Adapted from Bonneton et al. (2003) and Iwema et al. (2007). (A) USP-RXR protein tree, with an abnormal position of Mecopterida at the base of the bilaterians. Further analysis showed that this long branch, corresponding to an acceleration of the evolutionary rate in Mecopterida USP, correlates with changes in the ligand-binding abilities. (B) Species tree based on classical neutral markers such as ribosomal RNA genes), showing the real position of Mecopterida within arthropods. The three different colours refer to the different binding abilities of the ligand-binding pocket (LBP). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

very striking tree topology since all the USPs found in Diptera and Lepidoptera were not clustered with other arthropod sequences as expected, but at the base of all bilaterian RXRs (Bonneton et al., 2003). At the level of sequence identity this example is quite spectacular since, in the ligand-binding domain the RXR from a beetle (e.g. *Tribolium*) is more closely related to the human RXR than to the *Drosophila* USP. We found that in fact this corresponds to a very strong acceleration of evolutionary rate that occurred during the evolution of mecopterida, a monophyletic group with approximately 25% of extant insect species, including Trichoptera (caddisflies), Mecoptera (scorpionflies) and Siphonaptera (fleas) in addition to Diptera and Lepidoptera (Bonneton et al., 2006). Careful structural and functional analysis of RXR from *Tribolium* and *Drosophila* allowed us to reconstruct a complex and dynamic evolutionary history shaped by several functional shifts during which several events of loss and gain of ligand-binding occurred (Iwema et al., 2007). Interestingly, this event is not restricted to RXR since several similar shifts were also recently observed in other nuclear receptors, all acting in the ecdysone cascade (Bonneton et al., 2008). This example illustrates that such evolutionary shifts are not just complicating factors for phylogenetic analysis, they also provide very fruitful cases in which the evolution of ligand/receptor couples can be scrutinized. In this case, this reveals an unsuspected evolutionary flexibility of NRs in terms of ligand binding.

The literature now contains several cases of evolutionary shifts that all provide illuminating examples of the power of natural selection acting at the level of endocrine pathways. Among these, the nuclear receptors provide two recent examples of receptors, ERs and TRs, that were believed to be recent but for which orthologues were recently found in non-model protostomes. Interestingly, in both cases these protostome receptors do not bind the *bonafide* vertebrate ligand suggesting a more complex history than anticipated. Indeed the ligand binding evolution of estrogen receptor is still far from understood (see contrasting views in Baker, 2003; Thornton et al., 2003; Paris et al., 2008b) and similar questions are suggested by the recent characterization of thyroid hormone receptors in platyhelminthes, molluscs or even in a deuterostome, the cephalochordate amphioxus (Wu et al., 2007; Paris et al., 2008a). At a much smaller taxonomic level the careful analysis of the ligand-binding specificity of PXR in mammals also provides an example of shifts in the ligand-binding ability of a nuclear receptor (Krasowski et al., 2005; Reschly et al., 2007). For other receptor systems, the interested reader could, among a long potential list, refer to the cases of the Growth Hormone receptor in primates (Liu et al., 2001; Wallis, 2001), the TSH receptor which exhibits striking functional differences between mammals and actinopterygian fishes, but also among mammals (Farid and Szkudlinski, 2004) or the genes of the ectodysplasin pathway in vertebrates (Pantalacci et al., 2008).

3.4. Recombination events

Genomic data now available greatly facilitates assessment of gene orthology. In some multigenic families, the duplication rate is so high and variable for different genes that even a cautious phylogenetic analysis is not sufficient to establish orthology relationships. Such a problem occurs when examining the vertebrate CYP2 family that contains genes encoding xenobiotic metabolising enzymes, which have undergone many duplications events at various taxonomic levels (e.g. mammal-specific, rodent-specific, and mouse-specific duplications, with primate-specific or/and rat-specific duplications in parallel) (Nelson, 2005; Thomas, 2007). In such cases, data on the chromosomic localisation can provide very useful information since the analysis of neighboring genes allows one to assess the orthology of the whole syntenic region and, by extension, of the genes of interest. This was indeed the case of the CYP2 genes since some of the genes are located in genomic clusters, where it is believed that recombination events occur quite often, giving rise to the wide diversity of CYP2 genes (Thomas, 2007). Such events will erase the phylogenetic signal and make the functional comparisons between proteins that underwent different duplication events meaningless.

Another example of careful analyses of syntenic regions that led to large scale conclusions on complex evolutionary scenario are the evolutionary analyses of the major histocompatibility cluster (MHC) by Pontarotti's group (Danchin et al., 2003). In the context of fish genome duplication the example of MyoD gene family evolution also illustrates how the use of synteny could help to infer orthology (Macqueen and Johnston, 2008). This type of analysis is particularly useful when studying genes encoding short peptide hormones, which are often difficult to unambiguously identify at a large evolutionary scale. The recent analyses of the POMC/AGPR/MCR gene repertoire in fugu (Klovins et al., 2004) and the study of the origin and coevolution between NPY receptor and prolactin-releasing hormone-receptors in vertebrates (Lagerström et al., 2005), also show the utility of complementary use of data from phylogenetic analysis coupled with careful synteny analysis. The construction of the "gene rosace" diagram that visualizes the respective relationships between genes located in different regions facilitates this type of analysis (see Jaillon et al., 2004, for an explanatory illustration). As for gene analyses, such studies take profit of in depth phylogenomic knowledge about the analysed genes, and also are greatly improved by comparisons between more than two genomes (for detailed examples see Kasahara et al., 2007; Muffato and Crollius, 2008).

When possible, phylogenetic approach should integrate in a known phylogenetic species and gene trees all the information possibly available at the leave of the trees. Information are of course the

sequences but also functional information such as binding properties and other physiological data. In some cases it is possible to calculate the probability to have this information at the node of the tree and from the information available at the node propagate dawn it to the non-annotated leaves (see Thornton, 2004, for a detailed review). But even the finest probabilistic approaches are useless if the sampling is inappropriate.

4. The elusive ligands

The previous examples may give the impression that working on the evolution of receptor genes is full of obstacles that complicate analysis and interpretation. Thus, one solution to study evolution of a specific endocrine signaling pathway could be to test directly if the ligand of interest is present in a wide variety of organisms. This attractive solution is nevertheless also paved with obstacles that are not easily solved and for which a comparative genomics approach does not always provide clues. The two main lines of evidence for the existence of a given ligand in a given animal are (i) the detection of an effect of this ligand in a given physiological or developmental process, most often considered from the known effect of this ligand in mammals and (ii) the direct detection of the ligand through genomic or analytical chemistry analysis.

4.1. Effects

The rationale behind these studies, which are commonly done in traditional comparative endocrinology is that if a given molecule has an effect in a given species this should be taken as a clear indication that there is a receptor for this molecule and thus a signaling pathway and a "physiological role". Somehow mirroring the original receptor definition by Paul Ehrlich, for which a receptor was defined as any molecule that is able to bind exogenous elements in human cells (reviewed in Prüll, 2003), people tend to consider that any molecule of human origin acting on a living system is effectively proof that this molecule binds to another one, which can be called a "receptor". But even if this assumption is true, it cannot be used as an indication that this "receptor" is related in anyway to the human receptor for this molecule. The problem is that many authors have applied this concept without taking into account this important caveat.

A striking example is the case of the arboreal mycorrhizes fungi of the order Glomerales, such as *Glomus intraradices*, that live in symbiosis with many trees and play crucial roles at the root/soil interface. It has been shown that the roots of the plant have a positive effect in stimulating the growth of the fungal hyphe during pre-symbiosis steps (see Requena et al., 2007, for a review). The precise mechanism how plant signals are perceived by the fungi is still unknown but it has been shown that flavonoids have a strong chimio-tactic effect and that this effect can be mimicked by estrogens and blocked by antiestrogens (Poulin et al., 1997; Scervino et al., 2005). This was taken as an indication that the genome of these fungi contained an estrogen receptor and that a specific receptor/interacting protein with a binding site for flavonoid or structurally related compounds (estrogens and antiestrogens) exists in these fungi (Requena et al., 2007; Poulin et al., 1997; Catford et al., 2006). However, our current knowledge of NR evolution indicates that the genome of *Glomus* (which is currently being sequenced) should be devoid of NR genes (Escriva et al., 1997). Among the possible candidates for mediating the effect of estrogens in *Glomus*, one of the most probable would be an orthologue of the estrogen binding protein already found in other fungi, including *Candida albicans* (Madani et al., 1994; Cheng et al., 2006). This protein, well known under the name of "Old yellow enzyme" is one of the most ancient enzymatic systems characterized, and is a cytoplasmic oxydoreductase whose activity can be regulated by estrogen binding (Cheng et al., 2006 and references therein). This example demonstrates that whereas there is an effect of estrogen in this species, this could not be taken as an indication that the endogenous receptor is in any way evolutionary related to the nuclear estrogen receptor found in human.

RXRs provide here again an example of the difficulties associated with the identification of endogenous ligands. These receptors were first described in the early 1990s as orphan receptors whose activity could be modulated by all-*trans*-retinoic acid (see Laudet and Gronemeyer, 2005, for references). It was later shown that 9-*cis*-retinoic acid, an isomeric derivative of all-*trans*-retinoic acid, is able to bind with a high affinity to RXR. Indeed several rexinoids, specific ligands of RXRs, were developed as pharmacological compounds for treatment of diseases such as diabetes and insulin resistance (Pinaire and Reifel-Miller, 2007). Ironically the detection of 9-*cis*-retinoic acid *in vivo* has proven to be difficult, casting doubts on the *in vivo* relevance of this ligand (see Mic et al., 2003). This has been recently confirmed by genetic evidence in the mouse (Calléja et al., 2006). It is now thought that RXR acts rather as a fatty acid or fatty acid derivative sensor, although the exact identity and relevance of its ligand(s) remains a matter of debate (see references in Iwema et al., 2007). This example depicts that even in human, it is not because a molecule has an effect that it indicates its endogenous existence in a physiologically relevant manner!

These examples may appear artificial, but they illustrate a type of assumption that has been commonly used in some comparative endocrinology studies. The estrogen receptor of several molluscs such as *Aplysia*, *Octopus*, and others have been cloned and convincingly shown to be unable to bind to estrogens (Thornton et al., 2003; Keay et al., 2006; Kajiwara et al., 2006; Matsumoto et al., 2007; Bannister et al., 2007). Nevertheless, effects of estrogens are reported in these animals (e.g. sex reversal in several mollusc species) and this is taken as evidence in favor of the existence of a classical nuclear estrogen receptor (see Lafont and Mathieu, 2007 for references). However, the actual data suggests that if such a receptor exists it is not the unique ER orthologue already characterized. Among the several possible explanations we propose that: (i) the molecule used for the treatment is metabolized to another compound that is the active one. Indeed we recently discovered such a situation in amphioxus, where thyroid hormones are metabolized to a compound that binds the TR (Paris et al., 2008a); (ii) a NR that binds estrogen exists but this is not an orthologue of ER; (iii) the effect of estrogen is mediated by another receptor system, e.g. a transmembrane receptor (see Revankar et al., 2005 for an example) or a cytoplasmic protein such as in the case of *Glomus* discussed above. Thus one should treat claims of the existence of a signaling pathway based on treatment with human derived compound with much caution. Such data provide indications, but only indirect evidence that should be verified by more functional approaches including the cloning and characterisation of the relevant receptor.

4.2. Detection

Detecting a given ligand in one's favorite non-classical model organism is probably an excellent option for a comparative endocrinologist. One interesting case to illustrate the complexity of ligand evolution in different organisms is the one of the lamprey steroid hormones. For a long time, it was thought that as a vertebrate, lamprey should produce the classical steroids, estradiol, testosterone, and progesterone, and indeed, these were effectively identified by radioimmunoassays (RIA) and used for physiological (for an example see Bolduc and Sower, 1992) or functional (Thornton, 2001) studies. Nevertheless, in the 1980s, chromatographic tech-

niques and more recently, blood steroid profiles analyzed by high performance liquid chromatography (HPLC), contradicted this view and showed that the main circulating steroids are 15-alpha-hydroxylated (Kime and Callard, 1982; Lowartz et al., 2003). The actual nature of the active steroid hormones and their mechanism of action in lamprey are still a matter of debate (reviewed in Bryan et al., 2008). This example shows that the diversity of ligands in non-model species is often underestimated. In this case the presence of the classical steroids is not questioned, but they are not the physiologically relevant compounds.

Many reports have shown the presence of a typical human hormone in early vertebrates or even in invertebrates, steroids being the compounds for which the number of such reports is the highest. Most often the detection of these products is based on the use of antibodies, generated from human compounds and widely used on human tissue samples for medical application. These antibodies are mainly used for immunohistochemical staining and/or radioimmunoassay and two recent examples illustrate the caveats and uncertainties of such approaches. We have nevertheless to strongly insist on the fact that these examples were not selected because they are particularly dubious but rather because they illustrate numerous studies done in classical comparative endocrinology.

The first example concerns the detection of immunoreactivity for progesterone in the giant Rohde cells of the amphioxus *Branchiostoma belcheri* (Takeda et al., 2003). Using rabbit polyclonal antibodies against progesterone conjugated to bovine serum albumin from two commercial suppliers, the authors performed immunohistochemical staining on adult amphioxus and show staining in giant neurons known as the Rohde cells. The authors cautiously conclude that "progesterone-like" substances are likely to be present in these neurons, suggesting the existence of neurosteroids in this species. They performed a number of controls to avoid problems due to non-specific binding: they replaced the primary antibody by normal rabbit serum, they omitted the primary antibody from the staining reaction and they used PBS instead of the primary antibody. Furthermore, they also performed absorption tests with the antibodies, progesterone, and BSA, respectively, corroborating the fact that the staining is specific. The main problem with this type of paper is that we have no information on the actual specificity of the antibodies. Are these reagents able to recognize only progesterone itself or do they show cross-reactivity with other closely related molecules such as pregnenolone, deoxycorticosterone or 17β-hydroxyprogesterone? What if the amphioxus contains no progesterone but related steroids, such as those found in lamprey? The authors of this study were cautious and spoke of "progesterone-like substances". It is interesting to note that the amphioxus genome contains several, but not all the enzymes implicated in steroidogenesis (e.g. a CYP21 orthologue is missing) and that several of these enzymes are duplicated (e.g. 3β-HSD for which six genes were found in amphioxus, compared to three in human and the SDR family that contains enzymes with 17β-HSD and 11β-HSD activities for which 31 genes are known in human for more than 100 in amphioxus) (Holland et al., 2008). However, only one steroid receptor, located at the base of the vertebrate GR, MR, PR and AR, is present in the amphioxus genome, and, to date, nothing is known about its specificity. Thus it would not be surprising if amphioxus steroidogenesis has been elaborated independently from vertebrate steroidogenesis, and searching for human compounds in amphioxus may simply be unfruitful.

The second example concerns the presence of estrogens in molluscs, and more precisely in the cephalopod *Octopus vulgaris*, an interesting case since it is much referred to in the debate on the ligand-binding evolution of estrogen receptors (Thornton et al., 2003; Keay et al., 2006; Paris et al., 2008b). It has been shown that, in this species, 17β-estradiol and progesterone are found in oviduct and ovarian tissues, and that the concentration of these hormones in females correlates with phases of the reproductive cycle (Di Cosmo et al., 2001; D'Aniello et al., 1996). The doses of hormones in these tissues are measured using a radioimmunoassay and we are faced with the same uncertainties about the specificity of the detection method in this study as in the previous one. In addition, the level of hormones detected both at the level of radioimmunoassay or by HPLC is very close to the lower limits of detection (D'Aniello et al., 1996). From all the data accumulated in *Octopus*, one can conclude that steroid hormone metabolism does exist in this species, but it is difficult to be more conclusive on the precise identity of the steroids that will be found (see below).

These two examples emphasize the difficulties associated with detection based on the use of antibodies. In fact, before the genome era the same situation was found for NR and many reports claimed for the detection of steroid receptors in a wide variety of non-metazoans organisms such as plants or fungi (see Agarwal et al., 1994; Milanesi et al., 2001; Milanesi and Boland, 2006). Despite the fact that, as in the above mentioned studies, all the controls were correct, all these studies displayed artefacts: there is no gene related to steroid receptors outside metazoans (Escriva et al., 1997, 2004). We believe that basically the same situation holds for the immunological detection of NR ligands in invertebrates: perhaps some of these reports are correct but in the absence of more firm evidence this cannot be confirmed (Lafont and Mathieu, 2007). Thus the use of antibodies as an evolutionary probe, even if this is attractive because these experiments are technically easy to perform, should be avoided.

4.3. Other lines of evidence

Other methods of analysis for the presence of specific compounds can be used. This is for example the ability of a given tissue to metabolize a given compound, i.e. to carry out a given reaction (e.g. hydroxylation at the position 3β indicative of a 3β-HSD activity see Di Cosmo et al., 2001). Once again the problem in these experiments is the conclusions that can reasonably be drawn. That a given type of activity exists *in vivo* is interesting but given the diversity of steroidogenic enzymes (or enzymes using different substrates) in metazoans one cannot take for granted that because a reaction occurs with a given labelled intermediate it actually uses this very substrate *in vivo*.

Another approach often used is searching for the presence of the gene encoding an enzyme responsible for a critical step in the synthesis pathway of a given ligand. For example the existence of an aromatase in amphioxus (Castro et al., 2005) argues for the existence of estrogens in this species. This is convincing in this particular case since the phylogeny of the isolated clone was carefully assessed including synteny analysis. This is much less clear in other cases, such as the recent description of putative CYP11A and CYP17 in amphioxus (Mizuta and Kubokawa, 2007). The phylogeny of these clones, taking into account a more exhaustive dataset, with sequences from non-chordate species, shows that the phylogenetical inferences of this report were inexact and that in both cases, the cloned gene are orthologous to a new group of deuterostome CYPs, that was lost in vertebrates (Markov et al., in preparation).

Our critical arguments may well in turn be criticised as, we agree, it is easy to operate previous publications without offering positive alternatives. It seems clear that the best studies are those done with careful analytical chemistry methods. For example HPLC coupled with mass spectrometry allows the clear detection and identification of compounds and offers a very strong and rigorous alternative (Bridgham et al., 2006 see also Lafont and Mathieu, 2007 for a similar conclusion). The problem with this method is

of course that it is much more demanding and expensive than antibody detection. But they have the unique advantage of unambiguously assessing the presence and diversity of the compounds found in a given organism. The recent characterization of steroids present in the nematode *C. elegans* provides an interesting example of the variety of compounds that can be revealed through such a detailed analysis (Motola et al., 2006).

5. Conclusion—evolutionary vs. comparative approaches: an evolutionary shift for comparative endocrinology

Most often, endocrinology studies are realized with a long-term medical objective. This leads to a deformation of the evolutionary perspective, which may be too much "human-centered". Doing comparative endocrinology should therefore first be accompanied by an effort to go again this natural anthropocentric view and to consider equally the evidences coming from different taxa. This may seems obvious, but in fact "rampant" anthropocentrism and a graded view of evolution (e.g. that evolution is progressing toward complexity) is often difficult to avoid as nicely pointed out by Gould (1996).

In fact a key solution to these methodological and interpretation difficulties relies in the comparative strategy itself that should incorporate evidences coming from various organisms and scientific approaches in an integrated manner. An interesting example of the power of such an approach is the case of the suiform aromatases (Gaucher et al., 2004). In this study, by combining bioinformatic, molecular evolution, paleontology, cladistics, structural biology and organic chemistry analysis, the authors propose that the conservation of three subfunctionalised aromatase paralogues in pigs is the result of a selection for Suoidea with larger litter than their ancestors. This selective event has allowed their survival during the global climatic shift that began in the Eocene. Bioinformatic analyses (estimation of divergence times, detection of positive selection by K_a/K_s analysis) were correlated with the presence of residues that were subject to positive selection in the substrate-binding site and with previous experimental data about different substrate specificities for these enzymes. Additionally, a detailed examination of the palaeontological record and of the number of pups in modern artiodactyls as well, correlated with data on global climate changes, led to this quite audacious hypothesis, that is consistent with data of many different research fields. It would be interesting – and possible, thanks to the availability of genome data – to check if such a correlation occurs in other vertebrates and could be statistically significant.

We argue that future comparative endocrinology studies should combine large-scale evolutionary analysis, with several standardised phylogenetic and chemical methods to ensure the robustness of the conclusions. The classical two-by-two comparison that is prone to artefactual conclusions based on partial analysis and the use of poorly refined detection methods, such as those based on antibodies should be replaced by such multidisciplinary studies, even if these approaches are experimentally much more difficult. Comparative endocrinology is thus now facing a real challenge: to perform multidisciplinary evolutionary approaches that will effectively offer solutions, using rigorous technical and conceptual basis, to the long-standing questions of the origin of the endocrine signalling pathways.

Acknowledgements

We are grateful to ARC, CNRS, EMBO, MENRT, and Region Rhone-Alpes for financial support. Work from our laboratory is also funded by the EU Cascade Network of Excellence and the integrated project Crescendo. We thank Barbara Demeneix, François Bonneton and Hector Escriva for critical reading of the manuscript.

References

Agarwal, M.K., Mirshahi, M., Braq, S., Jullienne, A., Leblanc, N., Stibon, F., Guern, J., 1994. *Nicotiana tabacum* contains a putative mineralocorticoid receptor. Biochem. Biophys. Res. Commun. 200, 1230–1238.

Aspöck, G., Kagoshima, H., Niklaus, G., Bürglin, T.R., 1999. *Caenorhabditis elegans* has scores of hedgehog-related genes: sequence and expression analysis. Genome Res. 9, 909–923.

Baker, M.E., 2003. Evolution of adrenal and sex steroid action in vertebrates: a ligand-based mechanism for complexity. Mol. Cell. Endocrinol. 25, 396–400.

Bannister, R., Beresford, N., May, D., Routledge, E.J., Jobling, S., Rand-Weaver, M., 2007. Novel estrogen receptor-related transcripts in *Marisa cornuarietis*, a freshwater snail with reported sensitivity to estrogenic chemicals. Environ. Sci. Technol. 41, 2643–2650.

Bélanger, A., Hum, D.W., Beaulieu, M., Lévesque, E., Guillemette, C., Tchernof, A., Bélanger, G., Turgeon, D., Dubois, S., 1998. Characterization and regulation of UDP-glucuronosyltransferases in steroid target tissues. J. Steroid Biochem. Mol. Biol. 65, 301–310.

Bertrand, S., Brunet, F.G., Escriva, H., Parmentier, G., Laudet, V., Robinson-Rechavi, M., 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. Mol. Biol. Evol. 21, 1923–1937.

Bertrand, S., Thisse, B., Tavares, R., Sachs, L., Chaumot, A., Bardet, P., Escrivà, H., Duffraisse, M., Marchand, O., Safi, R., Thisse, C., Laudet, V., 2007. Unexpected novel relational links uncovered by extensive developmental profiling of nuclear receptor expression. PLoS Genet. 3, e188.

Bolduc, T.G., Sower, S.A., 1992. Changes in brain gonadotropin-releasing hormone, plasma estradiol 17-beta, and progesterone during the final reproductive cycle of the female sea lamprey, *Petromyzon marinus*. J. Exp. Zool. 264, 55–63.

Bonneton, F., Zelus, D., Iwema, T., Robinson-Rechavi, M., Laudet, V., 2003. Rapid divergence of the ecdysone receptor in Diptera and Lepidoptera suggests coevolution between ECR and USP-RXR. Mol. Biol. Evol. 20, 541–553.

Bonneton, F., Brunet, F.G., Kathirithamby, J., Laudet, V., 2006. The rapid divergence of the ecdysone receptor is a synapomorphy for Mecopterida that clarifies the Strepsiptera problem. Insect Mol. Biol. 15, 351–362.

Bonneton, F., Chaumot, A., Laudet, V., 2008. Annotation of *Tribolium* nuclear receptors reveals an increase in evolutionary rate of a network controlling the ecdysone cascade. Insect Biochem. Mol. Biol. 38, 416–429.

Bridgham, J.T., Carroll, S.M., Thornton, J.W., 2006. Evolution of hormone–receptor complexity by molecular exploitation. Science 312, 97–101.

Bryan, M.B., Scott, A.P., Li, W., 2008. Sex steroids and their receptors in lampreys. Steroids 73, 1–12.

Calléja, C., Messaddeq, N., Chapellier, B., Yang, H., Krezel, W., Li, M., Metzger, D., Mascrez, B., Ohta, K., Kagechika, H., Endo, Y., Mark, M., Ghyselinck, N.B., Chambon, P., 2006. Genetic and pharmacological evidence that a retinoic acid cannot be the RXR-activating ligand in mouse epidermis keratinocytes. Genes Dev. 20, 1525–1538.

Cardoso, J.C.R., Pinto, V.C., Vieira, F.A., Clark, M.S., Power, D.M., 2006. Evolution of secretin family GPCR members in the metazoa. BMC Evol. Biol. 6, 108.

Cardoso, J., de Vet, E., Louro, B., Elgar, G., Clark, M., Power, D., 2007. Persistence of duplicated PAC1 receptors in the teleost, *Sparus auratus*. BMC Evol. Biol. 7, 221.

Carlsson, P., Mahlapuu, M., 2002. Forkhead transcription factors: key players in development and metabolism. Dev. Biol. 250, 1–23.

Castro, L.F.C., Santos, M.M., Reis-Henriques, M.A., 2005. The genomic environment around the Aromatase gene: evolutionary insights. BMC Evol. Biol. 5, 43.

Catford, J.G., Staehelin, C., Larose, G., Piché, Y., Vierheilig, H., 2006. Systemically suppressed isoflavonoids and their stimulating effects on nodulation and mycorrhization in alfalfa split-root systems. Plant Soil 285, 257–266.

Cheng, G., Yeater, K.M., Hoyer, L.L., 2006. Cellular and molecular biology of *Candida albicans* estrogen response. Eukaryot. Cell. 5, 180–191.

Crollius, H.R., Weissenbach, J., 2005. Fish genomics and biology. Genome Res. 15, 1675–1682.

D'Aniello, A., Cosmo, A.D., Cristo, C.D., Assisi, L., Botte, V., Fiore, M.M.D., 1996. Occurrence of sex steroid hormones and their binding proteins in *Octopus vulgaris* lam. Biochem. Biophys. Res. Commun. 227, 782–788.

Danchin, E.G.J., Abi-Rached, L., Gilles, A., Pontarotti, P., 2003. Conservation of the MHC-like region throughout evolution. Immunogenetics 55, 141–148.

Danchin, E., Gouret, P., Pontarotti, P., 2006. Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. BMC Evol. Biol. 6, 5.

Dean, A.M., Thornton, J.W., 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nat. Rev. Genet. 8, 675–688.

Dehal, P., Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3, e314.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.

Di Cosmo, A., Cristo, C.D., Paolucci, M., 2001. Sex steroid hormone fluctuations and morphological changes of the reproductive system of the female of *Octopus vulgaris* throughout the annual cycle. J. Exp. Zool. 289, 33–47.

Escriva, H., Safi, R., Hänni, C., Langlois, M.C., Saumitou-Laprade, P., Stehelin, D., Capron, A., Pierce, R., Laudet, V., 1997. Ligand binding was acquired dur-

ing evolution of nuclear receptors. Proc. Natl. Acad. Sci. U.S.A. 94, 6803–6808.
Escriva, H., Delaunay, F., Laudet, V., 2000. Ligand binding and nuclear receptor evolution. Mol. Cell. Endocrinol. 22, 717–727.
Escriva, H., Bertrand, S., Laudet, V., 2004. The evolution of the nuclear receptor superfamily. Essays Biochem. 40, 11–26.
Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., Laudet, V., 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2, e102.
Farid, N.R., Szkudlinski, M.W., 2004. Minireview: structural and functional evolution of the thyrotropin receptor. Endocrinology 145, 4048–4057.
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.
Garcia-Fernández, J., Holland, P.W., 1994. Archetypal organization of the amphioxus Hox gene cluster. Nature 370, 563–566.
Gaucher, E.A., Graddy, L.G., Li, T., Simmen, R.C.M., Simmen, F.A., Schreiber, D.R., Liberles, D.A., Janis, C.M., Benner, S.A., 2004. The planetary biology of cytochrome P450 aromatases. BMC Biol. 2, 19.
Gould, S.J., 1996. Full House. Harmony Books, New York.
Heyland, A., Price, D.A., Bodnarova-Buganova, M., Moroz, L.L., 2006. Thyroid hormone metabolism and peroxidase function in two non-chordate animals. J. Exp. Zool. B Mol. Dev. Evol. 306, 551–566.
Holland, L.Z., Albalat, R., Azumi, K., Benito-Gutierrez, E., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L.J., Garcia-Fernandez, J., Ferrier, D.E.K., Gibson-Brown, J.J., Gissi, C., Godzik, A., Hallbook, F., Hirose, D., Hosomichi, K., Ikuta, T., Inoko, H., Kasahara, M., Kasamatsu, J., Kawashima, T., Kimura, A., Kobayashi, M., Kozmik, Z., Kubokawa, K., Laudet, V., Litman, G.W., McHardy, A.C., Meulemans, D., Nonaka, M., Olinski, R.P., Pancer, Z., Pestarino, M., Rast, J.P., Rigoutsos, I., Roch, G., Saiga, H., Sasakura, Y., Satake, M., Satou, Y., Schubert, M., Sherwood, N., Shiina, T., Takatori, N., Tello, J., Vopalensky, P., Wada, S., Xu, A., Ye, Y., Yoshida, K., Yoshizaki, F., Yu, J.K., Zhang, Q., Zmasek, C.M., Putnam, N.H., Rokhsar, D.S., Satoh, N., Holland, P.W.H., 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res. 18, 1100–1111.
Iwema, T., Billas, I.M.L., Beck, Y., Bonneton, F., Nierengarten, H., Chaumot, A., Richards, G., Laudet, V., Moras, D., 2007. Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. EMBO J. 26, 3770–3782.
Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., et al., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946–957.
Kajiwara, M., Kuraku, S., Kurokawa, T., Kato, K., Toda, S., Hirose, H., Takahashi, S., Shibata, Y., Iguchi, T., Matsumoto, T., Miyata, T., Miura, T., Takahashi, Y., 2006. Tissue preferential expression of estrogen receptor gene in the marine snail, *Thais clavigera*. Gen. Comp. Endocrinol. 148, 315–326.
Kakizawa, T., Nishio, S., Triqueneaux, G., Bertrand, S., Rambaud, J., Laudet, V., 2007. Two differentially active alternative promoters control the expression of the zebrafish orphan nuclear receptor gene Rev-erbalpha. J. Mol. Endocrinol. 38, 555–568.
Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., Kohara, Y., 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714–719.
Keay, J., Bridgham, J.T., Thornton, J.W., 2006. The *Octopus vulgaris* estrogen receptor is a constitutive transcriptional activator: evolutionary and functional implications. Endocrinology 147, 3861–3869.
Kime, D.E., Callard, G.V., 1982. Formation of 15 alpha-hydroxylated androgens by the testis and other tissues of the sea lamprey, *Petromyzon marinus*, in vitro. Gen. Comp. Endocrinol. 46, 267–270.
Klovins, J., Haitina, T., Fridmanis, D., Kilianova, Z., Kapa, I., Fredriksson, R., Gallo-Payet, N., Schiöth, H.B., 2004. The melanocortin system in Fugu: determination of POMC/AGRP/MCR gene repertoire and synteny, as well as pharmacology and anatomical distribution of the MCRs. Mol. Biol. Evol. 21, 563–579.
Krasowski, M.D., Yasuda, K., Hagey, L.R., Schuetz, E.G., 2005. Evolution of the pregnane × receptor: adaptation to cross-species differences in biliary bile salts. Mol. Endocrinol. 19, 1720–1739.
Lafont, R., Mathieu, M., 2007. Steroids in aquatic invertebrates. Ecotoxicology 16, 109–130.
Lagerström, M.C., Fredriksson, R., Bjarnadóttir, T.K., Fridmanis, D., Holmquist, T., Andersson, J., Yan, Y., Raudsepp, T., Zoorob, R., Kukkonen, J.P., Lundin, L., Klovins, J., Chowdhary, B.P., Postlethwait, J.H., Schiöth, H.B., 2005. Origin of the prolactin-releasing hormone (PRLH) receptors: evidence of coevolution between PRLH and a redundant neuropeptide Y receptor during vertebrate evolution. Genomics 85, 688–703.
Laudet, V., 1997. Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. J. Mol. Endocrinol. 19, 207–226.
Laudet, V., Gronemeyer, H., 2005. The Nuclear Receptor FactsBook. Academic Press, London.
Lalueza-Fox, C., Römpler, H., Caramelli, D., Stäubert, C., Catalano, G., et al., 2007. Melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. Science 318, 1453–1455.
Li, Y., Wallis, M., Zhang, Y., 2005. Episodic evolution of prolactin receptor gene in mammals: coevolution with its ligand. J. Mol. Endocrinol. 35, 411–419.
Liu, J.C., Makova, K.D., Adkins, R.M., Gibson, S., Li, W.H., 2001. Episodic evolution of growth hormone in primates and emergence of the species specificity of human growth hormone receptor. Mol. Biol. Evol. 18, 945–953.
Lowartz, S., Petkam, R., Renaud, R., Beamish, F.W.H., Kime, D.E., Raeside, J., Leatherland, J.F., 2003. Blood steroid profile and in vitro steroidogenesis by ovarian follicles and testis fragments of adult sea lamprey, *Petromyzon marinus*. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 134, 365–376.
Macqueen, D.J., Johnston, I.A., 2008. An update on MyoD evolution in teleosts and a proposed consensus nomenclature to accommodate the tetraploidization of different vertebrate genomes. PLoS ONE 3, e1567.
Madani, N.D., Malloy, P.J., Rodriguez-Pombo, P., Krishnan, A.V., Feldman, D., 1994. *Candida albicans* estrogen-binding protein gene encodes an oxidoreductase that is inhibited by estradiol. Proc. Natl. Acad. Sci. U.S.A. 91, 922–926.
Markov, G., Lecointre, G., Demeneix, B., Laudet, V., 2008. The «street light syndrome», or how protein taxonomy can bias experimental manipulations. Bioessays 30, 349–357.
Matsumoto, T., Nakamura, A.M., Mori, K., Akiyama, I., Hirose, H., Takahashi, Y., 2007. Oyster estrogen receptor: cDNA cloning and immunolocalization. Gen. Comp. Endocrinol. 151, 195–201.
McPartland, J.M., Norris, R.W., Kilpatrick, C.W., 2007. Coevolution between cannabinoid receptors and endocannabinoid ligands. Gene 397, 126–135.
Mic, F.A., Molotkov, A., Benbrook, D.M., Duester, G., 2003. Retinoid activation of retinoic acid receptor but not retinoid X receptor is sufficient to rescue lethal defect in retinoic acid synthesis. Proc. Natl. Acad. Sci. U.S.A. 100, 7135–7140.
Milanesi, L., Monje, P., Boland, R., 2001. Presence of estrogens and estrogen receptor-like proteins in *Solanum glaucophyllum*. Biochem. Biophys. Res. Commun. 289, 1175–1179.
Milanesi, L., Boland, R., 2006. Presence of vitamin D3 receptor (VDR)-like proteins in *Solanum glaucophyllum*. Physiol. Plant. 128, 341–350.
Mizuta, T., Kubokawa, K., 2007. Presence of sex steroids and cytochrome P450 genes in amphioxus. Endocrinology 148, 3554–3565.
Motola, D.L., Cummins, C.L., Rottiers, V., Sharma, K.K., Li, T., Li, Y., Suino-Powell, K., Xu, H.E., Auchus, R.J., Antebi, A., Mangelsdorf, D.J., 2006. Identification of ligands for DAF-12 that govern dauer formation and reproduction in *C. elegans*. Cell 124, 1209–1223.
Moyle, W.R., Campbell, R.K., Myers, R.V., Bernard, M.P., Han, Y., Wang, X., 1994. Co-evolution of ligand–receptor pairs. Nature 368, 251–255.
Muffato, M., Crollius, H.R., 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays 30, 122–134.
Nelson, D.R., 2005. Gene nomenclature by default, or BLASTing to Babel. Hum. Genom. 2, 196–201.
Norris, D.O., 2007. Vertebrate Endocrinology. Elsevier Academic Press, San Diego.
O'Halloran, D.M., Fitzpatrick, D.A., McCormack, G.P., McInerney, J.O., Burnell, A.M., 2006. The molecular phylogeny and functional significance of a nematode specific clade of heterotrimeric G-protein α-subunit genes. J. Mol. Evol. 63, 87–94.
Ogasawara, M., Lauro, R.D., Satoh, N., 1999. Ascidian homologs of mammalian thyroid peroxidase genes are expressed in the thyroid-equivalent region of the endostyle. J. Exp. Zool. 285, 158–169.
Ogasawara, M., 2000. Overlapping expression of amphioxus homologs of the thyroid transcription factor-1 gene and thyroid peroxidase gene in the endostyle: insight into evolution of the thyroid gland. Dev. Genes. Evol. 210, 231–242.
Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, Heidelberg.
Panopoulou, G., Poustka, A.J., 2005. Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. Trends Genet. 21, 559–567.
Pantalacci, S., Chaumont, C., Benoît, G., Sadier, A., Delsuc, F., Douzery, E., Laudet, V., 2008. Conserved features and evolutionary shifts of the EDA signaling pathway involved in vertebrate skin appendage development. Mol. Biol. Evol. 25, 912–928.
Paris, M., Escriva, H., Schubert, M., Brunet, F., Brtko, J., Cieselski, F., Roecklin, D., Vivat-Hannah, V., Cravedi, J.-P., Scanlan, T.S., Renaud, J.-P., Holland, N.D., Laudet, V., 2008a. Amphioxus metamorphosis and the origin of the thyroid hormone signaling pathway. Curr. Biol. 18, 825–830.
Paris, M., Pettersson, K., Schubert, M., Bertrand, S., Pongratz, I., Escriva, H., Laudet, V., 2008b. An amphioxus orthologue of the estrogen receptor that does not bind estradiol: insight into estrogen receptor evolution. BMC Evol. Biol., in press.
Park, P.S., Palczewski, K., 2005. Diversifying the repertoire of G protein-coupled receptors through oligomerization. Proc. Natl. Acad. Sci. U.S.A. 102, 8793–8794.
Pinaire, J.A., Reifel-Miller, A., 2007. Therapeutic potential of retinoid × receptor modulators for the treatment of the metabolic syndrome. PPAR Res., 941–956.
Poulin, M.J., Simard, J., Catford, J.G., Labrie, F., Piche, Y., 1997. Response of symbiotic endomycorrhizal fungi to estrogens and antiestrogens. Mol. Plant Microb. Interact., 10.
Prüll, C.-R., 2003. Part of a scientific master plan? Paul Ehrlich and the origins of his receptor concept. Med. Hist. 47, 332–356.
Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., Jurka, J., Genikhovich, G., Grigoriev, I.V., Lucas, S.M., Steele, R.E., Finnerty, J.R., Technau, U., Martindale, M.Q., Rokhsar, D.S., 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86–94.
Reschly, E.J., Bainy, A.C.D., Mattos, J.J., Hagey, L.R., Bahary, N., Mada, S.R., Ou, J., Venkataramanan, R., Krasowski, M.D., 2007. Functional evolution of the vitamin D and pregnane X receptors. BMC Evol. Biol. 7, 222.

Requena, N., Serrano, E., Ocón, A., Breuninger, M., 2007. Plant signals and fungal perception during arbuscular mycorrhiza establishment. Phytochemistry 68, 33–40.
Revankar, C.M., Cimino, D.F., Sklar, L.A., Arterburn, J.B., Prossnitz, E.R., 2005. A transmembrane intracellular estrogen receptor mediates rapid cell signaling. Science 307, 1625–1630.
Rikke, B.A., Murakami, S., Johnson, T.E., 2000. Paralogy and orthology of tyrosine kinases that can extend the life span of *Caenorhabditis elegans*. Mol. Biol. Evol. 17, 671–683.
Robinson-Rechavi, M., Maina, C.V., Gissendanner, C.R., Laudet, V., Sluder, A., 2005. Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. J. Mol. Evol. 60, 577–586.
Rogozin, I.B., Wolf, Y.I., Carmel, L., Koonin, E.V., 2007. Analysis of rare amino acid replacements supports the coelomata clade. Mol. Biol. Evol. 24, 2594–2597.
Scervino, J.M., Ponce, M.A., Erra-Bassells, R., Vierheilig, H., Ocampo, J.A., Godeas, A., 2005. Arbuscular mycorrhizal colonization of tomato by Gigaspora and Glomus species in the presence of root flavonoids. J. Plant. Physiol. 162, 625–633.
Simionato, E., Ledent, V., Richards, G., Thomas-Chollier, M., Kerner, P., Coornaert, D., Degnan, B.M., Vervoort, M., 2007. Origin and diversification of the basic helix–loop–helix gene family in metazoans: insights from comparative genomics. BMC Evol. Biol. 7, 33.
Simões-Costa, M.S., Azambuja, A.P., Xavier-Neto, J., 2008. The search for non-chordate retinoic acid signaling: lessons from chordates. J. Exp. Zool. B Mol. Dev. Evol. 310, 54–72.
Sluder, A.E., Maina, C.V., 2001. Nuclear receptors in nematodes: themes and variations. Trends Genet. 17, 206–213.
Takeda, N., Kubokawa, K., Matsumoto, G., 2003. Immunoreactivity for progesterone in the giant Rohde cells of the amphioxus, *Branchiostoma belcheri*. Gen. Comp. Endocrinol. 132, 379–383.
Thomas, J.H., 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3, e67.
Thornton, J.W., 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. Proc. Natl. Acad. Sci. U.S.A. 98, 5671–5676.
Thornton, J.W., Need, E., Crews, D., 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301, 1714–1717.
Thornton, J.W., 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat. Rev. Genet. 5, 366–375.
Volff, J., 2005. Genome evolution and biodiversity in teleost fish. Heredity 94, 280–294.
Waddell, P.J., Kishino, H., Ota, R., 2007. Phylogenetic methodology for detecting protein interactions. Mol. Biol. Evol. 24, 650–659.
Wallis, M., 2001. Episodic evolution of protein hormones in mammals. J. Mol. Evol. 53, 10–18.
Wittbrodt, J., Meyer, A., Schartl, M., 1998. More genes in fish? Mol. Cell. Endocrinol. 20, 511–512.
Wu, W., Niles, E.G., El-Sayed, N., Berriman, M., LoVerde, P.T., 2006. *Schistosoma mansoni* (Platyhelminthes, Trematoda) nuclear receptors: sixteen new members and a novel subfamily. Gene 366, 303–315.
Wu, W., Niles, E.G., LoVerde, P.T., 2007. Thyroid hormone receptor orthologues from invertebrate species with emphasis on *Schistosoma mansoni*. BMC Evol. Biol. 7, 150.
Wyder, S., Kriventseva, E., Schröder, R., Kadowaki, T., Zdobnov, E., 2007. Quantification of ortholog losses in insects and vertebrates. Genome Biol. 8, R242.
You, L., 2004. Steroid hormone biotransformation and xenobiotic induction of hepatic steroid metabolizing enzymes. Chem. Biol. Interact. 147, 233–246.

CHAPTER 2

WHAT DOES EVOLUTION TEACH US ABOUT NUCLEAR RECEPTORS?

GABRIEL MARKOV[1,2], FRANÇOIS BONNETON[1], AND VINCENT LAUDET[1]

[1]*Institut de Génomique Fonctionnelle de Lyon; Université de Lyon; Université Lyon 1; CNRS; INRA; Ecole Normale Supérieure de Lyon; Lyon cedex, France*

[2] *USM 501 – Evolution des Régulations Endocriniennes. Muséum National d'Histoire Naturelle, Paris, France*

Abstract: In this chapter we first summarise the current knowledge about the phylogenetic spectrum of nuclear receptors (NRs). Then, we discuss how studying their diversity can be helpful to make insights about their evolution. Significant attention is paid to the evolution of ligand-binding ability. Recent evolutionary and functional data have challenged the traditional concept of ligand, providing a more complex view of the mechanisms by which the transcriptional activity of NRs can be modulated. Finally, we argue that the evolutionary analysis of NRs has contributed to a conceptual shift of our understanding of nuclear receptors, from highly specific endocrine regulators to a promiscuous metabolic rheostat.

2.1. INTRODUCTION

Nuclear receptors (NRs) are classically defined as ligand-activated transcription factors that allow the regulation of target genes by small lipophilic molecules such as hormones (e.g. thyroid hormones or steroids), morphogen (e.g. retinoic acid) or dietary components (e.g. fatty acids). All built with a similar organization, NRs are nevertheless regulated by a wide diversity of compounds and are implicated in a tremendous diversity of physiological and metabolic processes. The question of the origin of such a system has received much attention because of its intrinsic interest, but also because it provides a nice experimental and conceptual framework to understand the origin of complex regulatory systems. Indeed, the NR family is a model of choice to address evolutionary issues because these proteins are present, with various physiological roles, in a broad range of well-studied organisms. Due to their importance, sequence, structural and physiological data have all been used to answer evolutionary questions. Notably, since NRs are involved in the integration

C. Bunce, M.J. Campbell (eds.), *Nuclear Receptors*, Proteins and Cell Regulation 8, xxx–xxx
DOI 10.1007/978-90-481-3303-1_2,

of genomic and environmental processes, they are a critical link to understand the molecular basis of phenotypic plasticity.

2.2. NRs PHYLOGENY AND CLASSIFICATION

Before addressing questions about NR evolution, it is necessary to place the current knowledge on a solid phylogenetic framework. Indeed, classification is the first step required prior to any evolutionary analysis, since it is only by knowing the relationships between taxa (either proteins or organisms) that one can safely propose evolutionary hypothesis to be tested. The presence of two functionally conserved domains, the DNA-binding domain (DBD) and the ligand-binding domain (LBD) have proved highly informative for tree reconstruction, since it allows the generation of robust phylogenies [1–3].

The phylogeny of NRs provided a framework to establish a nomenclature for the family, a particularly useful tool given the increasing number of new NRs sequences coming from many different organisms [4]. According to this nomenclature, the NR family is divided into six subfamilies (NR1 to NR6). Each of these subfamilies is a robust monophyletic group, in which all receptors clustered in a subfamily originate from a single ancestor. The precise relationships between the six subfamilies are still unclear, blurring our views about the origin of the family itself (see below). It is interesting to note that two subfamilies (I and IV) cluster receptors that are able to interact with RXR in vertebrates, suggesting that this feature is not common to the whole family (Figure 2.3) [5]. Similarly, in vertebrates, subfamily III clusters members that are able to dimerize on palindromic elements. In contrast to this link between evolutionary history and DNA binding activity, no link between the ligand binding ability and the phylogeny has been detected. Steroid receptors are present in different subfamilies (I and III), and strongly related receptors (e.g. within the NR1 family), bind molecules as different as thyroid hormones (NR1A), retinoids (NR1B) or prostaglandins (NR1C).

Also noteworthy is the fact that the family contains proteins that lack one of the two conserved domains. In the official nomenclature, these proteins are artificially gathered into a specific subfamily (NR0) that has no biological meaning in itself (that is, all its members do not share a specific ancestral receptor [4]). Without a LBD, the protein is not a receptor, although it can act as a classical ligand-independent transcription factor. The well-known gap segmentation gene *knirps*, a transcriptional repressor in insects, is a good example of the NR0A group [6]. Without a DBD (group NR0B), the protein cannot act as a transcription factor. For example, the Small Heterodimer Partner (SHP; NR0B2) is an orphan corepressor of various transcription factors, including nuclear receptors [7] whereas DAX-1 (NR0B1) plays a role in sex determination mechanisms in mammals. These two paralogues are distantly related to the TLL group within subfamily II. Members of the NR0 subfamily provide interesting examples of how protein domains are reshuffled during evolution, a major source of molecular innovation [8]. In this view, proteins are composed of

modules that follow their own evolutionary path and the phylogeny of a given protein is not necessarily identical to the phylogenies of its constituent modules. In the NR family, this remains an exception.

2.3. NR COMPLEXITY IS NOT LIMITED TO VERTEBRATES

In addition to an established phylogeny, the understanding of NR evolution requires a better knowledge of the phylogenetic distribution of receptors in a various sets of metazoans [9, 10]. This also allows discovery of a much more diverse family than originally anticipated by studying only NRs from common model organisms such as either human, mouse or *Drosophila*.

Figure 2.1 summarise the current knowledge on the phylogenetic distribution of NRs, with the addition of some notable events concerning specific model organisms. It should be underlined that the two best-studied non-vertebrate models, i.e. *Drosophila* and *Caenorhabditis*, are members of the group ecdysozoa, which cannot be taken as a picture of the 'ancestral condition' of NR functioning. This figure shows that the last common ancestor of all bilaterian animals, *Urbilateria*, which most likely possessed about 22–25 receptors, and subsequently complex events of gene loss, gene duplication and domain shuffling occurred. Some receptors such as NR3A/ER or NR1A/TR, were lost in tunicates and ecdyzosoans, indicating that such animals should not be taken as representatives of a 'primitive state'. On the contrary, the physiological regulatory networks of these organisms are certainly as much derived from earlier organisms as vertebrate networks are. In vertebrates, the number of receptors and their phylogenetic relationships fit very well with whole genome duplication events [11]. There are also many examples of duplication of one peculiar receptor. The amphioxus *Branchiostoma floridae* has ten copies of the NR1H, due to lineage-specific duplication [12], and, more spectacular, the NR2A of *Caenorhabditis elegans* has about 250 copies [13, 14]. Concerning the nematode receptors, it should be mentioned that many of them have diverged considerably, with some receptors that are linked to no clear group, for example the 'NR1K' of *Onchocerca volvulus* [15]. Such sequences may be very divergent forms of NRs that are otherwise well conserved among metazoans. It should be possible to test this exciting hypothesis using recent data about the genomes of *Brugia malayi* [16], *Meloidogyne incognita* [17], and *Pristionchus pacificus* [18]. The very different life styles of these species may also allow correlation between biology (nutrition, ecology, reproduction) and NR duplications. The finding of NRs with 2 DBDs in various lophotrochozoans and in *Daphnia* opens other promising research fields, suggesting once again parallel losses in nematodes, insects and maybe vertebrates [19].

In short, the emerging complete picture of metazoan NRs indicates that, in spite of the conservation of some basic mechanisms, receptors present in various metazoan phyla are extremely diverse. The characterization of NRs throughout the whole metazoan biodiversity thus offer a view on the real flexibility of the NR structural

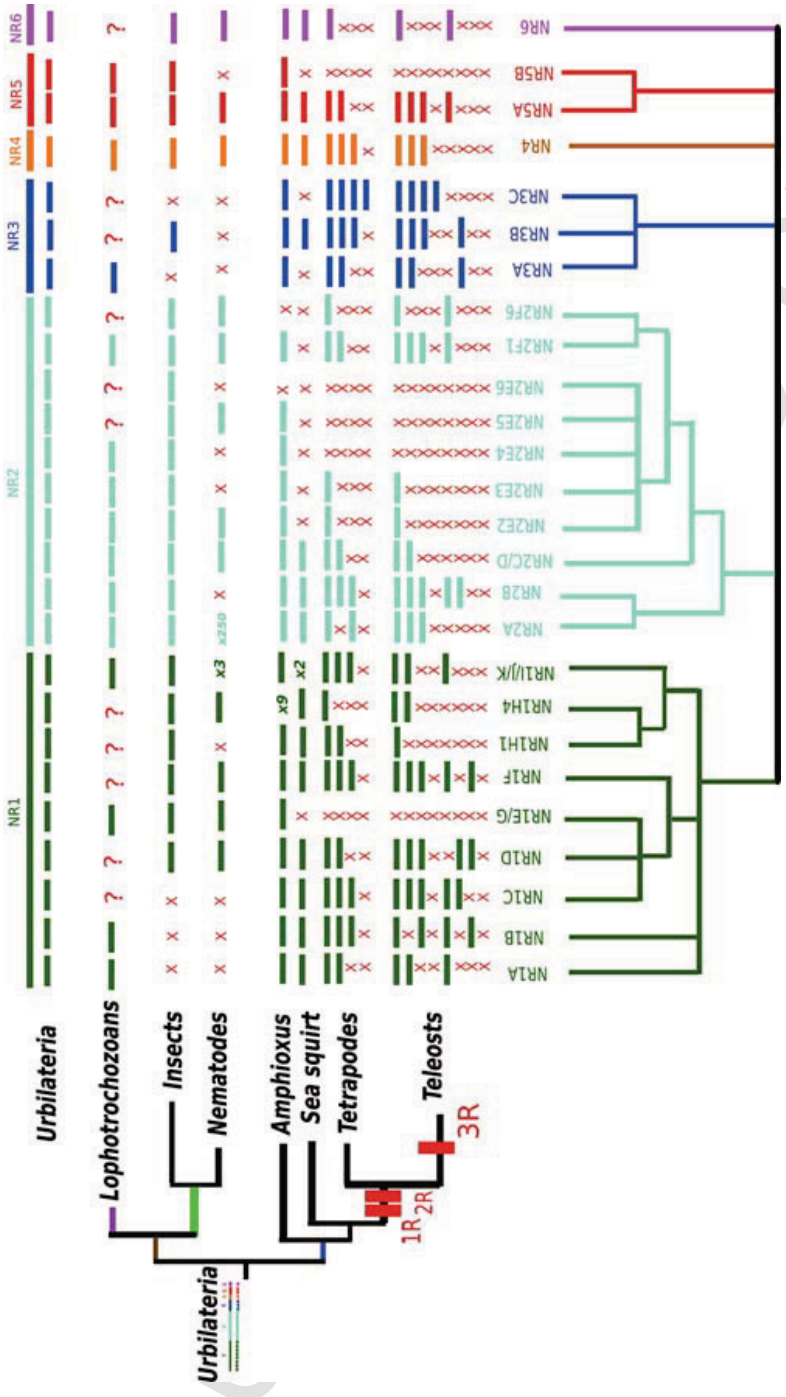


Figure 2.1. Phylogeny of the NR family and NR contain of some bilaterian genomic models. The following species were taken as representative of their taxon: *Tribolium castaneum* for insects [64], *Caenorhabditis elegans* for nematodes [10], *Homo sapiens* for tetrapodes [10] and *Danio rerio* for teleosts [11]. For lophotrochozoans, given the lack of extensive genomic search for NRs in the litterature, we merged partial data about various species [10, 65]. The first line represents the inferred NR contain of *Urbilateria*, the last common ancestor of all bilaterian animals [10]. Orthologous receptors are on the same column. Lost receptors are indicated by *red crosses*, whereas missing data are indicated by *red question mark*; lineage-specific duplications are indicated by « x*N* », where *N* is the number of paralogs. The three rounds or vertebrate whole-genome duplications that led to four paralog groups in tetrapodes and eight paralog groups in teleosts are indicated by 1R, 2R and 3R

modules. NRs are conserved proteins that were prone to some changes during evolution and a future challenge is to better understand to what extent these changes have contributed to phenotypic plasticity.

2.4. NR-LIKE ARE FOUND THROUGHOUT THE TREE OF LIFE

The origin of NRs is still unknown. A decade ago, a PCR screen indicated that NRs are found only in metazoans [9]. Recently it was confirmed that NRs are absent from the genome of the unicellular *Monosiga brevicollis*, which belongs to choanoflagellates, the sister group of animals [20]. The metazoan-specific DBD of nuclear receptors contains two C4-zinc fingers that are structurally related to the GATA C4-zinc fingers, which are found in all eukaryotes. It is possible that Nuclear Receptor DBD arose by duplication of a single ancestral C4-Zinc finger. By contrast, the LBD of nuclear receptors share no similarity with other domains outside animals. However, NR-like proteins were identified recently in the budding yeast. The heterodimeric transcription factors Oaf1/Pip2 are bound and regulated by fatty acids (oleate) through a mechanism that is very similar to PPAR/RXR [21]. Even more surprising is the suggestion, based on structure predictions, deletion and mutation analysis, that these proteins contain a LBD with a NR folding. Since the sequence identity is not significant, it is impossible to determine whether these resemblances are due to either homology or homoplasic evolution. Similarly, the yeast transcription factors Pdr1p/Pdr3p are regulated by xenobiotics, like the PXR nuclear receptor [22]. All these four proteins contain a zinc-finger DBD (Zn6Cys2) and orthologs of the putative LBD were found in other ascomycetes [21] (Figure 2.2). Therefore, fungi can use ligand-regulated transcription factors that share many functional and structural characteristics with NRs of animals (Figure 2.2).

It is interesting to recall that other ligand-activated transcription factors exist in animals. Indeed, the aryl hydrocarbon receptor (AHR) is a member of the bHLH-PAS family that contains a DBD of the basic-Helix-Loop-Helix type (bHLH) and a Per-Arnt-Sim (PAS) domain involved in the ligand binding activity [23] (Figure 2.2). Like NRs, the AHR can be bound and activated by a wide diversity of small lipophilic ligands that can act as either signalling molecules or xenobiotics (including dioxin).

Unfortunately, structural data of the AHR LBD are currently not yet available. Cross-talk interactions exist between NRs and AHR in the steroids and retinoids pathways of mammals [24, 25]. Similar relationships have been found in insects between the bHLH-PAS protein Methoprene-tolerant (Met) and the ecdysone receptor, which is a heterodimer between two nuclear receptors: ECR (NR1H) and USP (NR2B) [26]. Finally, NR coactivators such as the p160 proteins (SRC1, TIF2 and ACTR) are also bHLH-PAS proteins. Therefore, the possibility of interactions between NRs and bHLH-PAS pathways might well be a common theme in animals. These examples show that any information obtained for one of the two groups of receptor can lead to interesting suggestions for the other group, despite totally different DBD and LBD.

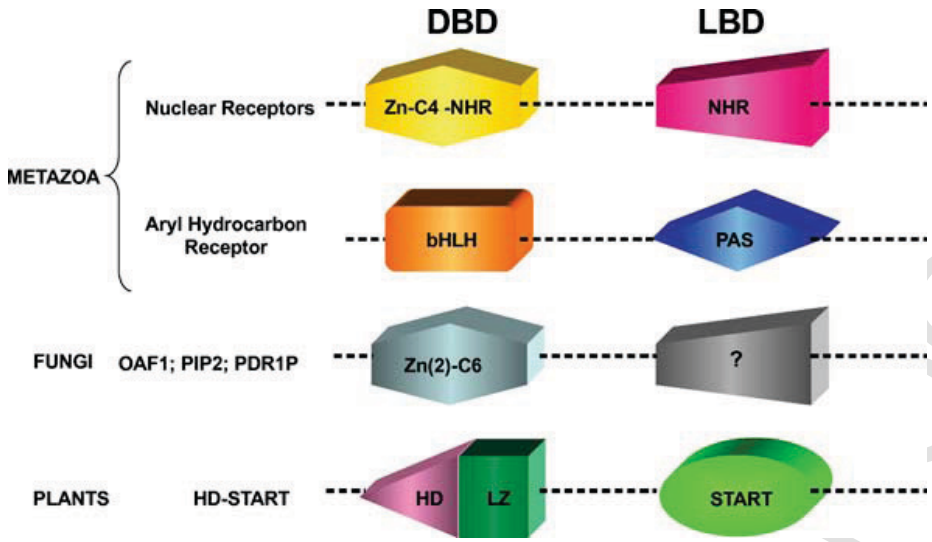


Figure 2.2. NR and NR-like transcription factors in eukaryotes. The domain structures of some known metazoan ligand-activated transcription factors are compared. They all possess a DBD and a LBD, but from different domain families, so they are not homologous

Plants also possess proteins that combine a DBD and a LBD: the HD-START family (Figure 2.2). These transcription factors contain a homeodomain associated with a leucine zipper important for dimerisation. The Steroidogenic Acute Regulatory-related lipid Transfer (START) domain was first identified in mammals as a lipid-sterol binding motif [27]. The crystal structure of several mammalian START domains revealed a hydrophobic tunnel that allows the transport of sterol or phospholipids [28]. The HD and START domains are combined with various other protein motifs in eukaryotes and prokaryotes. However, surprisingly, only within the plant kingdom are they associated together [29]. Several of these HD-START proteins are implicated in cell differentiation during plant development, possibly by linking the lipid metabolic state of the cell to the regulation of transcription [29]. Unfortunately, START sequences of plants are not closely related to those of animals, and their structure is unknown, as well as their ligands. Hydrophobic steroid hormones are very similar in plants and animals but the identified receptors are totally different [30]. Nevertheless, the plant-specific HD-START family may have an analogous role to the one of NRs in animals.

Finally, it is important to recall that what seems to be an exception in eukaryotes is actually the rule in eubacteria, where the regulation of transcription is based on proteins that bind to specific DNA sequences in a ligand-dependent manner [31]. One of the best example of this mechanism is the *lac* operon, the first discovered case of gene regulation. The repressor (lacI) undergoes a conformational transition in response to bound ligands that increases or decreases its affinity for the operator DNA of the *lac* operon [32]. NRs act fundamentally in a similar allosteric fashion. The structure of the repressor revealed a modular structure with four domains: a DBD of the HLH type; a hinge region; the LBD forming a sugar binding pocket;

a C-terminal helix important for tetramerization [3]. This general organisation is familiar to those studying NRs, which are usually organised in 4–5 domains: A/B, C (DBD), D (hinge), E (LBD) and sometimes a F domain.

Beyond these simplistic comparisons, the key point is that ligand-dependent gene regulation is probably a very ancient mechanism. In that perspective, the mode of action of NRs could be qualified as 'primitive', in the sense that it would represent an example of an older and widespread mechanism already present in the late common ancestor to prokaryotes and eukaryotes.

2.5. WHAT IS A NR-LIGAND?

In biochemistry, a ligand is a substance that is able to bind to and form a complex with a biomolecule in a biological context. In a narrower sense, it is a signal-triggering molecule binding to a site on a target protein, by intermolecular forces such as ionic bonds, hydrogen bonds and Van der Waals forces. The docking (association) is usually reversible (dissociation). Actually, irreversible covalent binding between a ligand and its target molecule is rare in biological systems.

The nature of the NR ligand is a question that has been strongly biased by the history of the discovery of NRs. The first NR ligands were classical hormones of the endocrinology field, like steroid hormones or thyroid hormones, hence the name often given to the family: the steroid/thyroid hormone receptor family. When the receptor for ecdysone was cloned, it was clear for everyone in the field that NRs were high affinity receptors (at the nanomolar range) for very specific compounds with a hormonal function [33]. The identification of the first orphan receptors in no way changed this paradigm and many pharmaceutical companies performed high-throughput screens on orphan receptors in order to discover new hormonal ligands. Nevertheless, the discovery of the metabolic receptors and, prominently, of the PPARs that were shown to bind a wide diversity of compounds (including fatty acids with an affinity in the micromolar range), provided the first clue that the situation was more complex than previously expected. This view is reinforced by data showing that 9-cis RA may only be a pharmacological ligand of RXR, at least in mammals, and that its ligands are rather fatty acids [34]. The characterization of xenobiotic regulators such as PXR and CAR have also much broadened our view since these receptors bind to an extremely wide variety of unrelated compounds such as rifampicine or RU486.

In addition, several intriguing results have highlighted the tremendous diversity of interactions that can exist between NRs and small molecules. We are now far from the key/lock model of a stable and simple interaction between a hormone (the key) and a receptor (the lock)! Among recent observations that modified these views on ligand binding it is interesting to mention the ability of ligands to bind unique ligand binding pockets; e.g. FMOC-Leu on PPARγ [35], and the existence of a still much discussed second ligand binding site in an estrogen receptor [36]; the regulation of a receptor activity by gas (NO and CO), which controls the redox status of

a heme molecule that is permanently bound to the E75 ligand binding pocket [37]; the several cases of structural ligands that are small molecules, often fatty acids, that are bound in the ligand binding pockets of USP and HNF4 [38, 39]. To finish on this rapid panorama of unusual binding modes one has to recall that classical receptors such as estrogen receptors are in fact promiscuous, since they recognize a large number of exogenous compounds (the endocrine disruptors) that can regulate their activity in very subtle ways. If some of these compounds are indeed artificial (BPA, DDT), others occur naturally (phytoestrogens) suggesting that may have been part of the ancestral regulatory system of NR activity by food. The historic idea of highly selective ligands controlling NR activities is now inconsistent with the fact that several endogenous ligands of estrogen receptors exist: 5α-androstane-3β,17β-diol is a natural agonist of ERβ [40] and 17OH-cholesterol, a naturally occurring steroid compound, is an endogenous antagonist of ERβ, [41].

All these data suggest that, at their origin, NRs were probably not hormonal receptors with high affinity for very specific compounds. Rather this is a feature that was acquired later during evolution. We propose that NRs instead act as a sensor, by interacting with a wide variety of compounds, to transfer, as transcriptional activity, subtle metabolic balances in the respective amounts of various compounds. Viewed in this manner, there is a continuum between classical hormones, endogenous regulators, exogenous regulators including pharmacological ligands, food derivatives, endocrine disruptors and even structural ligands that are permanently bound to the receptors.

2.6. EVOLUTION OF LIGAND BINDING

There are some well-known examples about refinement of the ligand-binding activity within closely related receptors. For the RAR/NR1B, which are retinoic-acid receptors in vertebrates, it was shown that the ability to bind slightly different molecules was acquired through definite mutations in the LBP [42].

The RXR/USP receptor, that binds 9cis-RA in metazoans, underwent at least two major shifts during its evolution in insects. First, a loss of the ligand-binding ability, after the split between grasshopers (where USP can bind 9cis-RA) and more derived insects, such as coleoptera, hemiptera and mecopterida. This first shift seems to be due only to punctual mutations that prevent the formation of a ligand binding pocket. Secondly, during the emergence of the mecopterida clade, a new and large ligand-binding pocket appeared, allowing the binding of a structural ligand. Insertions in the LBD of USP were probably responsible of this second evolutionary shift [43]. Therefore, all the possible situations regarding ligand binding (classical ligand, structural ligand, real orphan) occur for USP-RXR. Currently, it is the only NR exhibiting such plasticity.

The careful analysis of the ligand-binding specificity of PXR in mammals also provides an example of shifts in NR ligand-binding ability. In vitro and cell culture binding assays between various vertebrate species showed that the biliary salt

receptor PXR underwent a broadening of its specificity, being able to bind a various set of androstanes, pregnanes, C27 bile alcohol sulfates and some xenobiotics in the common vertebrate ancestor, whereas its specificity is restricted to C24 bile acids in amniotes [44, 45].

Despite numerous examples of transitions from one ligand to another, the original acquisition of ligand-binding ability remains unknown. The observation, that there are no obvious correlations between known ligand specificity and phylogenetic relationships between the receptors (see Figure 2.3a), led to the proposition that the evolution of the ligand-binding specificity of NRs involved several independent gains and losses of ligand-binding ability ([9], reviewed in [46, 47]). According to this notion, the LBD, which is conserved in all NRs, irrespective of the fact that they have or not a ligand, should rather be seen as an allosteric domain controlling the activity of the transcription factor. It is precisely because this conformational change

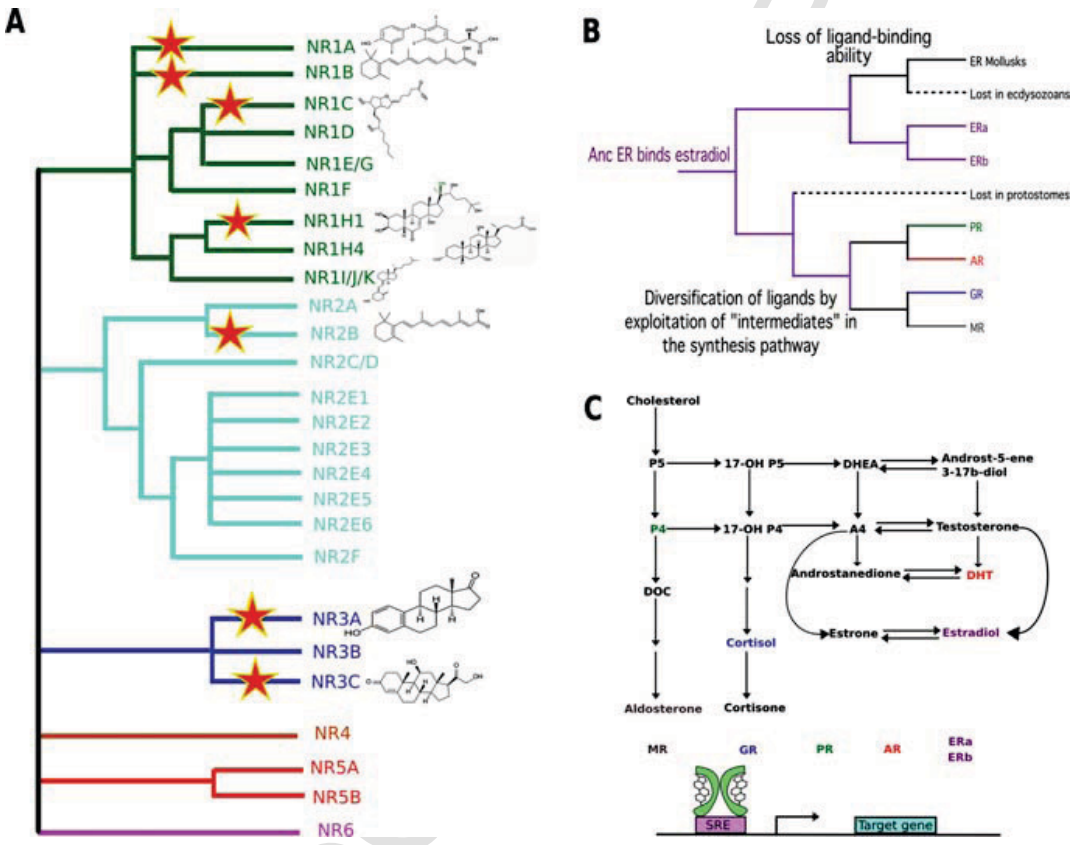


Figure 2.3. Different models about the evolution of ligand-binding ability. **a**: a general model for the whole family, independent acquisition of ligand-binding ability. Ligand for some human receptors are indicated on the tree. The lack for obvious correlations between known ligand specificity and phylogenetic relationships between the receptors led to the proposition that the evolution of the ligand-binding specificity of NRs involved several independent gains, that are here indicated by red stars **b**: a model for NR3A and NR3C. Gene losses, duplications and shifts in ligand-binding ability are indicated on a phylogenetic tree, with colors refering to ligands indicated in **c**. **c**: steroid biosynthesis pathway in human. Ligand and their receptors are indicated with the same colour

is conserved that the LBD sequence is conserved. It is interesting to note that several reports indicate that the conformational change can be triggered by several process in addition to ligand binding [48]. Indeed, phosphorylation or protein-protein interactions can induce conformational changes. Thus, ligand binding appears as just one possible trigger.

An hypothesis, based on reconstruction of ancestral sequences at internal nodes of the evolutionary tree and functional characterization of the 'resurrected' receptor suggests that the ancestral steroid receptor was liganded and that orphan receptors secondarily lost the ability to bind a ligand [49, 50]. According to the ligand exploitation model, the terminal product in a biosynthetic pathway is the first compound for which a receptor evolves; selection for this hormone also selects for the synthesis of intermediates (see Figure 2.3b, c), and duplicated receptors then evolved affinity for these intermediates. This model accounts for the divergence observed in ligand specificity of the steroid receptors, namely AR/NR3C4, GR/NR3C1, MR/NR3C2, PR/NR3C3 and ERs/NR3A [50, 51]. It also suggests that ligands for some 'orphan' receptors may be found among intermediates in the synthesis of ligands for evolutionary related receptors. The ligand exploitation model considers that 17β-estradiol (E2) was the ligand of the ancestral receptor and that E2 was present and active in a wide range of metazoans before the diversification of steroid receptors [50]. However this axiom is not supported by the current data (see for example a discussion in [52] and [53]). There are clear evidences that different steroids are synthesized in vertebrates, insects and nematodes. For example, moulting is controlled by dafachronic acids in *C. elegans* [54] and by ecdysteroids (ecdysone and related compounds) in insects [55]. In humans, the main active steroids are dihydroxytestosterone (DHT), progesterone (P4), cortisol and aldosterone that bind the four members of the NR3C group (MR, GR, AR, PR), and E2 that binds to ERs (Figure 2.3b). Even within vertebrates, there are some variations in the identity of active steroid hormones. For example, in teleosts, there are two different active androgens: DHT and 11-ketotestosterone (11KT) [56], while aldosterone is not present [57]. Strikingly, despite these known variations of steroids identity among vertebrates, many authors have searched for the presence and putative roles of 'human'-type steroids such as estradiol or progesterone, in all metazoan groups. It is important to realize that, to date, none of the biochemical evidences for the presence of vertebrate steroids in lophotrochozoans and cnidarians has been substantiated by cloning and biochemical characterization of enzymes responsible of their biosynthesis. Moreover, many of the techniques that were used to detect endocrine activity are prone to artefacts and misidentification [52, 53]. Thus, the identity of steroids present in non-model 'invertebrates' is still an open and important question. This evolutionary variability in ligands may be found for other hormonal systems. Indeed, the recent characterization of the thyroid hormone receptor signalling in amphioxus has shown that the ligand of amphioxus TR is not T3 itself, but TRIAC, a derivative of T3 [58, 59].

A way to reconcile the different views could be to suppose that the ancestral receptor was not an orphan but rather a sensor that was able to bind with low affinity a

wide range of molecules, probably provided by food [60], and that ligand binding specificity with high affinity evolved sometimes through selection of synthesis pathways for molecules that had a strong positive effect on animal physiology, and thus became endogenous synthesized hormones. This view is consistent with the fact that low affinity ligands are now known for a large set of receptors, which were previously thought to be orphans but are in fact sensors (e.g. NR2A/HNF4 in mammals [61]). Their abundance makes the hypothesis of an ancestral sensor more parsimonious than previous speculations (Figure 2.3a) that only distinguished between high affinity ligand binding receptors and orphans with no ligand binding ability at all. Some evolutionary studies at a subfamily scale [62] also reinforce the view that it may be a continuum between true orphan, sensors able to bind many molecules with low specificity and true endocrine receptors, which bind specifically one signalling molecule only.

2.7. CONCLUSION: EVOLUTION AS A REFLECTION FRAME TO UNDERSTAND NRs

The evolutionary story of NR is far from being fully elucidated but important recent progress has occurred leading to a radical shift in our view of NR signalling in recent years. NRs appear as very dynamic at the evolutionary level, being able to become adapted to a wide variety of physiological, metabolic and developmental roles. These molecules are thus now very promising evolutionary models. Striking conclusions from recent studies are that (i) non-usual genetic models such as lophotrochozoans and cnidarians continue to reveal provocative genomic and functional insight (ii) there is no obligatory co-linearity, at an evolutionary scale, between a given receptor and ligand, and (iii) physiological roles of NRs cannot be fully understood without an integrative view, taking into account the genetic environment in which receptors are evolving [63].

REFERENCES

1. Laudet, V., Hänni, C., Coll, J., Catzeflis, F., and Stéhelin, D. (1992). Evolution of the nuclear receptor gene superfamily. *EMBO J* 11(3), 1003–1013.
2. Amero, S. A., Kretsinger, R. H., Moncrief, N. D., Yamamoto, K. R., and Pearson, W. R. (1992). The origin of nuclear receptor proteins: A single precursor distinct from other transcription factors. *Mol Endocrinol* 6(1), 3–7.
3. Lewis, M. (2005). The *lac* repressor. C R Biol 328(6):521–548, DOI10.1016/j.crvi.2005.04.004.
4. Nuclear Receptors Nomenclature Committee (1999). A unified nomenclature system for the nuclear receptor superfamily. *Cell* 97(2), 161–163.
5. Brelivet, Y., Kammerer, S., Rochel, N., Poch, O., and Moras, D. (2004). Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level. *EMBO Rep* 5(4), 423–429, DOI10.1038/sj.embor.7400119.
6. Arnosti, D. N., Gray, S., Barolo, S., Zhou, J., and Levine, M. (1996). The gap protein knirps mediates both quenching and direct repression in the *Drosophila* embryo. *EMBO J* 15(14), 3659–3666.

7. Båvner, A., Sanyal, S., Gustafsson, J. A., and Treuter, E. (2005). Transcriptional corepression by SHP: Molecular mechanisms and physiological consequences. *Trends Endocrinol Metab* 16(10), 478–488, DOI10.1016/j.tem.2005.10.005.
8. Miyata, T. and Suga, H. (2001). Divergence pattern of animal gene families and relationship with the cambrian explosion. *Bioessays* 23(11), 1018–1027, DOI10.1002/bies.1147.
9. Escriva, H., Safi, R., Hänni, C., Langlois, M. C., Saumitou-Laprade, P., Stehelin, D., Capron, A., Pierce, R., and Laudet, V. (1997). Ligand binding was acquired during evolution of nuclear receptors. *Proc Natl Acad Sci U S A* 94(13), 6803–6808.
10. Bertrand, S., Brunet, F. G., Escriva, H., Parmentier, G., Laudet, V., and Robinson-Rechavi, M. (2004). Evolutionary genomics of nuclear receptors: From twenty-five ancestral genes to derived endocrine systems. *Mol Biol Evol* 21(10), 1923–1937, DOI10.1093/molbev/msh200.
11. Bertrand, S., Thisse, B., Tavares, R., Sachs, L., Chaumot, A., Bardet, P. L., Escrivà, H., Duffraisse, M., Marchand, O., Safi, R., Thisse, C., and Laudet, V. (2007). Unexpected novel relational links uncovered by extensive developmental profiling of nuclear receptor expression. *PLoS Genet* 3(11), e188, DOI10.1371/journal.pgen.0030188.
12. Schubert, M., Brunet, F., Paris, M., Bertrand, S., Benoit, G., and Laudet, V. (2008). Nuclear hormone receptor signaling in amphioxus. *Dev Genes Evol*, DOI10.1007/s00427-008-0251-y.
13. Sluder, A. E. and Maina, C. V. (2001). Nuclear receptors in nematodes: Themes and variations. *Trends Genet* 17(4), 206–213.
14. Robinson-Rechavi, M., Maina, C. V., Gissendanner, C. R., Laudet, V., and Sluder, A. (2005). Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J Mol Evol* 60(5), 577–586, DOI10.1007/s00239-004-0175-8.
15. Yates, R. A., Tuan, R. S., Shepley, K. J., and Unnasch, T. R. (1995). Characterization of genes encoding members of the nuclear hormone receptor superfamily from *Onchocerca volvulus*. *Mol Biochem Parasitol* 70(1–2), 19–31.
16. Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L., Guiliano, D. B., Miranda-Saavedra, D., Angiuoli, S. V., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N. M., Wortman, J. R., Feldblyum, T., Tallon, L., Schatz, M., Shumway, M., Koo, H., Salzberg, S. L., Schobel, S., Pertea, M., Pop, M., White, O., Barton, G. J., Carlow, C. K. S., Crawford, M. J., Daub, J., Dimmic, M. W., Estes, C. F., Foster, J. M., Ganatra, M., Gregory, W. F., Johnson, N. M., Jin, J., Komuniecki, R., Korf, I., Kumar, S., Laney, S., Li, B. W., Li, W., Lindblom, T. H., Lustigman, S., Ma, D., Maina, C. V., Martin, D. M. A., McCarter, J. P., McReynolds, L., Mitreva, M., Nutman, T. B., Parkinson, J., Peregrín-Alvarez, J. M., Poole, C., Ren, Q., Saunders, L., Sluder, A. E., Smith, K., Stanke, M., Unnasch, T. R., Ware, J., Wei, A. D., Weil, G., Williams, D. J., Zhang, Y., Williams, S. A., Fraser-Liggett, C., Slatko, B., Blaxter, M. L., and Scott, A. L. (2007). Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* 317(5845), 1756–1760, DOI10.1126/science.1145406.
17. Abad, P., Gouzy, J., Aury, J. M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V. C., Caillaud, M. C., Coutinho, P. M., Dasilva, C., Luca, F. D., Deau, F., Esquibet, M., Flutre, T., Goldstone, J. V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T. R., Markov, G. V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T. J., Blaxter, M., Bleve-Zacheo, T., Davis, E. L., Ewbank, J. J., Favery, B., Grenier, E., Henrissat, B., Jones, J. T., Laudet, V., Maule, A. G., Quesneville, H., Rosso, M. N., Schiex, T., Smant, G., Weissenbach, J., and Wincker, P. (2008). Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol* 26(8), 909–915, DOI10.1038/nbt.1482.
18. Dieterich, C., Clifton, S. W., Schuster, L. N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte, H., Yang, S. P., Wilson, R. K., and Sommer, R. J. (2008). The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* 40(10), 1193–1198, DOI10.1038/ng.227.

19. Wu, W., Niles, E. G., Hirai, H., and LoVerde, P. T. (2007). Evolution of a novel subfamily of nuclear receptors with members that each contain two DNA binding domains. *BMC Evol Biol* 7, 27, DOI10.1186/1471-2148-7-27.
20. King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J. B., Morris, A., Nichols, S., Richter, D. J., Salamov, A., Sequencing, J. G. I., Bork, P., Lim, W. A., Manning, G., Miller, W. T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I. V., and Rokhsar, D. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451(7180), 783–788, DOI10.1038/nature06617.
21. Phelps, C., Gburcik, V., Suslova, E., Dudek, P., Forafonov, F., Bot, N., MacLean, M., Fagan, R. J., and Picard, D. (2006). Fungi and animals may share a common ancestor to nuclear receptors. *Proc Natl Acad Sci U S A* 103(18), 7077–7081, DOI10.1073/pnas.0510080103.
22. Thakur, J. K., Arthanari, H., Yang, F., Pan, S. J., Fan, X., Breger, J., Frueh, D. P., Gulshan, K., Li, D. K., Mylonakis, E., Struhl, K., Moye-Rowley, W. S., Cormack, B. P., Wagner, G., and Näär, A. M. (2008). A nuclear receptor-like pathway regulating multidrug resistance in fungi. *Nature* 452(7187), 604–609, DOI10.1038/nature06836.
23. Furness, S. G. B., Lees, M. J., and Whitelaw, M. L. (2007). The dioxin (aryl hydrocarbon) receptor as a model for adaptive responses of bHLH/PAS transcription factors. FEBS Lett 581(19):3616–3625, DOI10.1016/j. febslet. 200704.011.
24. Murphy, K. A., Quadro, L., and White, L. A. (2007). The intersection between the aryl hydrocarbon receptor (AHR)- and retinoic acid-signaling pathways. *Vitam Horm* 75, 33–67, DOI10.1016/S0083-6729(06)75002-6.
25. Ohtake, F., Baba, A., Takada, I., Okada, M., Iwasaki, K., Miki, H., Takahashi, S., Kouzmenko, A., Nohara, K., Chiba, T., Fujii-Kuriyama, Y., and Kato, S. (2007). Dioxin receptor is a ligand-dependent E3 ubiquitin ligase. *Nature* 446(7135), 562–566, DOI10.1038/nature05683.
26. Konopova, B. and Jindra, M. (2007). Juvenile hormone resistance gene methoprene-tolerant controls entry into metamorphosis in the beetle *Tribolium castaneum*. *Proc Natl Acad Sci U S A* 104(25), 10,488–10,493, DOI10.1073/pnas.0703719104.
27. Soccio, R. E. and Breslow, J. L. (2003). Star-related lipid transfer (start) proteins: Mediators of intracellular lipid metabolism. *J Biol Chem* 278(25), 22,183–22,186, DOI10.1074/jbc.R300003200.
28. Tsujishita, Y. and Hurley, J. H. (2000). Structure and lipid transport mechanism of a star-related domain. *Nat Struct Biol* 7(5), 408–414, DOI10.1038/75192.
29. Schrick, K., Nguyen, D., Karlowski, W. M., and Mayer, K. F. X. (2004). Start lipid/sterol-binding domains are amplified in plants and are predominantly associated with homeodomain transcription factors. *Genome Biol* 5(6), R41, DOI10.1186/gb-2004-5-6-r41.
30. Kushiro, T., Nambara, E., and McCourt, P. (2003). Hormone evolution: The key to signalling. *Nature* 422(6928), 122, DOI10.1038/422122a.
31. Beckett, D. (2001). Regulated assembly of transcription factors and control of transcription initiation. *J Mol Biol* 314(3), 335–352, DOI10.1006/jmbi.2001.5134.
32. Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3, 318–356.
33. Koelle, M. R., Talbot, W. S., Segraves, W. A., Bender, M. T., Cherbas, P., and Hogness, D. S. (1991). The *Drosophila EcR* gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell* 67(1), 59–77.
34. Calléja, C., Messaddeq, N., Chapellier, B., Yang, H., Krezel, W., Li, M., Metzger, D., Mascrez, B., Ohta, K., Kagechika, H., Endo, Y., Mark, M., Ghyselinck, N. B., and Chambon, P. (2006). Genetic and pharmacological evidence that a retinoic acid cannot be the RXR-activating ligand in mouse epidermis keratinocytes. *Genes Dev* 20(11), 1525–1538, DOI10.1101/gad. 368706.
35. Rocchi, S., Picard, F., Vamecq, J., Gelman, L., Potier, N., Zeyer, D., Dubuquoy, L., Bac, P., Champy, M. F., Plunket, K. D., Leesnitzer, L. M., Blanchard, S. G., Desreumaux, P., Moras, D., Renaud, J. P.,

and Auwerx, J. (2001). A unique PPARgamma ligand with potent insulin-sensitizing yet weak adipogenic activity. *Mol Cell* 8(4), 737–747.

36. Wang, Y., Chirgadze, N. Y., Briggs, S. L., Khan, S., Jensen, E. V., and Burris, T. P. (2006). A second binding site for hydroxytamoxifen within the coactivator-binding groove of estrogen receptor beta. *Proc Natl Acad Sci U S A* 103(26), 9908–9911.
37. Thummel, C. S. (2005). Powered by gas–a ligand for a fruit fly nuclear receptor. Cell 122(2): 151–153, DOI10.1016/j.cell.2005.07.007.
38. Clayton, G. M., Peak-Chew, S. Y., Evans, R. M., and Schwabe, J. W. (2001). The structure of the ultraspiracle ligand-binding domain reveals a nuclear receptor locked in an inactive conformation. *Proc Natl Acad Sci U S A* 98(4), 1549–1554, DOI10.1073/pnas.041611298.
39. Dhe-Paganon, S., Duda, K., Iwamoto, M., Chi, Y. I., and Shoelson, S. E. (2002). Crystal structure of the HNF4 alpha ligand binding domain in complex with endogenous fatty acid ligand. *J Biol Chem* 277(41), 37,973–37,976, DOI10.1074/jbc.C200420200.
40. Weihua, Z., Lathe, R., Warner, M., and Gustafsson, J. A. (2002). An endocrine pathway in the prostate, ERbeta, AR, 5alpha-androstane-3beta,17beta-diol, and CYP7B1, regulates prostate growth. *Proc Natl Acad Sci U S A* 99(21), 13,589–13,594, DOI10.1073/pnas.162477299.
41. Umetani, M., Domoto, H., Gormley, A. K., Yuhanna, I. S., Cummins, C. L., Javitt, N. B., Korach, K. S., Shaul, P. W., and Mangelsdorf, D. J. (2007). 27-hydroxycholesterol is an endogenous serm that inhibits the cardiovascular effects of estrogen. *Nat Med* 13(10), 1185–1192, DOI10.1038/nm1641.
42. Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., and Laudet, V. (2006). Neofunctionalization in vertebrates: The example of retinoic acid receptors. *PLoS Genet* 2(7), e102, DOI10.1371/journal.pgen.0020102.
43. Iwema, T., Billas, I. M. L., Beck, Y., Bonneton, F., Nierengarten, H., Chaumot, A., Richards, G., Laudet, V., and Moras, D. (2007). Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. *EMBO J* 26(16), 3770–3782, DOI10.1038/sj.emboj.7601810.
44. Krasowski, M. D., Yasuda, K., Hagey, L. R., and Schuetz, E. G. (2005). Evolution of the pregnane X receptor: Adaptation to cross-species differences in biliary bile salts. *Mol Endocrinol* 19(7), 1720–1739, DOI10.1210/me.2004-0427.
45. Reschly, E. J., Bainy, A. C. D., Mattos, J. J., Hagey, L. R., Bahary, N., Mada, S. R., Ou, J., Venkataramanan, R., and Krasowski, M. D. (2007). Functional evolution of the vitamin D and pregnane X receptors. *BMC Evol Biol* 7, 222, DOI10.1186/1471-2148-7-222.
46. Escriva, H., Delaunay, F., and Laudet, V. (2000). Ligand binding and nuclear receptor evolution. *Bioessays* 22(8), 717–727, DOI3.0.CO;2-I.
47. Baker, M. E. (2003). Evolution of adrenal and sex steroid action in vertebrates: A ligand-based mechanism for complexity. *Bioessays* 25(4), 396–400, DOI10.1002/bies.10252.
48. Popov, V. M., Wang, C., Shirley, L. A., Rosenberg, A., Li, S., Nevalainen, M., Fu, M., and Pestell, R. G. (2007). The functional significance of nuclear receptor acetylation. Steroids 72(2):221–230, DOI10.1016/j.steroids.2006.12.001.
49. Thornton, J. W. (2001). Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci U S A* 98(10), 5671–5676, DOI10.1073/pnas.091553298.
50. Thornton, J. W., Need, E., and Crews, D. (2003). Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. *Science* 301(5640), 1714–1717, DOI10.1126/science.1086185.
51. Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312(5770), 97–101, DOI10.1126/science.1123348.
52. Lafont, R. and Mathieu, M. (2007). Steroids in aquatic invertebrates. *Ecotoxicology* 16(1), 109–130, DOI10.1007/s10646-006-0113-1.
53. Markov, G. V., Paris, M., Bertrand, S., and Laudet, V. (2008). The evolution of the ligand/receptor couple: A long road from comparative endocrinology to comparative genomics. *Mol Cell Endocrinol* 293(1–2), 5–16, DOI10.1016/j.mce.2008.06.011.
54. Rottiers, V., Motola, D. L., Gerisch, B., Cummins, C. L., Nishiwaki, K., Mangelsdorf, D. J., and Antebi, A. (2006). Hormonal control of *C. elegans* dauer formation and life span by a rieske-like oxygenase. Dev Cell 10(4):473–482, DOI10.1016/j.devcel.2006.02.008.

55. Rewitz, K. F., Rybczynski, R., Warren, J. T., and Gilbert, L. I. (2006). The halloween genes code for cytochrome P450 enzymes mediating synthesis of the insect moulting hormone. *Biochem Soc Trans* 34(Pt 6), 1256–1260, DOI10.1042/BST0341256.
56. Lokman, P. M., Harris, B., Kusakabe, M., Kime, D. E., Schulz, R. W., Adachi, S., and Young, G. (2002). 11-oxygenated androgens in female teleosts: Prevalence, abundance, and life history implications. *Gen Comp Endocrinol* 129(1), 1–12.
57. Bury, N. R., Sturm, A., Rouzic, P. L., Lethimonier, C., Ducouret, B., Guiguen, Y., Robinson-Rechavi, M., Laudet, V., Rafestin-Oblin, M. E., and Prunet, P. (2003). Evidence for two distinct functional glucocorticoid receptors in teleost fish. *J Mol Endocrinol* 31(1), 141–156.
58. Paris, M., Escriva, H., Schubert, M., Brunet, F., Brtko, J., Ciesielski, F., Roecklin, D., Vivat-Hannah, V., Jamin, E. L., Cravedi, J. P., Scanlan, T. S., Renaud, J. P., Holland, N. D., and Laudet, V. (2008). Amphioxus postembryonic development reveals the homology of chordate metamorphosis. *Curr Biol* 18(11), 825–830, DOI10.1016/j.cub.2008.04.078.
59. Paris, M. and Laudet, V. (2008). The history of a developmental stage: Metamorphosis in chordates. *Genesis* 46(11), 657-672, DOI10.1002/dvg.20443.
60. Ben-Shlomo, I. and Hsueh, A. J. W. (2005). Three's company: Two or more unrelated receptors pair with the same ligand. *Mol Endocrinol* 19(5), 1097–1109, DOI10.1210/me.2004-0451.
61. Wisely, G. B., Miller, A. B., Davis, R. G., Thornquest, A. D., Johnson, R., Spitzer, T., Sefler, A., Shearer, B., Moore, J. T., Miller, A. B., Willson, T. M., and Williams, S. P. (2002). Hepatocyte nuclear factor 4 is a transcription factor that constitutively binds fatty acids. *Structure* 10(9), 1225–1234.
62. Reschly, E. J. and Krasowski, M. D. (2006). Evolution and function of the NR1I nuclear hormone receptor subfamily (VDR, PXR, and CAR) with respect to metabolism of xenobiotics and endogenous compounds. *Curr Drug Metab* 7(4), 349–365.
63. Laudet, V. (1997). Evolution of the nuclear receptor superfamily: Early diversification from an ancestral orphan receptor. *J Mol Endocrinol* 19(3), 207–226.
64. Bonneton, F., Brunet, F. G., Kathirithamby, J., and Laudet, V. (2006). The rapid divergence of the ecdysone receptor is a synapomorphy for mecopterida that clarifies the strepsiptera problem. *Insect Mol Biol* 15(3), 351–362, DOI 10.1111/j.1365-2583.2006.00654.x.
65. Wu, W., Niles, E. G., El-Sayed, N., Berriman, M., and LoVerde, P. T. (2006). *Schistosoma mansoni* (Platyhelminthes, Trematoda) nuclear receptors: Sixteen new members and a novel subfamily. *Gene* 366, 303–315.

Molecular and Cellular Endocrinology 334 (2011) 21–30

Contents lists available at ScienceDirect

Molecular and Cellular Endocrinology

journal homepage: www.elsevier.com/locate/mce

Review

Origin and evolution of the ligand-binding ability of nuclear receptors

Gabriel V. Markov[a,b], Vincent Laudet[a,*]

[a] Molecular Zoology Team, Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland Lyon Sud, France
[b] UMR 7221 - Evolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, Paris, France

ARTICLE INFO

Article history:
Received 24 June 2010
Received in revised form 22 October 2010
Accepted 22 October 2010

Keywords:
Nuclear receptors
Comparative endocrinology
Gene duplication
Phylogeny
Orthology

ABSTRACT

The origin of the ligand-binding ability of nuclear receptors is still a matter of discussion. Current opposing models are the early evolution of an ancestral receptor that would bind a specific ligand with high affinity and the early evolution of an ancestral orphan that was a constitutive transcription factor. Here we review the arguments in favour or against these two hypotheses, and we discuss an alternative possibility that the ancestor was a ligand sensor, which would be able to explain the apparently contradictory data generated in previous models for the evolution of ligand binding in nuclear receptors.

Contents

Nuclear receptors (NRs) are classically defined as ligand-activated transcription factors that allow the regulation of target genes by small lipophilic molecules such as hormones (thyroid hormones, steroids), morphogens (retinoic acid) or dietary components (fatty acids). Although built with a similar organization, NRs are nevertheless regulated by a wide diversity of compounds and are implicated in a tremendous diversity of physiological and metabolic processes (de Lera et al., 2007; Germain et al., 2006; Huang et al., 2010). The origins and ancestral function of nuclear receptors, especially in terms of ligand-binding ability, have been a matter of debate for a long time. This question is crucial to understand what could be the common features shared by all NRs and to be able to distinguish, in the action of NRs, what is a remnant of phylogenetic constraint, and what has a specific adaptative value. Indeed, along with the bHLH-PAS family (Hahn, 2002), the NRs are the only transcription factors that are able to make a direct link between gene regulation and the metabolic environment, and they may have played an important role in the diversification of animals as multicellular heterotrophs. The debate on the origin of ligand-binding ability of NR began at time when there was a clear-cut dichotomy between hormonal receptors, that bind a ligand with nanomolar affinity, and orphans receptors with no known ligands. However, structural data during the past decade have significantly refined our understanding of NR binding affinities. Additional to the two classical categories it became clear that there also exists

* Corresponding author at: Molecular Zoology, Institut de Génomique Fonctionnelle de Lyon, UMR 5242 du CNRS, INRA, IFR128 BioSciences Lyon-Gerland, Université de Lyon, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France. Tel.: +33 4 72 72 81 90; fax: +33 4 72 72 80 80.
E-mail address: Vincent.Laudet@ens-lyon.fr (V. Laudet).

0303-7207/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.mce.2010.10.017

receptors with micromolar affinity for a variety of components, that were termed "nutritional sensors", such as PPAR (NR1C), LXR (NR1H2/3), FXR (NR1H4/5). It was also found that some receptors have their pockets filled either by side-chain amino-acids, such as Nurr1 (NR4A2) or HR38 (NR4A4) or by an hydrophobic molecule that does not trigger transcriptional changes, for which the term "structural ligand" was coined, such as HNF4α (NR2A1) or mecopteridan USP (NR2B4) (Benoit et al., 2004; Sladek, this issue). Even the chemical nature of the ligand was expanded with the discovery that hemes are also NR ligands (Burris, 2008). Now is a good time to incorporate such data in discussions about the origin and evolution of ligand-binding ability in NRs.

Here we reassess this question by reviewing the distribution of ligand binding abilities within the NR family, with a focus on well studied subfamilies, in which changes in ligand binding specificity allow us to draw hypotheses on the mechanisms that would explain the appearance of the first NRs. Finally we address the question of ligand synthesis evolution, which brings some complementary clarifying information to the debate.

1. Nuclear receptor phylogeny and the origin of the ancestral orphan hypothesis

The way of regarding the relationships between NRs and their ligands has been historically biased by the fact that the first identified ligands were mammalian steroid and thyroid hormones, hence the name often given to the family: the steroid/thyroid hormone receptor family (Evans, 1988). When EcR (NR1H1), the receptor for the insect hormone ecdysone was identified in *Drosophila*, it was clear to everyone that NRs were high affinity receptors (at the nanomolar range) for hormones in all animals (Koelle et al., 1991). The question then arose of the origin of the ligand-binding ability. Based on the observation that two kinds of major NR ligands, steroids and retinoids, are products of terpenoid metabolism, the first hypothesis was that the ancestral receptor would have been liganded by a terpenoid molecule (Moore, 1990). At that time, the only known non-terpenoid NR ligands were thyroid hormones, and there were many candidate ligands within terpenoid molecules with signalling roles in various eukaryotes, such as juvenile hormone in insects or abscissic acid and gibberellins in plants. In this context, the first orphan receptors that were cloned were mainly considered as receptors waiting for a yet unknown high-affinity ligand (Giguère et al., 1988). When the possibility was raised that orphans could be constitutively active, the hypothesis that ligands could be derived from intracellular metabolism, was preferred (O'Malley, 1989).

With the first phylogenies of the family, it became clear that all NRs share a common ancestor, and it became possible to speculate on the ancestral state of the first nuclear receptor (Amero et al., 1992; Laudet et al., 1992). This led to the proposal that the evolution of the ligand binding specificity of NRs involved several independent gains and losses of ligand-binding ability from an ancestral orphan (Fig. 1, Escriva et al., 1997, reviewed in Escriva et al., 2000, and Baker, 2003). This view was supported by three types of arguments.

First, ***orthologs*** of classically liganded vertebrate receptors, such as TR (NR1A), RAR (NR1B) and steroid receptors from the NR3 family (ER, GR, MR, PR, AR) were not identified outside vertebrates, suggesting a late appearance of liganded receptors during animal evolution.[1] It was later shown that ***homologs*** of these receptors are present in some mollusks and platyhelminthes, but to date it remains true that they are not present outside bilaterians. This

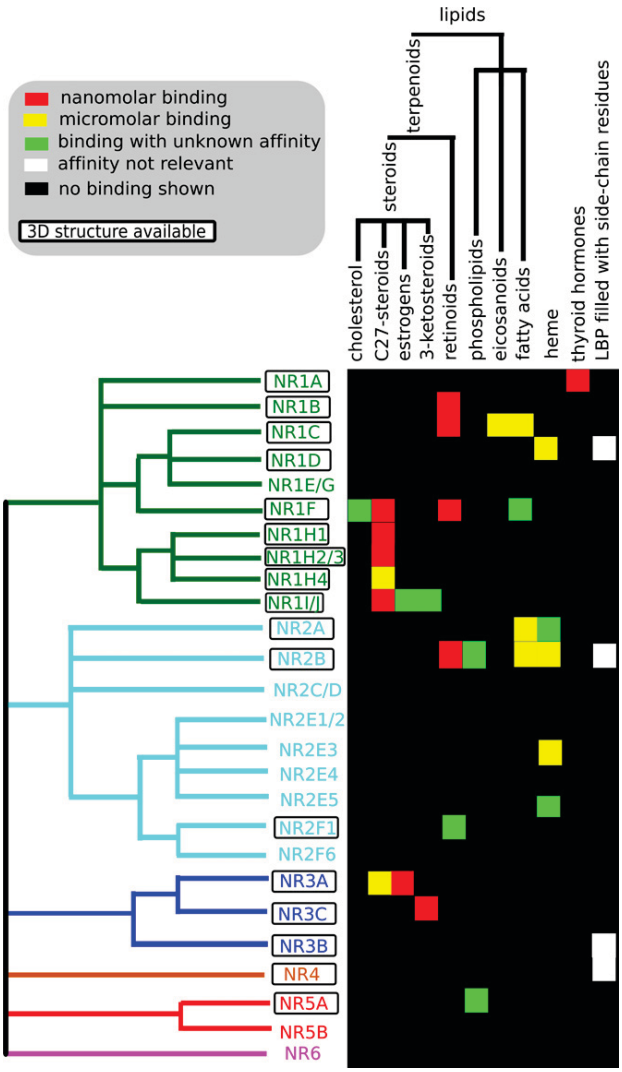


Fig. 1. Phylogenetic distribution of NR ligands. A phylogeny of putative bilaterian NRs, adapted from Bertrand et al., 2004. The presence of a ligand in front on the receptor does not mean that all the orthologs bind it, but only that the binding was shown for some of them. Boxed receptors are those for which at least one crystal structure is available. The binding ability for each type of ligand is indicated by the colour code. Red, nanomolar affinity; yellow, micromolar affinity; green, binding without data on the affinity, or affinity criterion not relevant (for pockets filled with side-chain amino-acids).

means that the early steps of NR diversification, leading to the common set of 25 bilaterian NRs (Bertrand et al., 2004) may have taken place in a context where classical hormone receptors did not exist, as animals living at that time did not have an internal circulatory system linking differentiated organs. However, this does not necessary imply that the first receptors were true orphans. Among the NRs that exist in cnidarians, there are HNF4 (NR2A), which has a "structural ligand" in mammals, that may mirror the ancestral situation (Sladek, 2002; Benoit et al., 2004; Yuan et al., 2009), and also RXR (NR2B), which can be viewed as a sensor (discussed below) due to the broad diversity of its substrates (Mic et al., 2003; Calléja et al., 2006).

Second, the hypothesis that orphan receptors evolved early also raises the questions of the mechanism that would allow them to be activated, and what structural constraints allow for the ligand binding domain (LBD) to be conserved in orphan receptors. In 1997 came the first results of NR regulation through conformational changes

[1] *Note*: Evolutionary notions that may need further explanation are indicated in ***bold italic*** at their first occurrences and are defined in Box 1.

Box 1: Definition of some evolutionary concepts used in the text

Convergence
Two similar structures are convergent when their similarity is not due to common ancestry, but due to independent formation from different structures.

Homology
Two structures are homologs when they share common ancestry. For genes – and by extension for proteins – there are two major kinds of homologs: orthologs and paralogs.

Long branch attraction
This is a phenomenon in phylogenetic analyses when rapidly evolving lineages are inferred to be closely related, regardless of their true evolutionary relationships.

Orthology
Two genes are orthologs when they share common ancestry after a speciation event, e.g. TRα from mouse and TRα from rat.

Paralogy
Two genes are paralogs when they share common ancestry after a gene duplication event, e.g. TRα from mouse, TRβ from mouse, and unduplicated TR from amphioxus.

Parsimony
An hypothesis is parsimonious when it uses the simplest way to explain an observation. For example, concerning gene family evolution, a scenario that implies 2 gene duplications and one loss is more parsimonious that a scenario that implies one duplication and four losses.

in a ligand-independent way, for example due to phosphorylation (Rochette-Egly, 2003), or other types of post-translational modifications. This provided an explanatory mechanism for the regulation of orphan receptors by something other than ligand binding, which could explain the structural conservation of the LBD in absence of ligand binding. This fact has gained increasing support: it is clear that the ligand is just one possible trigger for the conformational change or more appropriately termed the allosteric transition (Faus and Haendler, 2006).

The third argument for the ancestral orphan receptor view was that there seemed to be nothing in common between the synthesis pathways of diverse ligands such as thyroid hormones, retinoic acids and steroids. This should be now reassessed in light of our more complete understanding of the wide diversity of NR ligands (Fig. 1, and Sladek, this issue). For example, even if some of these data still need to be confirmed, retinoids can apparently bind mammalian receptors other than RAR (NR1B) and RXR (NR2B), with comparable affinity for PPARβ/δ (NR1C2) and lower affinities for the two other mammalian PPARs (NR1C), RORβ (NR1F2), or COUP-TFII (NR2F2) (reviewed in Theodosiou et al., 2010). NR2B (RXR/USP) is sensitive to a wide range of different molecules, not only retinoids but also fatty acids and phospholipids, and, in fact, retinoid binding to RXR may have no relevance *in vivo* (Mic et al., 2003; Calléja et al., 2006). PPARs bind fatty acids and their eicoisanoid derivatives. Furthermore, there is evidence for crosstalk between retinoic acid and fatty acid signalling based on alternate activation of RAR or PPARβ/δ depending on retinoic acid concentration (Schug et al., 2007). It is worth noting that retinoids are transported to cells as esters and that the dissociation of the ester produces not only retinoic acid, but also releases a fatty acid. This raises the possibility that other unknown cross-talk due to the sharing of one of many ligands may connect the different NR-signalling pathways. Rev-erbs (NR1D1/2) and their ortholog in insects, E75 (NR1D3), were shown to bind hemes, a very big molecule when compared to other ligands in mammals, as in *Drosophila* (Burris, 2008). Recent data indicate that other NRs, such as RXRα (NR2B1) in mammals (Gotoh et al., 2008) or HR51 (NR2E3) bind heme with micromolar affinity, whereas HNF4 (NR2A) and HR83 (NR2E5) bind heme with lower affinity in *Drosophila* (de Rosny et al., 2008). This suggests that heme could be a more widespread ligand than previously expected. Even historical "hormone-receptors", can be activated by a variety of ligands. For example, the Vitamin D Receptor (NR1I1) is also activated by a bile acid (Makishima et al., 2002), and 5α-androstane-3β,17β-diol appears to be a natural agonist of the estrogen receptor ERβ (NR3A2) (Weihua et al., 2002). On the other hand, many steroid hormones activate the "xenobiotic receptor" PXR (NR1I2) at a micromolar range (Ekins et al., 2008). Therefore, generally there is no exclusive pairing between ligands and receptors: one ligand can activate many receptors and one receptor can be activated by many ligands. Furthermore, a receptor can be both a sensor or a liganded receptor, depending on the context (Fig. 1).

Even if the various NR ligands are members of different chemical families from a nomenclature viewpoint, they share common physical properties, such as hydrophobicity and a volume between 250 and 550 $Å^3$. Indeed, there are other protein families that interact with the NR ligands (Fig. 2), where different ***paralogs*** bind different ligands. Retinoic acid, retinol, fatty acids, eicosanoids,

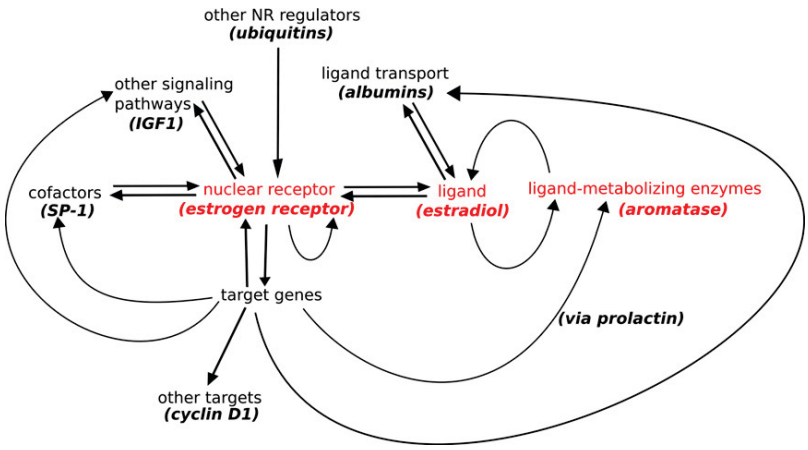


Fig. 2. The interaction network between NRs and their ligands. The basic interaction network in which NRs are involved comprises NR target genes, for which transcription is activated or inhibited by NRs, cofactors, that bind to NRs during activation or repression, other signalling pathways, that led to post-transcriptional modifications of NRs, and ligands, whose presence is controlled by ligand-metabolising enzymes, and ligand transport proteins, intracellular such as FABP or extracellular such as albumins. NRs can also act on other signalling pathways through non-genomic mechanisms, that do not involve their binding to DNA. Members of the estrogen receptor network are indicated as an example.

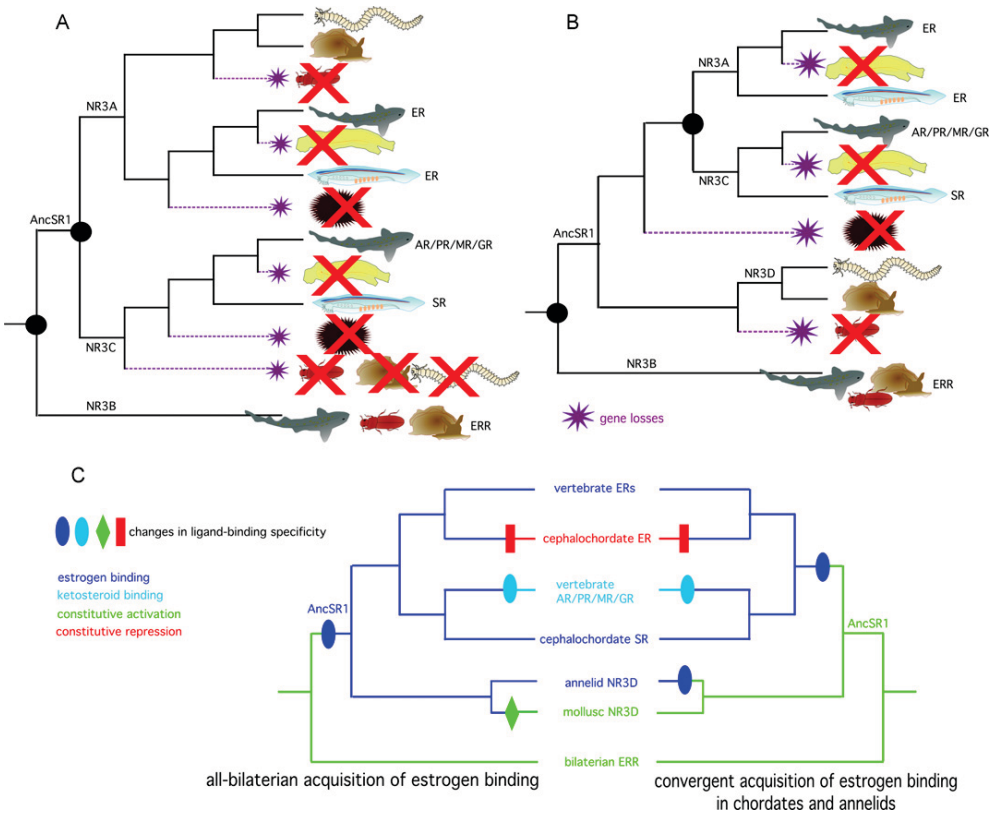


Fig. 3. Uncertainities in parsimony-based scenarios about the ligand-binding abilities of the ancestral steroid receptor (AncSR1). (A) If there was a gene duplication of AncSR1 leading to NR3A and NR3C in the common ancestor of all bilaterian animals, six independent gene losses are necessary to explain the current gene distribution. These losses are indicated by the purple stars at the end of dotted branches. (B) If NR3A and NR3C are the products of a chordate-specific gene duplication, four gene losses would be sufficient to explain the observed gene distribution. Thus, this scenario is more parsimonious than (A), but would require the renaming of lophotrochozoan "ER" as NR3D, to stress that they are not more related to vertebrate NR3A than to vertebrate NR3C. (C) Under the tree topology presented in (B) and taking into account the fact that ERR is a constitutive activator, acquisition of steroid binding in all bilaterians (on left side) or convergent acquisition in chordates and annelids (on right side) are both equiparsimonious scenarios. In both cases, constitutive repression in cephalochordate ER (red box) and a shift in ligand specificity in vertebrate AR/PR/MR/GR (light blue oval) would be acquired from a liganded receptor. On the contrary, the early acquisition of estrogen-binding in AncSR1 (dark blue oval on left) would necessitate a reversion to constitutive activity in mollusk NR3D. For discussion on the topology presented in (A), see also Eick and Thornton, this issue.

heme and bile acids are all bound by proteins of the FABP family (Zimmerman and Veerkamp, 2002), that are involved in their intracellular transport, whereas the extracellular albumins bind all kinds of NR ligands (Baker, 2002a). Short-chain dehydrogenase reductases (SDR) are also known to metabolise both retinoids and steroids (Baker, 2001). Proteins of the CYP family are involved in the metabolism of retinoids, steroids, fatty acids, eicosanoids and xenobiotics. Important substrate shifts between closely related paralogs are also well known (Brown et al., 2008). For example, CYP8A1 and CYP8B1, which are the result of a vertebrate-specific duplication, metabolise respectively, prostacyclin (an eicosanoid) and a bile-acid precursor (Thomas, 2007). These shifts in substrates may indicate that these ligands share some common properties allowing a rapid switch from one ligand to another on an evolutionary timescale.

Finally, the orphan hypothesis raises the question on how an unliganded transcription factor shifts to a ligand binding receptor. This can only be addressed by focusing on some precise case studies.

2. Origin and evolution of the steroid receptors and their implications on the ancestral NR

There are many well studied cases of variations on binding ability between similar ligands (Bridgham et al., 2006; Escriva et al., 2006; Paris et al., 2008a; Reschly et al., 2008a), and some cases of transition from a liganded receptor to an orphan (Krylova et al., 2005; Iwema et al., 2007) or transition from an orphan to a receptor with a structural ligand (Iwema et al., 2007), but these studies do not deal with the transition from an orphan to a liganded receptor. This is partly due to the fact that, except for insects and nematodes, functional data on non-vertebrate animals are still scarce, and this is further complicated by the lack of data regarding the physiological significance of some putative ligands, and by uncertainties regarding the topology of the NR trees (see the numerous polytomies in Fig. 1). The only families where there is sufficient genomic sampling and understanding of the physiological significance of the ligands are the NR3 subfamily (Fig. 3) and the NR1H/I/J group (Fig. 4), that contain the steroid receptors, making them the best proxies to address the question of ligand binding acquisition.

2.1. Hypotheses on the binding-ability of the ancestral steroid receptor in the NR3 subfamily

The N3 family contains receptors for vertebrate sex and adrenal steroids, but also some mollusk receptors that do not bind sex and adrenal steroids. However, a resurrected ancestral estrogen receptor was found to be activated by estrogens, implying that estrogen binding would have been secondarily lost in some mollusks (Thornton et al., 2003). This was followed by the further characterization of mollusk constitutive activators in this family (Keay et al., 2006), whereas secondary losses of ligand binding-ability was further documented in the rodent LRH-1, a receptor

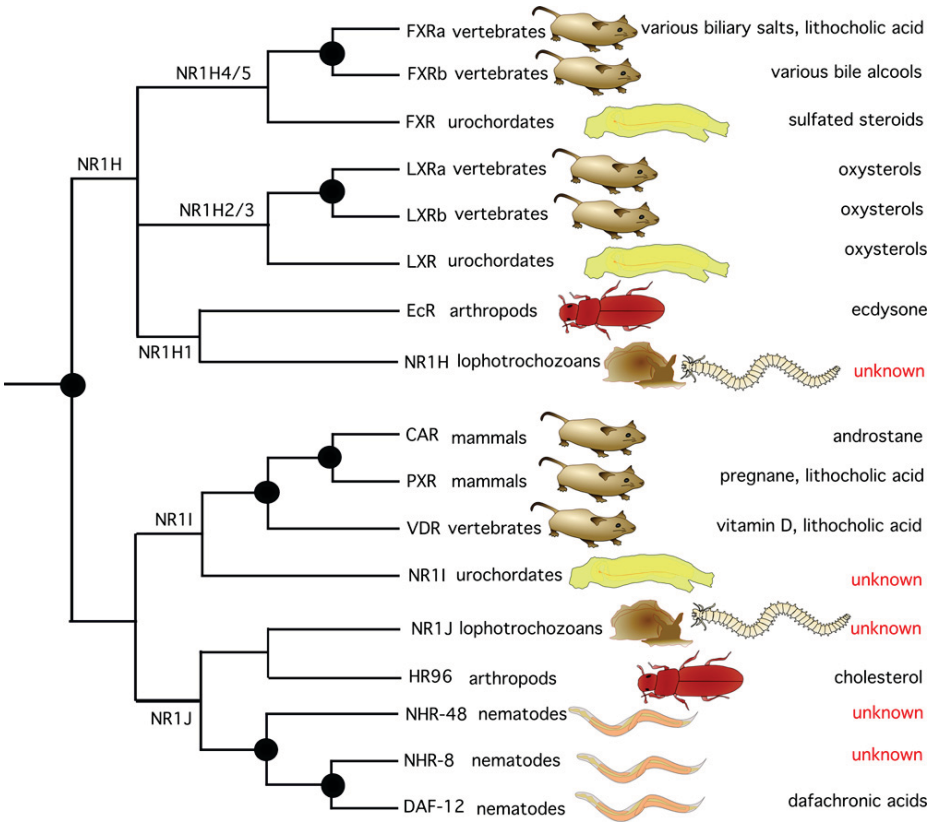


Fig. 4. Current data about steroid binding in the NR1H/I/J family. Trivial names of various NR1H/I/J members are plotted on the tree, together with indication on the species where there are present, and with information about the ligand-binding ability of the receptor. Gene duplication events are indicated with black disks. In the NR1H family, all chordate receptors bind steroids, and the NR1H of insects also binds a steroid, so there is high probability that lophotrochozoan NR1H also bind an oxydated cholesterol derivative. Similarly, since there are steroid-binding receptors in vertebrate NR1I and nematode NR1J, whereas the insect ortholog NR1J may bind cholesterol, the currently uncharacterized NR1J from lophotrochozoan would be an obvious candidate for steroid binding.

from the NR5A2 group (Krylova et al., 2005). Taken together, these data were interpreted as evidence against the ancestral orphan receptor theory, and it was proposed that, on the contrary, constitutive activation has evolved several times in parallel from a ligand-dependant nuclear receptor ancestor (Keay et al., 2006).

Hypotheses about steroid binding in the NR3 subfamily are strongly dependent on the understanding of the relationships between the various receptors (Fig. 3). The NR3 subfamily diversified specifically in bilaterians (Baker, 2008). The first duplication produced the common ancestor of bilaterian ERR (NR3B) and the common ancestor of the bilaterian steroid receptor, AncSR1 (Thornton, 2001; Thornton et al., 2003). There is also clear evidence that vertebrate ER has an ortholog in amphioxus (Paris et al., 2008b), and that another amphioxus receptor, named SR (Bridgham et al., 2008), is orthologous to the vertebrate ancestor gene that gave rise to the current GR (NR3C1), MR (NR3C2), PR (NR3C3) and AR (NR3C4).

Things become more complicated when dealing with the recently cloned NR3 in mollusks and annelids. The first analyses of mollusk sequences provided support for their grouping with the chordate ER (NR3A), and they were also named NR3A in some publications (Thornton et al., 2003; Paris et al., 2008b). This topology is shown in Fig. 3A. But the addition of two cloned annelid receptors decreased the support for this grouping (Keay and Thornton, 2009). In this paper, the authors acknowledged that they "*cannot rule out the possibility that the protostome ERs could be equally orthologous to the entire SR family*"(topology shown in Fig. 3B).

Two additional lines of evidence may be worth taking into account to discriminate between both scenarios. First, when one considers the minimal number of secondary gene losses, it appears that there would have been at least six independent losses if lophotrochozan "ERs" are orthologous to NR3A (Fig. 3A), whereas if it were orthologous to the chordate gene that gives rise to NR3A and NR3C after duplication, there would have been only four losses (Fig. 3B). This second scenario is therefore more ***parsimonious***. The second line of evidence is that vertebrate NR3C has undergone an acceleration of evolutionary rate, which has led to reconstruction artifacts at the base of NR3 phylogeny. At that time, when only mammal sequences where available, NR3C branched basally to a group containing NR3A and NR3B (Laudet, 1997). In fact, similar artifacts are also present for other NR subfamilies, even when methods that diminish the effect of ***long-branch attraction***, such as maximum-likelihood, are used. For example, a recent paper on the phylogeny of bilaterian RXR/USP (NR2B) showed nematode, platyhelminthe and mecopteridan insect sequences branching erroneously at the basis of bilateria, and this was interpreted using the same reasoning that favours the topology in Fig. 3B (Tzertzinis et al., 2010).

Whatever the true topology, this debate raises an important nomenclature issue. In such ambiguous cases, it would be preferable to give the controversial sequence a name that does not favour one or the other hypothesis. This is why we suggest that mollusk and annelid sequences, that where up to now unofficially designated as "ER" or "NR3A" should preferably be named "NR3D", as we do in Fig. 3B and C, and as it has already been done for other ambiguous cases, such as vertebrate VDR/PXR/CAR (NR1I) and insect HR96 (NR1J) (Nuclear Receptors Nomenclature Committee, 1999; see also Fig. 4). This formal problem should not be underestimated,

given the fact that non-neutral names can significantly bias further experimental research (Markov et al., 2008a,b). In spite of the mentioned uncertainties, and maybe due to the nomenclature bias, the evolution of the steroid binding ability in the NR3 family was, to date, only discussed based on the topology presented in Fig. 3A. It was proposed that the early acquisition of estrogen-binding in bilaterians was more parsimonious than the late ***convergent*** acquisition from a constitutive activator occurring three times independently in vertebrates, cephalochordates and annelids (Keay and Thornton, 2009, see also Eick and Thornton, this issue). However, if we take into account the topology proposed in Fig. 3B and the fact that ERR is a constitutive activator in mammals and mollusks (Giguère et al., 1988; Bannister et al., 2007), we find that the early acquisition of estrogen binding in bilaterians (Fig. 3C, left) and the late acquisition of estrogen binding in chordates (Fig. 3C, right) are equally parsimonious hypotheses. In both cases, four evolutionary steps are required. Thus, additional data are required to discriminate between the two possibilities.

2.2. Acquisition of hormonal binding from a steroid sensing background in the NR1H/I/J group

Steroid-binding nuclear receptors are not restricted to the NR3 subfamily. In the NR1 subfamily (Fig. 4), there is a group of steroid-binding receptors containing the arthropod ecdysone receptor EcR (NR1H1), the vertebrate oxysterol-binding LXR (NR1H2 and NR1H3), the vertebrate bile acid receptor FXRα (NR1H4) and the vertebrate bile alcohol receptor FXRβ (NR1H5). The orthologs of LXR and FXR in the urochordate *Ciona intestinalis* were recently shown to bind oxysterols and sulfated steroids (Reschly et al., 2008a; Reschly et al., 2008b). The NR1I/J group also contains the vertebrate PXR (NR1I2) and VDR (NR1I1), which bind vitamin D, bile acids and other cholesterol derivatives, the nematode DAF-12, which binds dafachronic acids (also a kind of steroids) and the insect HR96 (NR1J1), which binds cholesterol (Horner et al., 2009). This group also contains many not yet characterized receptors from amphioxus (Schubert et al., 2008), sea urchin (Howard-Ashby et al., 2006) and various nematodes (Abad et al., 2008), each with lineage-specific duplications. Thus, because most of the characterized receptors from this group, either in chordates or in ecdysozoans, are able to bind cholesterol or steroid derivatives, it is highly likely that the common ancestor of this group was also able to do so.

Functional data allow us to be more precise at least for the NR1I/J group. Members of this group regulate the xenobiotic response in vertebrates, *Drosophila* (HR96; King-Jones et al., 2006) and nematodes (nhr-8; Lindblom et al., 2001), thus indicating that the common ancestor of bilaterian NR1I/J may have had this ancestral function. If this hypothesis is true, it would be necessary to explain how the nematode DAF-12 and vertebrate VDR shifted from xenobiotic sensing, which implies the binding of many different ligands at the micromolar range, to hormone binding, that implies a binding of one specific molecule with high affinity. For nematode DAF-12, data are currently insufficient to draw an evolutionary scenario. But concerning VDR, it is possible to compare its properties to that of its paralogs PXR and CAR and also to compare the properties of VDR in various vertebrates. The three receptors share binding to some targets implicated in xenobiotic responses, such as the CYP3A genes. CYP3A are involved in hydroxylation of various xenobiotics and endobiotics, such as lithocholic acid, a cytotoxic secondary bile acid that is produced by mammalian intestinal bacteria. Additionally, the sea lamprey VDR is able to activate the transcription of a reporter gene bearing a response element of the mammalian CYP3A4 xenobiotic-metabolising enzyme, but not that of rat osteocalcin, which is a mammalian VDR target gene involved in bone physiology (Whitfield et al., 2003). This is consistent with the lack in lamprey of a calcified skeleton and plasma levels of calcitriol that are 7–8 times higher than those of other vertebrates, which correlates with the lower affinity of lamprey VDR for this ligand. Thus, even if calcitriol is present in lamprey and able to bind VDR, it may not have an hormonal function in that animal. The physiologically relevant lamprey VDR ligand may be a bile acid or another steroid-like molecule. We suggest that results from NR3 family steroid receptors should be interpreted in a similar way. Physiologically relevant estrogen-binding would be a chordate-specific feature. Estrogen binding by annelid NR3D may be viewed as a purely pharmacological property, as observed for the calcitriol-binding VDR in lamprey. Or estrogens may be one of the numerous ligands that can activate annelid NR3D during a xenobiotic response, as it is the case for estrogen binding by the vertebrate PXR or CAR.

3. NR ligand synthesis evolution and its implications on the state of AncSR

Another important element to consider in this debate is the evolution of NR ligand synthesis pathways. Thanks to the increase of genome sequences and functional characterization of enzymes in several species, much data are now available to tackle this long standing question.

3.1. Independent acquisition of steroidogenic synthesis pathways

One critical aspect to understand the context of ancestral NR binding is to decipher the timing of the appearance of NR ligands. Indeed it would be difficult to have a complete view of NR origins and evolution without integrating ligand synthesis. Again, the most numerous data exist for steroids, that were reported in almost all animal groups (Fig. 5), the most recent example being progesterone in rotifers (Stout et al., 2010) and even in plants. Unfortunately, many searches for "human"-type steroid hormones such as estradiol or progesterone throughout metazoan groups have been prone to artifacts and/or misidentification. To date, biochemical evidence (immunological and/or chromatographic methods linked to mass spectrometry) for presence of vertebrate steroids in lophotrochozoans, ecdysozoans and cnidarians have not been substantiated by molecular characterization of enzymes directly involved in their *de novo* biosynthesis (Lafont and Mathieu, 2007; Markov et al., 2008b). A recent search for orthologs of the enzymes implicated in the synthesis of vertebrate sex and adrenal steroids, as well as enzymes implicated in the synthesis of ecydsone and dafachronic acids showed that these enzymes appeared specifically following gene duplications in the vertebrate, arthropod and nematode lineages, respectively (Markov et al., 2009). This is also true for some key enzymes that are involved in the synthesis of oxysterols, bile acids and vitamin D. All these enzymes are members of multigenic families such as the cytochrome P450 (CYP), the short-chain dehydrogenases-reductases (SDR), the 3β-hydroxysteroid dehydrogenases (HSD3B) and the 5α-reductases (SRD5A). Of course, the phylogeny of these families is far from being fully elucidated, but there are nevertheless some robust nodes concerning two key enzymes in vertebrate sex and adrenal steroid synthesis: CYP19 and CYP11A.

The CYP19, also named aromatase, is a chordate-specific enzyme (Reitzel and Tarrant, 2010), that may either have been secondarily lost in other animals or, following the same reasoning as presented above for vertebrate NR3C, may be a highly derived chordate protein that branches basally to other metazoan CYPs due to long-branch attraction. However, in both cases, this indicates that there is no evidence for a CYP19 ortholog that would perform aromatase activity outside chordates. This should be put in perspective with the reports from aromatase activities in cnidarians or mollusks (Fig. 5). This means that this aromatase activity may be performed

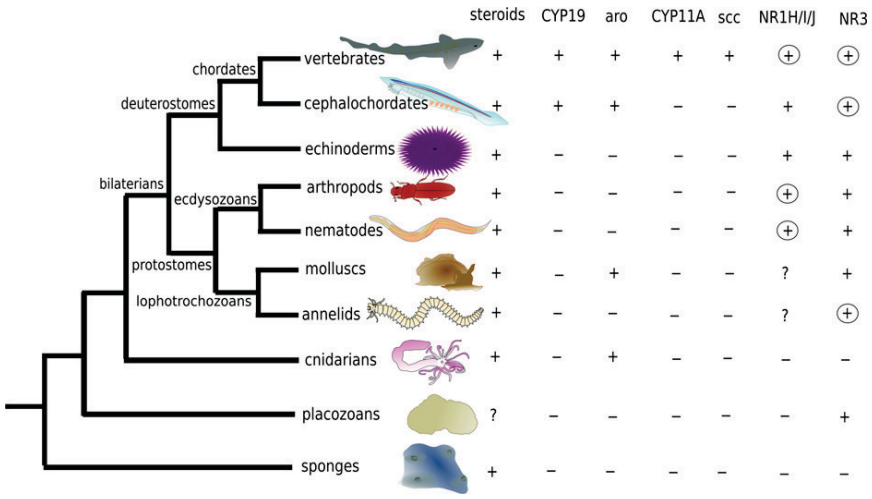


Fig. 5. Distribution of potential steroid receptors and steroidogenic enzymes among metazoans. Endogenous steroids are reported in all metazoans except for the placozoan *Trichoplax adhaerens*, but enzymes catalysing vertebrate-specific steroid reactions have a more restricted distribution. Biochemical reports on aromatase activity (aro) are available for cnidarians and mollusks, both of which lack a gene encoding a CYP19 enzyme, suggesting that this activity is due to a paralogous enzyme. However, side-chain cleavage of cholesterol (scc) has been reported only in vertebrates, and the side-chain cleavage enzyme CYP11A also vertebrate-specific. Among candidate steroid nuclear receptors, NR1H/I/J may be present in all bilaterians, but steroid binding (⊕) is known only in ecdysozoans and vertebrates. NR3 are present in bilaterians and placozoans, but steroid binding was reported only in annelids and vertebrates (⊕).

by another enzyme than CYP19, and on an endogenous substrate that is not necessarily testosterone, as in vertebrates. Moreover, because all cnidarians studied to date lack any NR1I/J/H or any NR3 (Reitzel and Tarrant, 2009), if there is any NR-mediated steroid signalling in these animals, it would not be mediated by a classical bilaterian-type NR. Thus, there is no reason to suppose that the physiologically active steroids in cnidarians, if they exist, would be of vertebrate-type, instead of being one of the various steroids that were identified in some cnidarians (Lafont and Mathieu, 2007). Concerning mollusks, that lack steroid-binding NR3, candidate steroid-binding receptors may be found in the NR1H/I/J, but once again, that would not be in favour of a vertebrate-type estrogen ligand, rather than one of the various hydroxylated steroids once identified in some mollusks. Concerning annelids, evidence of biochemical activity implicated in estrogen synthesis is poorly documented, and most steroid reports concern ecdysteroids (Lafont and Mathieu, 2007). In a recent report trying to link "vertebrate-like" steroid hormone levels to sexual maturity indexes in *Nereis diversicolor*, the authors acknowledge in the discussion that "*Quantification of steroid hormones in worms cannot permit [them] to take into account exogenous compounds accumulated in [worm tissues] and which can act as endocrine disruptors. The presence of these latter contaminants is well documented in the Seine estuary*" (Durou and Mouneyrac, 2007).

The CYP11A, also known as side-chain cleavage enzyme, is unambiguously vertebrate-specific (Fig. 5), having arisen from a duplication of an ancestral CYP11 gene leading to CYP11A and CYP11B, that is involved in glucocorticoid synthesis. To date, a lack of phylogenetic resolution prevents us from establishing which CYP is the most closely related to these two vertebrate enzymes in other animals, even in chordates. But due to the fact that both duplicates have different activities, there is no reason to imagine that the activity of this CYP should be more CYP11A-like than CYP11B-like. Interestingly, a recent study succeeded in reporting *in vitro* steroidogenic biochemical activities for most of the members of the sex steroid synthesis pathway, except from the side-chain cleavage activity, for which the only indirect evidence is a report of mRNA expression in amphioxus ovaries (Mizuta et al., 2008). However, this so-called "CYP11A" is not even the best candidate for an ortholog of ancestral vertebrate CYP11, but rather is a distant paralog, named CYP374A2, which is clearly orthologous to a sea urchin CYP (Markov et al., 2009). This means that, up to now, the estrogen synthesis pathway remains vertebrate-specific, even if it is possible that amphioxus synthesizes other steroids, for example aromatized steroids with a side chain, which could cross-react with the antibodies used to search for estrogen in amphioxus (Mizuta and Kubokawa, 2007). If it were firmly established that side-chain cleavage reaction was convergently acquired by an amphioxus enzyme, this other steroid could also be Δ5-androstenediol, which was shown to bind mammalian ERs with nanomolar affinity (Kuiper et al., 1997), as has already been suggested (Baker, 2002b). So for the moment, estrogens cannot be viewed as a physiological ligand for amphioxus SR because there is no sufficient evidence for estrogen in this species. Indeed, even if the amphioxus SR was shown to be activated by estradiol, it was at very high concentrations (Bridgham et al., 2008), which makes it possible that the real physiological activation mechanism for this receptor is through phosphorylation, interactions with coactivators or binding of another ligand. This case is reminiscent of the relationship between RXR and 9-cis retinoic acid in vertebrates: 9-cis RA is an excellent ligand of vertebrate RXR but has not been found in vertebrate extracts and does not regulate RXR activity *in vivo*.

3.2. A xenobiotic origin for vertebrate sex steroid hormones?

The lack of vertebrate-type steroids outside vertebrates leaves the question of the origin of steroid binding in the NR3 subfamily open. The late appearance of steroidogenic enzymes within highly multigenic and promiscuous families is in good agreement with the growing evidence that enzyme specificity should have evolved from low-specificity proteins catalysing a whole range of activities at low levels, to subfamilies with potent and highly specialized activities (Khersonsky et al., 2006). This is also consistent with the striking similarity between the xenobiotic response pathway in vertebrates and the sex and adrenal steroid metabolism (Fig. 6). Xenobiotic response is divided in phase I (hydroxylations by CYPs) and phase II (addition of additional hydrophilic residues on hydroxylated carbon), before transport outside the cell (Wada et al., 2009). Sex steroids are synthesized from cholesterol, mainly by a succession of hydroxylations catalysed by CYPs, and are directly degraded throughout a mechanism similar to the phase II xenobiotic response, involving enzymes from the SULT and

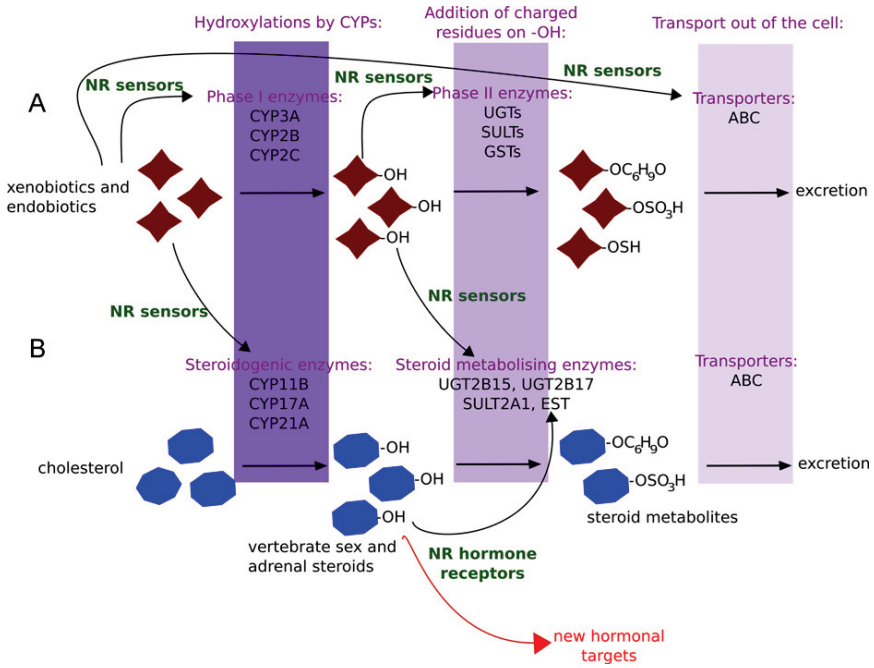


Fig. 6. Similarities between xenobiotic response and steroid metabolism in vertebrates. (A) Xenobiotic response is divided in phase I (hydroxylations by CYPs) and phase II (conjugation with charged species), before transport outside the cell. (B) Steroids are synthesized from cholesterol, mainly by a succession of hydroxylations catalysed by CYPs, and are directly degraded through a mechanism similar to phase II xenobiotic response, implicating sulfonations by SULTs and glucuronylations by UGTs. NR sensors like PXR, LXR or FXR are involved in transcriptional regulation of both steroidogenic and xenobiotic-metabolising CYPs, and they also regulate the transcription of enzymes that metabolise steroid hormones. We hypothesize that the similarity between both pathways reflects common ancestry. In this framework, the transcriptional control of steroid hormone metabolising enzymes by hormonal nuclear receptors can be viewed as conservation of ancestral regulatory pathways, whereas targets implicated in steroid signalling (such as peptide hormones) can be viewed as newly acquired targets after gene duplication.

UGT families (He et al., 2010). This similarity led to the hypothesis that both phase I of xenobiotic detoxification and steroid synthesis are homologous, meaning that steroid hormones may be recruited cholesterol metabolites (Baker, 2005; Markov et al., 2009). This would be in agreement with the proposition that thyroid hormones (T_3 and T_4) were dietary metabolites before of becoming hormones endogenously synthesized by the thyroid gland in vertebrates (Heyland et al., 2005), and with the fact that other high affinity NR ligands, such as retinoids and eicosanoids, are also derivatives from food components. Retinoids are known potent teratogens, when their concentration increases, and fatty acids are highly cytotoxic molecules due to their potential detergent effect on cell membranes (Theodosiou et al., 2010; Babin and Gibbons, 2009). Xenobiotics are often viewed and studied mainly as environmental man-made pollutants. Actually, for an heterotrophic organism, many of the molecules that are present in food can be viewed as xenobiotics, being molecules that are synthesized by an organism different from the consumer and having a possible negative effect on its metabolism. This is well illustrated by the known pleiotropic effects of genistein, a component of soy, in humans (Henley and Korach, 2006). Production of toxins is a very widespread defense mechanism from plants to bacteria, so we can reasonably infer that detoxification mechanisms would have evolved much earlier than the first metazoans. But the presence of an NR as a sensing transcription factor would have enabled the possibility of efficiently regulating the production of detoxifying enzymes in response to endocrine perturbation due to food nutrients (Baker, 2005). Moreover, such steroid xenobiotics may have allowed the coordination of the reproduction cycle with food availability, and endogenous synthesis of such molecules (or very close chemical analogs) may have stabilized the system.

4. Conclusion: the ancestral receptor may have been a nutritional sensor

The discussed data suggest that, at their origin, NRs were probably not hormonal receptors with high affinity for very specific compounds, a feature that was acquired later during evolution. We propose that the first NR was a lipid sensor, that is a receptor that could bind with micromolar affinity several different hydrophobic molecules, such as hemes, retinoids, steroids, fatty acids, eicosanoids or maybe other lipids that would have been dietary components of the early metazoans. Different from the previous proposition of a ancestral receptor with a structural ligand, that would bind permanently a lipophilic compound (Sladek, 2002), we think that this receptor was a true sensor, that was able to bind an interchangeable ligand. By interacting with a wide variety of compounds, this sensor would have been able to transfer as transcriptional activity subtle metabolic balances in the respective amounts of various compounds. Through duplications and neofunctionalization, it would either have secondarily lost the ligand-based regulation of transcriptional activation, or specialized into the highly specific binding of a particular molecule.

An important factor that remains to be integrated into such a framework is the evolution of signalling through bHLH-PAS transcription factors, which are the other known metazoan transcription factors recognised to activate transcription triggered by the binding of xenobiotics (Hahn, 2002) or gases bound to their hemes (Mukaiyama et al., 2006). Indeed, such xenobiotic-binding molecules would have overlapping roles, and it would not be possible to build a reliable evolutionary scenario without taking into account all the possible partners in the ancestral signalling pathway.

Acknowledgements

We are grateful to CNRS, EMBO, FRM, MENRT and Agence Nationale de la Recherche (AmphiNR program) for financial support. Work from our laboratory is also funded by the EU Cascade Network of Excellence and the integrated project Crescendo and the IAPP program SME-Receptor. We thank François Bonneton, Joanne Burden and the reviewers for critical reading of the manuscript and Gerard Benoit for useful suggestions. The mouse, nereis and amphioxus pictures in Figs. 3–5 are made or derived from works by Madeleine Price Ball, Dieter Tracey (IAN Image Library ian.umces.edu/imagelibrary/) and Piotr Michał Jaworski respectively and are released under the creative commons 3.0 licence (http://creativecommons.org/licenses/by-sa/3.0/).

References

Abad, P., Gouzy, J., Aury, J., Castagnone-Sereno, P., Danchin, E.G.J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., Caillaud, M., Coutinho, P.M., Dasilva, C., Luca, F.D., Deau, F., Esquibet, M., Flutre, T., Goldstone, J.V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T.R., Markov, G.V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T.J., Blaxter, M., Bleve-Zacheo, T., Davis, E.L., Ewbank, J.J., Favery, B., Grenier, E., Henrissat, B., Jones, J.T., Laudet, V., Maule, A.G., Quesneville, H., Rosso, M., Schiex, T., Smant, G., Weissenbach, J., Wincker, P., 2008. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat. Biotechnol. 26, 909–915.

Amero, S.A., Kretsinger, R.H., Moncrief, N.D., Yamamoto, K.R., Pearson, W.R., 1992. The origin of nuclear receptor proteins: a single precursor distinct from other transcription factors. Mol. Endocrinol. 6, 3–7.

Babin, P.J., Gibbons, G.F., 2009. The evolution of plasma cholesterol: direct utility or a "spandrel" of hepatic lipid metabolism? Prog. Lipid Res. 48, 73–91.

Baker, M.E., 2001. Evolution of 17beta-hydroxysteroid dehydrogenases and their role in androgen, estrogen and retinoid action. Mol. Cell. Endocrinol. 171, 211–215.

Baker, M.E., 2002a. Albumin, steroid hormones and the origin of vertebrates. J. Endocrinol. 175, 121–127.

Baker, M.E., 2002b. Recent insights into the origins of adrenal and sex steroid receptors. J. Mol. Endocrinol. 28, 149–152.

Baker, M.E., 2003. Evolution of adrenal and sex steroid action in vertebrates: a ligand-based mechanism for complexity. Bioessays 25, 396–400.

Baker, M.E., 2005. Xenobiotics and the evolution of multicellular animals: emergence and diversification of ligand-activated transcription factors. Integr. Comp. Biol. 45, 172–178.

Baker, M.E., 2008. *Trichoplax*, the simplest known animal, contains an estrogen-related receptor but no estrogen receptor: implications for estrogen receptor evolution. Biochem. Biophys. Res. Commun. 375, 623–627.

Bannister, R., Beresford, N., May, D., Routledge, E.J., Jobling, S., Rand-Weaver, M., 2007. Novel estrogen receptor-related transcripts in *Marisa cornuarietis*, a freshwater snail with reported sensitivity to estrogenic chemicals. Environ. Sci. Technol. 41, 2643–2650.

Benoit, G., Malewicz, M., Perlmann, T., 2004. Digging deep into the pockets of orphan nuclear receptors: insights from structural studies. Trends Cell. Biol. 14, 369–376.

Bertrand, S., Brunet, F.G., Escriva, H., Parmentier, G., Laudet, V., Robinson-Rechavi, M., 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. Mol. Biol. Evol. 21, 1923–1937.

Bridgham, J.T., Carroll, S.M., Thornton, J.W., 2006. Evolution of hormone-receptor complexity by molecular exploitation. Science 312, 97–101.

Bridgham, J.T., Brown, J.E., Rodríguez-Marí, A., Catchen, J.M., Thornton, J.W., 2008. Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. PLoS Genet. 4, e1000191.

Brown, C.M., Reisfeld, B., Mayeno, A.N., 2008. Cytochromes P450: a structure-based summary of biotransformations using representative substrates. Drug Metab. Rev. 40, 1–100.

Burris, T.P., 2008. Nuclear hormone receptors for heme: REV-ERBalpha and REV-ERBbeta are ligand-regulated components of the mammalian clock. Mol. Endocrinol. 22, 1509–1520.

Calléja, C., Messaddeq, N., Chapellier, B., Yang, H., Krezel, W., Li, M., Metzger, D., Mascrez, B., Ohta, K., Kagechika, H., Endo, Y., Mark, M., Ghyselinck, N.B., Chambon, P., 2006. Genetic and pharmacological evidence that a retinoic acid cannot be the RXR-activating ligand in mouse epidermis keratinocytes. Genes Dev. 20, 1525–1538.

de Lera, A.R., Bourguet, W., Altucci, L., Gronemeyer, H., 2007. Design of selective nuclear receptor modulators: RAR and RXR as a case study. Nat. Rev. Drug Discov. 6, 811–820.

de Rosny, E., de Groot, A., Jullian-Binard, C., Borel, F., Suarez, C., Pape, L.L., Fontecilla-Camps, J., Jouve, H., 2008. DHR51, the *Drosophila melanogaster* homologue of the human photoreceptor cell-specific nuclear receptor, is a thiolate heme-binding protein. Biochemistry.

Durou, C., Mouneyrac, C., 2007. Linking steroid hormone levels to sexual maturity index and energy reserves in *Nereis diversicolor* from clean and polluted estuaries. Gen. Comp. Endocrinol. 150, 106–113.

Eick, G.N., Thornton, J.W., this issue. Evolution of steroid receptors from an estrogen-sensitive ancestral protein. Mol. Cell. Endocrinol.

Ekins, S., Reschly, E.J., Hagey, L.R., Krasowski, M.D., 2008. Evolution of pharmacologic specificity in the pregnane X receptor. BMC Evol. Biol. 8, 103.

Escriva, H., Safi, R., Hänni, C., Langlois, M.C., Saumitou-Laprade, P., Stehelin, D., Capron, A., Pierce, R., Laudet, V., 1997. Ligand binding was acquired during evolution of nuclear receptors. Proc. Natl. Acad. Sci. U.S.A. 94, 6803–6808.

Escriva, H., Delaunay, F., Laudet, V., 2000. Ligand binding and nuclear receptor evolution. Mol. Cell. Endocrinol. 22, 717–727.

Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., Laudet, V., 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2, e102.

Evans, R.M., 1988. The steroid and thyroid hormone receptor superfamily. Science 240, 889–895.

Faus, H., Haendler, B., 2006. Post-translational modifications of steroid receptors. Biomed. Pharmacother. 60, 520–528.

Germain, P., Staels, B., Dacquet, C., Spedding, M., Laudet, V., 2006. Overview of nomenclature of nuclear receptors. Pharmacol. Rev. 58, 685–704.

Giguère, V., Yang, N., Segui, P., Evans, R.M., 1988. Identification of a new class of steroid hormone receptors. Nature 331, 91–94.

Gotoh, S., Ohgari, Y., Nakamura, T., Osumi, T., Taketani, S., 2008. Heme-binding to the nuclear receptor retinoid X receptor alpha (RXRalpha) leads to the inhibition of the transcriptional activity. Gene 423, 207–214.

Hahn, M.E., 2002. Aryl hydrocarbon receptors: diversity and evolution. Chem. Biol. Interact. 141, 131–160.

He, J., Cheng, Q., Xie, W., 2010. Minireview: Nuclear receptor-controlled steroid hormone synthesis and metabolism. Mol. Endocrinol. 24, 11–21.

Henley, D.V., Korach, K.S., 2006. Endocrine-disrupting chemicals use distinct mechanisms of action to modulate endocrine system function. Endocrinology 147, S25–S32.

Heyland, A., Hodin, J., Reitzel, A.M., 2005. Hormone signaling in evolution and development: a non-model system approach. Bioessays 27, 64–75.

Horner, M.A., Pardee, K., Liu, S., King-Jones, K., Lajoie, G., Edwards, A., Krause, H.M., Thummel, C.S., 2009. The *Drosophila* DHR96 nuclear receptor binds cholesterol and regulates cholesterol homeostasis. Genes Dev. 23, 2711–2716.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006. Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. Dev. Biol. 300, 90–107.

Huang, P., Chandra, V., Rastinejad, F., 2010. Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. Annu. Rev. Physiol. 72, 247–272.

Iwema, T., Billas, I.M.L., Beck, Y., Bonneton, F., Nierengarten, H., Chaumot, A., Richards, G., Laudet, V., Moras, D., 2007. Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. EMBO J. 26, 3770–3782.

Keay, J., Bridgham, J.T., Thornton, J.W., 2006. The *Octopus vulgaris* estrogen receptor is a constitutive transcriptional activator: evolutionary and functional implications. Endocrinology 147, 3861–3869.

Keay, J., Thornton, J.W., 2009. Hormone-activated estrogen receptors in annelid invertebrates: implications for evolution and endocrine disruption. Endocrinology 150, 1731–1738.

Khersonsky, O., Roodveldt, C., Tawfik, D.S., 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. Curr. Opin. Chem. Biol. 10, 498–508.

Koelle, M.R., Talbot, W.S., Segraves, W.A., Bender, M.T., Cherbas, P., Hogness, D.S., 1991. The *Drosophila EcR* gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. Cell 67, 59–77.

King-Jones, K., Horner, M.A., Lam, G., Thummel, C.S., 2006. The DHR96 nuclear receptor regulates xenobiotic responses in *Drosophila*. Cell. Metab. 4, 37–48.

Krylova, I.N., Sablin, E.P., Moore, J., Xu, R.X., Waitt, G.M., MacKay, J.A., Juzumiene, D., Bynum, J.M., Madauss, K., Montana, V., Lebedeva, L., Suzawa, M., Williams, J.D., Williams, S.P., Guy, R.K., Thornton, J.W., Fletterick, R.J., Willson, T.M., Ingraham, H.A., 2005. Structural analyses reveal phosphatidyl inositols as ligands for the NR5 orphan receptors SF-1 and LRH-1. Cell 120, 343–355.

Kuiper, G.G., Carlsson, B., Grandien, K., Enmark, E., Häggblad, J., Nilsson, S., Gustafsson, J.A., 1997. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. Endocrinology 138, 863–870.

Lafont, R., Mathieu, M., 2007. Steroids in aquatic invertebrates. Ecotoxicology 16, 109–130.

Laudet, V., Hänni, C., Coll, J., Catzeflis, F., Stéhelin, D., 1992. Evolution of the nuclear receptor gene superfamily. EMBO J. 11, 1003–1013.

Laudet, V., 1997. Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. J. Mol. Endocrinol. 19, 207–226.

Lindblom, T.H., Pierce, G.J., Sluder, A.E., 2001. A *C. elegans* orphan nuclear receptor contributes to xenobiotic resistance. Curr. Biol. 11, 864–868.

Makishima, M., Lu, T.T., Xie, W., Whitfield, G.K., Domoto, H., Evans, R.M., Haussler, M.R., Mangelsdorf, D.J., 2002. Vitamin D receptor as an intestinal bile acid sensor. Science 296, 1313–1316.

Markov, G., Lecointre, G., Demeneix, B., Laudet, V., 2008a. The "street light syndrome", or how protein taxonomy can bias experimental manipulations. Bioessays 30, 349–357.

Markov, G.V., Paris, M., Bertrand, S., Laudet, V., 2008b. The evolution of the ligand/receptor couple: a long road from comparative endocrinology to comparative genomics. Mol. Cell. Endocrinol. 293, 5–16.
Markov, G.V., Tavares, R., Dauphin-Villemant, C., Demeneix, B.A., Baker, M.E., Laudet, V., 2009. Independent elaboration of steroid hormone signaling pathways in metazoans. Proc. Natl. Acad. Sci. U.S.A. 106, 11913–11918.
Mic, F.A., Molotkov, A., Benbrook, D.M., Duester, G., 2003. Retinoid activation of retinoic acid receptor but not retinoid X receptor is sufficient to rescue lethal defect in retinoic acid synthesis. Proc. Natl. Acad. Sci. U.S.A. 100, 7135–7140.
Mizuta, T., Kubokawa, K., 2007. Presence of sex steroids and cytochrome P450 genes in amphioxus. Endocrinology 148, 3554–3565.
Mizuta, T., Asahina, K., Suzuki, M., Kubokawa, K., 2008. In vitro conversion of sex steroids and expression of sex steroidogenic enzyme genes in amphioxus ovary. J. Exp. Zool. 309A, 83–93.
Moore, D.D., 1990. Diversity and unity in the nuclear hormone receptors: a terpenoid receptor superfamily. New Biol. 2, 100–105.
Mukaiyama, Y., Uchida, T., Sato, E., Sasaki, A., Sato, Y., Igarashi, J., Kurokawa, H., Sagami, I., Kitagawa, T., Shimizu, T., 2006. Spectroscopic and DNA-binding characterization of the isolated heme-bound basic helix-loop-helix-PAS-A domain of neuronal PAS protein 2 (NPAS2), a transcription activator protein associated with circadian rhythms. FEBS J. 273, 2528–2539.
Nuclear Receptors Nomenclature Committee, 1999. A unified nomenclature system for the nuclear receptor superfamily. Cell 97, 161–163.
O'Malley, B.W., 1989. Did eucaryotic steroid receptors evolve from intracrine gene regulators? Endocrinology 125, 1119–1120.
Paris, M., Escriva, H., Schubert, M., Brunet, F., Brtko, J., Cieselski, F., Roecklin, D., Vivat-Hannah, V., Cravedi, J.-P., Scanlan, T.S., Renaud, J.-P., Holland, N.D., Laudet, V., 2008a. Amphioxus metamorphosis and the origin of the thyroid hormone signaling pathway. Curr. Biol. 18, 825–830.
Paris, M., Pettersson, K., Schubert, M., Bertrand, S., Pongratz, I., Escriva, H., Laudet, V., 2008b. An amphioxus orthologue of the estrogen receptor that does not bind estradiol: insight into estrogen receptor evolution. BMC Evol. Biol. 8, 219.
Reitzel, A.M., Tarrant, A.M., 2009. Nuclear receptor complement of the cnidarian *Nematostella vectensis*: phylogenetic relationships and developmental expression patterns. BMC Evol. Biol. 9, 230.
Reitzel, A.M., Tarrant, A.M., 2010. Correlated evolution of androgen receptor and aromatase revisited. Mol. Biol. Evol. 27, 2211–2215.
Reschly, E.J., Ai, N., Welsh, W.J., Ekins, S., Hagey, L.R., Krasowski, M.D., 2008a. Ligand specificity and evolution of liver X receptors. J. Steroid Biochem. Mol. Biol. 110, 83–94.
Reschly, E.J., Ai, N., Ekins, S., Welsh, W.J., Hagey, L.R., Hofmann, A.F., Krasowski, M.D., 2008b. Evolution of the bile salt nuclear receptor FXR in vertebrates. J. Lipid Res. 49, 1577–1587.
Rochette-Egly, C., 2003. Nuclear receptors: integration of multiple signalling pathways through phosphorylation. Cell Signal. 15, 355–366.
Schubert, M., Brunet, F., Paris, M., Bertrand, S., Benoit, G., Laudet, V., 2008. Nuclear hormone receptor signaling in amphioxus. Dev. Genes Evol. 218, 651–665.
Schug, T.T., Berry, D.C., Shaw, N.S., Travis, S.N., Noy, N., 2007. Opposing effects of retinoic acid on cell growth result from alternate activation of two different nuclear receptors. Cell 129, 723–733.
Sladek, F., 2002. Desperately seeking . . .something. Mol. Cell 10, 219–221.
Sladek, F.M., this issue. What are nuclear receptor ligands? Mol. Cell. Endocrinol.
Stout, E.P., Clair, J.J.L., Snell, T.W., Shearer, T.L., Kubanek, J., 2010. Conservation of progesterone hormone function in invertebrate reproduction. Proc. Natl. Acad. Sci. U.S.A. 107, 11859–11864.
Theodosiou, M., Laudet, V., Schubert, M., 2010. From carrot to clinic: an overview of the retinoic signaling pathway. Cell. Mol. Life Sci. 67, 1423–1445.
Thomas, J.H., 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3, e67.
Thornton, J.W., 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. Proc. Natl. Acad. Sci. U.S.A. 98, 5671–5676.
Thornton, J.W., Need, E., Crews, D., 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301, 1714–1717.
Tzertzinis, G., Egaña, A.L., Palli, S.R., Robinson-Rechavi, M., Gissendanner, C.R., Liu, C., Unnasch, T.R., Maina, C.V., 2010. Molecular evidence for a functional ecdysone signaling system in *Brugia malayi*. PLoS Negl. Trop. Dis. 4, e625.
Wada, T., Gao, J., Xie, W., 2009. PXR and CAR in energy metabolism. Trends Endocrinol. Metab. 20, 273–279.
Weihua, Z., Lathe, R., Warner, M., Gustafsson, J.A., 2002. An endocrine pathway in the prostate, ERbeta, AR, 5alpha-androstane-3beta,17beta-diol, and CYP7B1, regulates prostate growth. Proc. Natl. Acad. Sci. U.S.A. 99, 13589–13594.
Whitfield, G.K., Dang, H.T.L., Schluter, S.F., Bernstein, R.M., Bunag, T., Manzon, L.A., Hsieh, G., Dominguez, C.E., Youson, J.H., Haussler, M.R., Marchalonis, J.J., 2003. Cloning of a functional vitamin D receptor from the lamprey (*Petromyzon marinus*), an ancient vertebrate lacking a calcified skeleton and teeth. Endocrinology 144, 2704–2716.
Yuan, X., Ta, T.C., Lin, M., Evans, J.R., Dong, Y., Bolotin, E., Sherman, M.A., Forman, B.M., Sladek, F.M., 2009. Identification of an endogenous ligand bound to a native orphan nuclear receptor. PLoS One 4, e5609.
Zimmerman, A.W., Veerkamp, J.H., 2002. New insights into the structure and function of fatty acid-binding proteins. Cell. Mol. Life Sci. 59, 1096–1116.

Research article

MBE-10-0988-Revised

Evolution of nuclear retinoic acid receptor alpha (RARα) phosphorylation sites. Serine gain provides fine-tuned regulation

Eric Samarut [1,2], Ismail Amal [1], Gabriel V. Markov[2,3], Roland Stote [1], Annick Dejaegere[1], Vincent Laudet[2] and Cécile Rochette-Egly [1,4]

1 IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), INSERM, U596; CNRS, UMR7104; Université de Strasbourg, 1 rue Laurent Fries, BP 10142, 67404 Illkirch Cedex, France.

2 Institut de Génomique Fonctionelle de Lyon; UMR 5242; INRA; Université de Lyon; Ecole Normale Supérieure de Lyon; 46 Allée d'Italie, 69364 Lyon, France

3 UMR 7221, Museum d'Histoire Naturelle, Paris, France.

4 Corresponding author. Tel. (33) 3 88 65 34 59; Fax. (33) 3 88 65 32 01; E-Mail: cegly@igbmc.fr

Running head: Evolution of RAR phosphorylation sites

Key words: nuclear retinoic acid receptor, phosphorylation, evolution

Abstract

The human nuclear retinoic acid (RA) receptor alpha (hRARα) is a ligand-dependent transcriptional regulator, which is controlled by a phosphorylation cascade. The cascade starts with the RA-induced phosphorylation of a serine residue located in the Ligand Binding Domain, S(LBD), allowing the recruitment of the cdk7/cyclinH/MAT1 subcomplex of TFIIH through the docking of cyclin H. It ends by the subsequent phosphorylation by cdk7 of an other serine located in the N-terminal domain, S(NTD). Here we show that this cascade relies on an increase in the flexibility of the domain involved in cyclin H binding, subsequently to the phosphorylation of S(LBD). Owing to the functional importance of RARα in several vertebrate species, we investigated whether the phosphorylation cascade was conserved in zebrafish (*Danio rerio*), which expresses two RARα genes, RARα-A and RARα-B. We found that in zebrafish RARαs, S(LBD) is absent, while S(NTD) is conserved and phosphorylated. Therefore we analyzed the pattern of conservation of the phosphorylation sites and traced back their evolution. We found that S(LBD) is most often absent outside mammalian RARα and appears late during vertebrate evolution. In contrast, S(NTD) is conserved, indicating that the phosphorylation of this functional site has been under ancient high selection constraint. This suggests that, during evolution, different regulatory circuits control RARα activity.

Introduction

Retinoic Acid (RA), the main active metabolite of vitamin A plays a critical role in many biological processes such as cell proliferation and differentiation, embryonic development and adult homeostasis (Bour, Taneja, and Rochette-Egly 2006; Mark, Ghyselinck, and Chambon 2009; Theodosiou, Laudet, and Schubert 2010). RA acts through nuclear receptors, RARs, which have been identified in a wide variety of animals.

There is one unique RAR ancestral gene for which an ortholog is known in some prostosomes such as mollusks (*Lottia gigantea*) and annelids (*Capitella capitata*) (Campo-Paysaa et al. 2008; Albalat and Canestro 2009), and in some invertebrate deusterostomes such as echinoderms (*Strongylocentrolus purpuratus* (Canestro et al. 2006; Marletaz et al. 2006)), cephalochordates (*Branchiostoma floridae* (Escriva et al. 2002a; Fujiwara 2006)], and urochordates (*Ciona intestinalis and Polyandrocarpa misakiensis* (Hisata et al. 1998; Fujiwara 2006)). Early during vertebrates evolution, the total number of genes markedly increased by two rounds (2R) of whole genome duplication (Dehal and Boore 2005). This is why vertebrates have 3 RAR paralogous genes that encode the three known subtypes of receptors: α (NR1B1), β (NR1B2) and γ (NR1B3) (Escriva et al. 2006; Germain et al. 2006). Note that in teleost fishes, a third round (3R) of whole genome duplication combined to gene losses occurred (Amores et al. 1998; Postlethwait et al. 1998), giving rise to 4 RAR genes in zebrafish (Bertrand et al. 2007). The history of RARs in regard to genome duplications have been addressed in several phylogenetic studies that clearly validated this evolutionary scenario (Escriva et al. 2002b; Jaillon et al. 2004; Robinson-Rechavi, Boussau, and Laudet 2004; Bertrand et al. 2007; Kuraku, Meyer, and Kuratani 2009). Note that the timing of the genome duplications inferred from a recent analysis of the RAR synteny group by Kuraku et al (Kuraku, Meyer, and Kuratani 2009).

RARs are ligand-dependent transcriptional regulators (for review, see (Rochette-Egly and Germain 2009) and references therein), which bind to specific sequence elements located in the promoters of target genes. They have a well-defined domain organization, consisting mainly of a central DNA-binding domain (DBD) linked to a C-terminal ligand-binding domain (LBD) and a N-terminal domain (NTD) (fig.1A). While the NTDs are naturally not structured and not conserved (Dyson and Wright 2005;

Lavery and McEwan 2005), DBDs and LBDs are highly structured and depict a significant degree of conservation between vertebrate species (Escriva et al. 2006; Campo-Paysaa et al. 2008; Theodosiou, Laudet, and Schubert 2010). Briefly, the DBD contains two typical cysteine-rich zinc-binding motifs and two alpha helices, which cross at right angles, folding into a globular conformation to form the core of the DBD. Concerning the LBD, it shows a common fold comprising 12 conserved alpha helices and a short beta turn, arranged in three layers to form an antiparallel « alpha-helical sandwich » (Renaud et al. 1995) (fig.1B). Ligand binding triggers conformational changes in the LBD that direct the dissociation/association of several coregulator protein complexes and thereby the transcription of target genes (Rochette-Egly and Germain 2009).

In addition to this scenario, a new concept emerged according to which RARs are also subjected to rapid phosphorylation cascades. Recent studies from our laboratory (Bruck et al. 2009) demonstrated that, via non-genomic effects, RA activates rapidly the p38MAPK/MSK1 pathway, which in turn leads to the phosphorylation of the RARα subtype (mouse and human) at two serine residues located in solvent-accessible regions of the receptor. One serine is located in the LBD [S(LBD)], in a loop between helices 9 and 10 (L9-10) (fig.1A and 1B) and belongs to an arginine-lysine-rich motif that corresponds to a consensus phosphorylation motif for MSK1 (fig. 1A). The other serine residue is located in the NTD [S(NTD)], in a proline-rich motif (fig. 1A) and is phosphorylated by cdk7 (Rochette-Egly et al. 1997; Bastien et al. 2000), which forms with cyclin H and MAT1 the CAK subcomplex of the general transcription factor TFIIH. Most interestingly, the correct positioning of cdk7 and thereby the efficiency of the NTD phosphorylation rely on the docking of cyclin H at a specific site of the LBD located in loop L8-9 and the N-terminal part of helix 9 (H9) (fig. 1A and 1B) (Bour et al. 2005).

In the case of human and mouse RARα, we previously demonstrated that the phosphorylation of the two serines results from a coordinated phosphorylation cascade starting with the phosphorylation by MSK1 of S(LBD) (fig.1B) (Bruck et al. 2009). Phosphorylation of this residue increases the binding efficiency of cyclin H to the nearby loop L8-9 (fig.1B), allowing the right positioning of cdk7 and the phosphorylation of the serine located in the NTD (fig.1A) by this kinase (Gaillard et al. 2006). Finally phosphorylation of S(NTD) leads to the recruitment of RARα to promoters (Bruck et al. 2009). Whether it also controls the association/dissociation of specific coregulators as

described for the other RAR subtypes (Vucetic et al. 2008; Lalevee et al. 2010) is still unknown.

In RARs, ligand binding is conserved at least in chordates (Escriva et al. 2006), indicating that the ligand-triggered conformational changes are a common feature of all chordate species. Interestingly, the high regulatory potential of the phosphorylation cascade also makes phosphorylations prime candidates for evolutionary studies. It must be noted that S(LBD) and S(NTD) are conserved in the different human and mouse RAR subtypes α, β and γ (Rochette-Egly 2003; Rochette-Egly and Germain 2009), but the above cascade has been described only in the context of RARα (Bruck et al. 2009), which has ubiquitous or quite widespread expression patterns. There are still no indications whether this cascade also occurs in the context of the other RAR paralogs (RARβ and RARγ), which show rather complex tissue specific expression (Dolle 2009). Therefore we analyzed the pattern of conservation of the S(NTD) and S(LBD) phosphorylation sites, focusing on the RARα subtype.

First we demonstrated that in non-mammalian vertebrates exemplified by zebrafish, the S(NTD) of RARα is conserved, while S(LBD), the phosphorylation of which increases the flexibility of L8-9 that is required for cyclin H binding, is absent. However this process was compensated by changes in the sequence of L8-9 mimicking the conformation/flexibility changes induced by phosphorylation. Then we traced back the evolution of chordate RARα phosphorylation sites. This work led to the conclusion that in RARα, S(NTD) is evolutionary conserved, indicating that the phosphorylation of this functional site has been under ancient strong selection constraint. However, S(LBD) is most often absent outside mammalian RARα. This indicates that the fine-tuned phosphorylation cascade of RARα, starting at S(LBD), appears late during vertebrate evolution. Thus, the evolution of phosphorylation sites appears to provide a reservoir of changes in order to provide additional levels of regulation of critical functional proteins.

Materials and Methods

Sequences alignment and ancestral sequences reconstructions

RAR protein sequences were found in the nuclear receptor database (NureXbase) (http://nurexbase.prabi.fr) and by Blast and gene homology (NCBI). Multiple sequence alignments were performed by the MUSCLE software (Edgar 2004) and analyzed with ClustalX. Sequence assignment was verified by phylogenetical reconstruction as in (Escriva et al. 2006). Ancestral sequences were estimated with PAML (Yang 1997) under the JTT+γ substitution model from a dataset of 71 sequences containing a 232 amino acid long portion of the LBD. Some sequences containing obvious predictions errors or indels at unambiguous positions were manually corrected by parsimony. Other sequences with too many uncertainties were excluded from the reconstruction dataset. PhyML (Guindon and Gascuel 2003) generated the starting tree.

Molecular Dynamics Simulations

Structure preparation

Given the lack of an experimental structure for the LDB of human (h) RARα in an agonist form (holo) at the start of this work, a model structure was assembled from closely related structures available in the Protein Data Bank (Berman et al. 2000). The majority of the structure that includes helix 1 (H1) to helix 10 (H10) was taken from the structure of hRARα (PDBID 1DKF) bound to the selective antagonist BMS614 (Bourguet et al. 2000). Structural information for an agonist conformation of H11 and H12 was taken from the structures of hRARγ (PDBID 1FCZ) and RARβ (PDBID 1XAP) in the agonist forms (Klaholz, Mitschler, and Moras 2000; Germain et al. 2004). Side chains specific to hRARα were positioned using the Scwrl3.0 software (Canutescu, Shelenkov, and Dunbrack 2003). The structures of 9-cis RA and of a fragment of the TRAP220 coactivator were obtained from the structure of RARβ (PDBID 1XDK) in an agonist conformation (Pogenberg et al. 2005). The protonation states of all titratable groups at physiological pH (7.4) were determined as described in (Schaefer, van Vlijmen, and Karplus 1998) and all were found to favor their standard protonation states. Our model shows a very high degree of correspondence with an experimental structure of hRARα in an agonist form, which has been recently deposited in the Protein Data Bank (3A9E) (Sato et al. 2010), after the termination of this work.

Structural models were also constructed for apo hRARα, , i.e. in the absence of ligand and coactivator peptide. Under these conditions, the C-terminal end of the LBD, in particular H12, extends toward the solvent where it displays significant conformational flexibility (Renaud and Moras 2000). In the absence of any experimental apo hRARα structure, a model was constructed that kept H1 to H10 in the same conformation as the holo structure, but repositioned the C-terminal end based on the apo RXR structure (PDB 3A9E) (Sato et al. 2010). This model was constructed using the Modeller 9v8 program (Sali and Blundell 1993). Given that in the apo structures, H12 is conformationally mobile, variants of this apo model were constructed with different initial positions of H12. As all the initial apo models gave similar simulation results, we presented the data corresponding to the initial structure (Sato et al. 2010).

The LBDs of zebrafish (zf) RARα (-A and -B) were constructed by modifying all residues that differ from hRARα, maintaining the backbone conformation and modifying the side chains using the SCRWL4 program (Canutescu, Shelenkov, and Dunbrack 2003). For zfRARα-A, this involved the following side-chain modifications: E183D, V184T, G185E, E186Q, L187M, E189D, K190R, A201S, N211S, Q216R, S219A, I222V, I335L, P345A, R347K, M350V, V361I, K365N, S369H, R370K. For zfRARα-B, the modifications were: E183D, V184T, G185E, E186K, L187M, K190Q, A201S, S214A, E215D, Q216H, S219A, I222V, E280D, I335L, P345S, R347K, M350E, V361I, K365N, S369H, R370K. A similar protocol was used to construct hRARα mutants (hRARαP345G/D346A and hRARαP345A). The models for the apo forms of zfRARαs and the hRARα mutants were constructed as above.

Molecular simulations

All molecular dynamic simulations were done using the CHARMM program (Brooks et al. 1983) and the all atom parameter set of CHARMM27 (MacKerell et al. 1998), with CMAP corrections (Mackerell, Feig, and Brooks 2004). Hydrogen atoms were added using the HBUILD module (Brunger and Karplus 1988). Bonds between heavy atoms and hydrogen atoms were constrained using SHAKE (Ryckaert, Ciccotti, and Berendsen 1977). We employed a shift-type cutoff at 14 Å for electrostatic interactions, and a switch-type cutoff at 12.0 Å for the van der Waals energy terms.

The system was energy minimized using the steepest descent algorithm after placing harmonic constraints on the backbone and side chain heavy atoms with force

constants of 50 and 100 kcal.mol^{-1}Å^{-2}, respectively. The force constants were systematically scaled by a factor of 0.65 and minimization was repeated until there were no constraints on the protein. The protein was then solvated with a shell of explicit TIP3P water molecules (Gaillard, Dejaegere, and Stote 2009) extending 12Å from the protein surface. The system was equilibrated in two phases. In the first phase, a 20 ps molecular dynamics simulation of the water around the fixed protein was performed with a time step of 2 fs. In the second phase, the entire solvated protein was heated to 300K and equilibrated. During heating, velocities were assigned every 50 steps from a Gaussian distribution function. During equilibration, velocities were scaled by a single factor only when the average temperature was lying outside the 300 ± 10 K window. This was followed by a 10ns production phase without any further intervention.

Simulations were stable as measured by the backbone root-mean-square coordinate differences (RMSD), which were all less than 1.24 Å with respect to the initial starting structure. The phosphate group was assigned a charge of -2 based on the pKa of 6.5 (Kast et al. 2010).

Using the above protocol, simulations were run for the LBD of hRARα unphosphorylated or phosphorylated at S(LBD) either in the apo or holo forms. Simulations were also run for hRARαP345G/D346A, hRARαP345A and zfRARα-A and -B in the apo-forms. Upon completion of the simulations, the root-mean-square fluctuations (RMSfl) as well as the root-mean-square coordinate differences (RMSD) were calculated from the trajectories.

Plasmids

The pSG5- and pGEX-2T-based expression vectors for hRARα1 have been previously described (Bour et al. 2005). The full length or truncated cDNAs of zfRARα-A and zfRARα-B were amplified by PCR and inserted into pSG5-hER-B10-tag or pGEX-2T vectors. The cDNA of hcyclin H (a gift from D. Busso, IGBMC) was inserted into pCX-HA-FLAG. The cDNA of zfcyclin H (Liu et al. 2007) was inserted into pCX-HA or pET-15b. All constructs were generated using standard cloning procedures and were verified by PCR, restriction enzyme analysis and DNA sequencing. The sequence of primers used for PCR amplifications are available upon request.

Antibodies

Mouse monoclonal antibodies recognizing hRARα phosphorylated at S77, cyclin H and the epitope B of the estrogen receptor (B10) were previously described (Ali et al. 1993; Bruck et al. 2009). Rabbit polyclonal recognizing the N-terminal part of cyclin H and anti-FLAG monoclonal antibodies were from Sigma.

Mouse monoclonal antibodies recognizing zfRARα-B phosphorylated at the conserved serine residue located in the N-terminal proline-rich domain (S72) were generated by immunization of Balb/c mice with a synthetic phosphopeptide (EEMVPSSPS(p)PPPPPRVYKPC). Six-week-old female BALB/c mice were injected intraperitoneally (thrice at two weeks intervals) with 100 µg of peptide coupled to ovalbumin and 100 µg of poly I/C as adjuvant. Mice with positive sera were reinjected four days prior to hybridoma fusion and spleens were fused with Sp2/0.Ag14 myeloma cells. After hybridoma cell selection and cloning (de StGroth and Scheidegger 1980), the culture supernatants were tested by differential ELISA with the phosphopeptide, the corresponding non-phosphopeptide and an irrelevant phosphopeptide. Positive clones were confirmed by immunoblotting and cloned twice on soft agar. Ascites fluids were prepared by injection of 2×10^6 hybridoma cells into pristane-primed BALB/c mice.

Cell lines, transfections and immunoprecipitation experiments

COS-1 cells were grown and transiently transfected as described (Bour et al. 2005). ZF13 cells were grown at 27°C, in Leibovitz L-15 medium (Invitrogen) supplemented with 5% Fetal Calf Serum and 15mM HEPES and transiently transfected by using FuGene 6 reagent (Roche). Immunoprecipitations were performed with cell extracts prepared from paraformaldehyde-fixed cells (Bruck et al. 2009).

In vitro binding and phosphorylation experiments

GST and GST fusion proteins expressed in *Escherichia coli* were immobilized onto glutathione-Sepharose beads and incubated with recombinant human cyclin H over expressed in insect Sf9 cells (Bour et al. 2005) or with purified bacterially expressed zfcyclin H. Bound proteins were immunoprobed and quantified by using the Chemigenius XE imaging system as described (Bour et al. 2005). Data were analyzed according to standard statistical procedures using Graph Pad Prism 5.0 and compared using the Tukey's test in conjunction with ANOVA.

In vitro phosphorylation experiments were performed with equimolar amounts of immobilized GST-RARα proteins (5µg). Phosphorylation by the purified cdk7/cyclin H complex (Bour et al. 2005) was performed as in (Bruck et al. 2009) and detected by immunoblotting with antibodies recognizing specifically the phosphorylated forms. Phosphorylation by recombinant active MSK1 (Millipore Upstate Chemicon) (30ng) was performed in the presence of γ[^{32}P] as described (Rochette-Egly et al. 1995) and visualized by autoradiography.

Results

In mammalian RARα, phosphorylation of S(LBD) increases the dynamics/flexibility of the cyclin H binding domain.

In hRARα, the upstream serine residue of the phosphorylation cascade, S369 [S(LBD)], is located in the LBD, in loop L9-10 within an arginine-lysine-rich motif (fig. 1A and 1B). This serine is in the vicinity of a specific domain of the LBD, encompassing loop L8-9 and the N-terminal tip of H9 and involved in the binding of cyclin H (fig.1A and 1B) (Bour et al. 2005). Phosphorylation of S(LBD) has been shown to increase the ability of hRARα to interact with cyclin H, with a characteristic downstream consequence on the phosphorylation by cdk7 of the serine located in the NTD [S(NTD)] (Gaillard et al. 2006; Bruck et al. 2009) (fig.1A). This is a typical model of substrate recognition by a protein-kinase through association via another substrate-binding subunit.

To further investigate the consequences of phosphorylation on the LBD of hRARα, molecular dynamic simulations (MD) were performed (see materials and methods) to analyze whether phosphorylation of S(LBD) generates conformational changes affecting the cyclin H binding domain located at a 30Å distance, in L8-9 (fig.1B). Simulations were first performed with the holo form of hRARα (i.e. in the presence of RA and of a coactivator peptide), which is closest to the *in vivo* experimental phosphorylation studies (Bruck et al. 2009).

Average structures of the native and phosphorylated forms of hRARα were calculated and the root-mean-square deviations (RSMD) from the initial structures were measured (fig. 2A). RMSD of the backbone atoms forming secondary structure was less than 1.0 Å, indicating that the overall structure of the LBD domain is conserved upon phosphorylation. However, local conformational changes of loop L8-9 were observed with RMSD values in the order of 4Å. These conformational changes are shown in Figure 2A where the average structures of unphosphorylated and phosphorylated hRARα are superposed. An upward displacement of L8-9 is clearly visible, linked to an upward bending of helix H9. These changes likely result from the locally enhanced electrostatic environment due to the -2 charge of the phosphate moiety.

More significant however, is the increase in the local conformational dynamics or flexibility of L8-9 as measured by the atomic root-mean-square fluctuations (RMSfl) averaged by-residue (Fidelak et al.). RMSfl are directly related to the temperature

factors determined during an x-ray crystallography structural study and are a direct calculation of local, short-time scale dynamics of L8-9. As shown in Figure 2B, in the absence of S(LBD) phosphorylation, L8-9 was generally more flexible than the neighboring helices, with an RMSfl in the order of 0.9Å. However, with S(LBD) phosphorylated, the average RMSfl of L8-9 increased by a factor of 2. Thus, the simulations clearly indicate that S(LBD) phosphorylation affects the cyclin H binding domain. In the absence of an experimental structure of the hRARα – cyclin H complex, the investigation of the detailed molecular mechanism of this allosteric signaling was, however, beyond the scope of this study.

Then, molecular dynamics simulations were repeated without or with S(LBD) phosphorylated, but with the apo-form of hRARα in order to assess whether phosphorylation can still affect L8-9 structural dynamics in the absence of RA and of a coactivator peptide and with H12 in an extended conformation (see materials and methods). RMSfl analysis shows that L8-9 exhibits an increased flexibility when the apo form was phosphorylated at S(LBD) (fig. 2C). This suggests that S(LBD) phosphorylation by itself can affect the conformational dynamics of the cyclin H binding domain of hRARα in the apo form.

In conclusion, from these results and our previous experimental results (Gaillard et al. 2006; Bruck et al. 2009), one can suggest that the increase in the flexibility of loop L8-9 observed upon phosphorylation of S(LBD), might facilitate the binding of cyclin H to this domain.

In zebrafish RARα, S(LBD) is absent but S(NTD) is conserved and phosphorylated.

Given the functional importance of RARα not only in mouse and human, but also in other vertebrate species such as zebrafish (Dolle 2009; Linville et al. 2009), we investigated whether the phosphorylation cascade was conserved in zebrafish (*Danio rerio*), which expresses two RARα genes, RARα-A and RARα-B.

Sequence alignment revealed that in zebrafish (zf) RARα-A and RARα-B, the S(LBD) residue was not present (fig. 1A). Instead, a histidine residue was found. However, the arginine-lysine-rich motif flanking this residue was well conserved (fig. 1A). Accordingly, *in vitro* phosphorylation experiments indicated that the LBDs of zfRARαs were not phosphorylated by MSK1, the upstream kinase involved in the

phosphorylation cascade (fig. 3A). As phosphorylation sites that are not strictly conserved at a specific position can be compensated by others, driving the same functions (Nguyen Ba and Moses 2010), we investigated whether other phosphorylation sites are present in the nearby region. However, an *in silico* prediction of potential phosphorylation sites in the LBDs of zfRARα-A and zfRARα-B, did not reveal any other compensatory phosphorylation sites.

In contrast, in zfRARα-A and zfRARα-B, the serine residue located in the NTD [S(NTD)] was conserved as well as its flanking region, i.e. the proline rich motif (fig. 1A). Then the question was whether, in zfRARαs, the conserved S(NTD) could be phosphorylated even in the absence of a phosphorylatable S(LBD). Most interestingly, *in vitro*, the S(NTD) of zfRARα, was phosphorylated by the purified cdk7/cyclinH complex, as assessed by immunoblotting with antibodies recognizing specifically this phosphorylated residue (fig. 3*B*, lanes 6 and 7). No signal was obtained with zfRARα deleted for the NTD, confirming the specificity of the antibodies (fig. 3*B*, lane 8). These results were confirmed *in vivo,* with B10-tagged zfRARα over expressed in zebrafish (ZF13) or mammalian (COS-1) cells and immunoprecipitated with an antibody recognizing specifically the phosphorylated receptor (fig. 3*C*). Collectively, these results indicate that, in zebrafish, RARα can be phosphorylated at S(NTD) even in the absence of phosphorylation of the LBD. This raises the question of how the phosphorylation cascade starting at the LBD can be bypassed in zebrafish.

In zebrafish RARα, the cyclin H binding domain is more flexible than in hRARα.

Given that in hRARα, phosphorylation of S(NTD) by cdk7 relies on the binding of the associated cyclin H, we investigated whether zfRARα could interact with cyclin H despite the absence of phosphorylation of the LBD.

The ability of zfRARα to interact with cyclin H was compared to that of hRARα in *in vitro* protein-protein interaction assays. In GST pulldown assays that use non-phosphorylated bacterially expressed fusion proteins, both zfRARα-A and zfRARα-B interacted with cyclin H but more efficiently than did hRARα (fig. 4A, lanes 4 and 5 and fig. 4B). Zebrafish RARαs also interacted with cyclin H in coimmunoprecipitation experiments performed with extracts from transfected COS-1 cells (Supplemental

fig.1A). In this case the interaction was as efficient as with hRARα in line with the fact that hRARα is phosphorylated at S369 in transfected cells (Supplemental fig.1B). Similar results were obtained whatever cyclin H was from human (Supplemental fig.1A) or zefrafish (Supplemental fig. 1C). Altogether these observations suggest that the conformation of zfRARα favors cyclin H binding.

Given that the efficiency of cyclin H binding to hRARα relies on the conformational features of L8-9 and the N-terminal tip of H9 (Bour et al. 2005), we compared this domain to that of zfRARα-A and zfRARα-B (fig. 1*A*). The overall sequence of the cyclin H binding domain is well conserved but the proline residue located at the N-terminal tip of H9 in hRARα (P345), is replaced by an alanine and a serine in zfRARα-A and zfRARα-B respectively. Similarly the methionine residue found in helix 9 at position 350 is replaced by a valine or a glutamic acid in zfRARαs. Altogether, these observations pinpoint some specific mutations that may be functionally relevant to affect the conformation of zfRARα L8-9 in order to favor cyclin H binding.

This led us to explore the conformational dynamics of zfRARαs in molecular dynamics simulations carried out with the apo forms that best match the experimental GST-pulldown conditions (i.e. in the absence of RA and of a coactivator peptide and with H12 extended into solution). Comparison of the RMSfl calculated from these simulations showed that both zfRARα-A and zfRARα-B exhibited an increased flexibility of L8-9 compared to hRARα also in the apo form (fig. 4C). Such results indicate that in zebrafish RARα, L8-9 is natively more flexible than in the human counterpart.

Then, given that in hRARα, substitution of P345 and D346 with a glycine and an alanine respectively, significantly increases cyclin H binding (fig. 4*A*, lane 7 and (Bour et al. 2005)), we analyzed the consequences of these changes in MD simulations. As shown in Figure 4D, the hRARαP345G/D346A mutant in the apo form showed an increased flexibility of L8-9 compared to WT RARα. Finally, the single substitution of P345 with an alanine in hRARα (as in zfRARα-A) was sufficient to increase the flexibility of L8-9 (Supplemental fig. S2). Collectively these results highlight the importance of the amino acid sequence in the flexibility of the cyclin H binding domain.

During vertebrate evolution, S(NTD) is conserved, but not S(LBD).

From the comparison of human and zebrafish RARαs, one can suggest that the serine located in the NTD would be conserved across vertebrates, while it would not be the case for the serine in the LBD. Therefore we investigated whether there is a constraint on these two RARα phosphorylation sites during evolution. Sequence comparison and prediction of ancestral sequences at all nodes of the chordate RAR tree were performed.

Figure 5 provides a global overview of the evolution of S(NTD) and S(LBD) in all known full length and functional chordate RARαs with available complete sequences, from invertebrate chordates such as amphioxus (*Branchiostoma floridae)* and ascidian tunicates (*Ciona savignyi*) through basal vertebrates such as lampreys (*Lethenteron japonicum*) and teleost fishes (e.g. *Danio rerio*) to mammals.

It appeared that S(NTD) is strictly conserved in all available complete chordate sequences (fig. 5), even in amphioxus (*Branchiostoma floridae)*. Note however that for some species such as *Eptatretus burgeri, Mordacia mordax, Callorhinchus callorynchus and Lepisosteus platyrincus,* the RARα sequences are incomplete (hyphens in fig.5), making difficult the introduction of these species in our evolutionary study. The situation was very similar for the RARβ and RARγ paralogs (fig.5). Interestingly, the flanking region, i.e. the proline-rich motif, was also conserved (fig.6). This indicates that S(NTD) has been under a high selective pressure through chordate evolution.

In contrast, S(LBD) was not present in RAR from cephalochordates (*Branchiostoma floridae)* and urochordates *(Ciona savignyi* and *Polyandrocarpa misakensis*) (fig.5) despite the conservation of the flanking arginine-lysine rich motif (fig. 6 and supplemental fig. S3). It was not present either in RARα from most vertebrates (teleost fish, amphibians, birds), (fig. 5). Instead, aspartic acid, glutamic acid, asparagine or histidine residues were found. Most interestingly, the presence of a serine in L9-10 seems to be specific to the main clades of mammals with an exception in the case of *Anolis carolinensis* (fig. 5 and supplemental fig. S3). The situation was very similar for RARγ (fig. 5). However, in the case of RARβ a serine was present in the LBD not only in mammals but also in teleost fishes and in *Lepisosteus platyrhincus* (fig. 5).

Given this complex pattern, it was difficult to infer directly (using simple parsimony reasoning) which aminoacid was at the position of S(LBD) before the two duplications (A1 and A2 in fig.7) that led to RARα in gnathostomes. Thus, we

reconstructed the ancestors using maximum likelihood. This allowed us to propose an evolutionary scenario (fig. 7) in which the inferred ancestor harbors an asparagine in L9-10. Then in vertebrates, this asparagine is replaced by a serine in teleost fishes RARβ as well as in mammalian RARα, RARβ and RARγ. This suggests that, late during vertebrate evolution, a change in selective pressure allowed the acquisition of an easily phosphorylatable residue in L9-10.

Note that S(NTD) was also found in TR (fig. 7), a nuclear receptor belonging to the same NR subfamily. Though part of a distinct motif, this serine is known as a functional phosphorylation site (Glineur et al. 1990), corroborating the ancient high selection constraint acting on this residue.

DISCUSSION

Protein phosphorylation is crucial for the regulation of many cellular events and it has been hypothesized that it would serve as a transcriptional clock, orchestrating rapid and dynamic exchanges of coregulators so that at the end, the right proteins are present with the right activity, at the right place and at the right time (Rochette-Egly and Germain 2009; Lalevee, Ferry, and Rochette-Egly 2010). Most interestingly, similar to changes in genes cis-regulatory modules that modify specific aspects of the expression pattern of a gene without affecting the function of the encoded protein (Hoekstra and Coyne 2007), specific changes of phosphorylation processes can modify a regulatory cascade without affecting the overall function of the protein (Basu, Wang, and Alt 2008). Therefore, phosphorylation sites may be important targets of evolutionary processes. Now, with the availability of high throughput datasets, it becomes possible to examine and test experimentally the evolution of large sets of proteins and phosphorylation sites.

Human and mouse RARα are typical transcriptional regulators that are modulated by RA binding and rapid concomitant RA-induced phosphorylation cascades starting at a serine located in the LBD [S(LBD)] and ending at an other serine in the NTD [S(NTD)].

The present work indicates that there is a strong conservation of S(NTD) in all chordate RARα sequences known to date, comprising cephalochordates (amphioxus), urochordates (*Ciona*) and vertebrates. Such an evolutionary constraint correlates with the importance of this phosphorylation site for RARα binding to DNA and RARα-mediated transcription (Bruck et al. 2009; Rochette-Egly and Germain 2009). It is worth noting that S(NTD) is also highly conserved in the other RAR paralogs, RARγ and RARβ, all along the chordate phylum, confirming the functional importance of this phosphorylation site. However, out of chordates, only a few RAR sequences have been identified, one in another deuterostome such as *Strongylocentrus purpuratus* (Canestro et al. 2006; Marletaz et al. 2006) and only two in prostosomes (one in a mollusk and one in an annelid) (Campo-Paysaa et al. 2008; Albalat and Canestro 2009), and functional data are still lacking. Therefore the phylogenic coverage is not sufficient yet to generalize our data out of chordates and to protostomes on a safe basis. Finally, given the overall conservation of the flanking proline-rich motif, S(NTD) appears to be phosphorylated by similar kinases in all species. In line with this, cdk7 and cyclin H are

highly conserved and functional homologs have been described even in yeast (Damagnez, Makela, and Cottarel 1995).

The original aspect of human RARα phosphorylation resides in the fact that the cdk7 kinase involved in the phosphorylation of S(NTD), recognizes the receptor through the binding of cyclin H at a specific domain located in a disordered loop of the LBD (L8-9). This process is controlled by the phosphorylation of S(LBD), a nearby serine residue located in an other disordered loop, L9-10 (Bour et al. 2005; Gaillard et al. 2006). Due to the importance of this phosphorylation cascade, we have combined evolutionary studies with molecular dynamics computer simulations and experimental analysis, to predict phosphorylation of S(LBD), flexibility of the cyclin H binding domain, cyclin H binding and the evolution of these processes.

The present study demonstrates that in human RARα, phosphorylation of S(LBD) increases the flexibility of the nearby L8-9 involved in cyclin H binding and thereby in the phosphorylation of S(NTD), whatever RARα is under an apo or holo form. Thus one can suggest that this process cooperates with the conformational changes induced by RA-binding for hRARα transcriptional activity (Fig. 8B).

However, despite the good conservation of the flanking basic K/R-rich motif, located at the end of the highly structured helix 9, our evolutionary analysis points out that S(LBD), located in L9-10, a disordered loop, is not universally conserved throughout vertebrates. Indeed, S(LBD) is present mostly in mammalian RARα whereas an asparagine is present at this position at the basis of the chordate RAR tree, as exemplified by amphioxus. The same conclusion was made for the paralog RARγ but not for RARβ since S(LBD) was also found in teleost RARβ. From these observations two major different mechanisms of evolution can be proposed. In the first one, S(LBD) might have appeared in the three RAR paralogs at the basis of vertebrates during the second round (2R) of duplication. Then this residue might have been lost independently in the three RARs from several vertebrates and during the third round (3R) of duplication in teleost RARα and RARγ. In the second mechanism, the asparagine might change to a serine in each of the 3 mammalian RARs (α, β and γ) after the two rounds of duplications and in teleost RARβ during the third round of duplication. This latter mechanism might be the most probable one, in line with the functional role of S(LBD). Nevertheles, whatever the mechanism is, our observations suggest a convergent evolution of a similar regulatory mechanisms that occured four times independently. However as we

still have no evidence whether the phosphorylation cascade described for RARα also occurs in the context of RARγ and RARβ, it is still premature to speculate on this convergence. Nevertheless, it is worth noting that due to its unstructured nature, loop L9-10 should respond rapidly and accurately to changing environmental conditions, i.e. requirement of a phosphorylation or not (see below).

Besides, the cyclin H docking site of RARα evolved subtly in parallel to S(LBD). Indeed, in zebrafish RARα, which did not acquire S(LBD), L8-9 harbors a flexible conformation that favors cyclin H binding without any requirement for a phosphorylation process in L9-10 (fig. 8A). Interestingly, amphoxius RAR also depicts a mutation in the cyclin H binding domain (fig. 6), suggesting that the flexibility of this domain might be also increased. In contrast, in mammalian RARα, the acquisition of a serine in L9-10 is associated to a drastic reduction in the dynamics of L8-9 and in its ability to interact with cyclin H. The appearance of such a rigidity makes necessary a fine-tuned regulation by the phosphorylation of S(LBD) (fig. 8B). It is worth noting that both loops L8-9 and L9-10 correspond to disordered domains that evolve faster than ordered ones (Schaefer, Schlessinger, and Rost 2010). In line with this, phosphosites frequently appear in such disordered regions, thus facilitating the evolution of kinase signaling circuits (Beltrao et al. 2009; Holt et al. 2009; Landry, Levy, and Michnick 2009).

Thus, we believe that during evolution, a selective pressure might push for rapid changes in the disordered loops L8-9 and L9-10 of the LBD, in order to maintain in a changing environment, the phosphorylation of the NTD that is essential for RARα transcriptional activity (fig. 8). Indeed, when L8-9 lost its flexibility, there was a strong pressure for compensation, i.e. the appearance of a phosphorylatable serine in L9-10. Of note, MSK1, the kinase involved in the phosphorylation of S(LBD) is a vertebrate kinase, but an ortholog has been identified in drosophila (Jin et al. 1999), suggesting that the phosphorylation machinery predates chordates RAR diversification.

In conclusion, the present work highlights the evolutionary potential of the RARα phosphorylation network, especially at the level of the kinase-substrate interaction. As the complex combinatorial control of hRARα phosphorylation by multiple kinases is a readily evolved network, one can predict that its deregulation might be at the basis of disease. In support of such an hypothesis we have shown that in Xeroderma Pigmentosum patients, RARα is not efficiently phosphorylated by cdk7 with

characteristic downstream consequences on the expression of RAR target genes (Keriel et al. 2002). This has been correlated at least in part to the clinical abnormalities of the patients but also to their high risk of skin cancer in response to UV.

Acknowledgments

We thank Dr. Yiping Liu (Shanghai Institute for Biological Science, China) for the gift of the zebrafish cyclin H cDNA, M. Oulad Abdelghani (IGBMC) for the mouse monoclonal antibodies and members of the cell culture facilities for help. Special thanks to all teams members for fruitful discussions and suggestions and to L. Azzab (IGBMC) and A. Perret (Universite de Strasbourg) for help in the development of simulation protocols. This work was supported by funds from CNRS, INSERM, the Association pour la Recherche sur le Cancer (ARC 3169), the Agence Nationale pour la Recherche (ANR-05-BLAN-0390-02 and ANR-09-BLAN-0127-01), the Fondation pour la Recherche Médicale (DEQ20090515423) and the Institut National du Cancer (INCa-PL09-194). The Institut du Developpement et des Ressources en Informatique Scientifique (IDRIS), the Centre Informatique National de l'Enseignement Supérieur (CINES) and the Centre d'Etude du Calcul Parallèle de Strasbourg (Université de Strasbourg) are aknowledged for generous allocations of computer time.

References

Albalat, R., and C. Canestro. 2009. Identification of Aldh1a, Cyp26 and RAR orthologs in protostomes pushes back the retinoic acid genetic machinery in evolutionary time to the bilaterian ancestor. Chem Biol Interact **178**:188-196.

Ali, S., Y. Lutz, J. P. Bellocq, M. P. Chenard-Neu, N. Rouyer, and D. Metzger. 1993. Production and characterization of monoclonal antibodies recognising defined regions of the human oestrogen receptor. Hybridoma **12**:391-405.

Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. L. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science **282**:1711-1714.

Bastien, J., S. Adam-Stitah, T. Riedl, J. M. Egly, P. Chambon, and C. Rochette-Egly. 2000. TFIIH interacts with the retinoic acid receptor gamma and phosphorylates its AF-1-activating domain through cdk7. J Biol Chem **275**:21896-21904.

Basu, U., Y. Wang, and F. W. Alt. 2008. Evolution of phosphorylation-dependent regulation of activation-induced cytidine deaminase. Mol Cell **32**:285-291.

Beltrao, P., J. C. Trinidad, D. Fiedler, A. Roguev, W. A. Lim, K. M. Shokat, A. L. Burlingame, and N. J. Krogan. 2009. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol **7**:e1000134.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res **28**:235-242.

Bertrand, S., B. Thisse, R. Tavares, L. Sachs, A. Chaumot, P. L. Bardet, H. Escriva, M. Duffraisse, O. Marchand, R. Safi, C. Thisse, and V. Laudet. 2007. Unexpected novel relational links uncovered by extensive developmental profiling of nuclear receptor expression. PLoS Genet **3**:e188.

Bour, G., E. Gaillard, N. Bruck, S. Lalevee, J. L. Plassat, D. Busso, J. P. Samama, and C. Rochette-Egly. 2005. Cyclin H binding to the RAR{alpha} activation function (AF)-2 domain directs phosphorylation of the AF-1 domain by cyclin-dependent kinase 7. Proc Natl Acad Sci U S A **102**:16608-16613.

Bour, G., R. Taneja, and C. Rochette-Egly. 2006. Mouse Embryocarcinoma F9 cells and Retinoic Acid. A model to study the molecular mechanisms of endodermal differentiation. Pp. 211-253 *in* R. Taneja, ed. Nuclear Receptors in development. Elsevier Press Inc.

Bourguet, W., V. Vivat, J. M. Wurtz, P. Chambon, H. Gronemeyer, and D. Moras. 2000. Crystal structure of a heterodimeric complex of RAR and RXR ligand-binding domains. Mol Cell **5**:289-298.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and K. M. 1983. CHARMM: A Program for Macromolecular Energy Minimization and Dynamics Calculations. J. Comp. Chem **4**:187-217.

Bruck, N., D. Vitoux, C. Ferry, V. Duong, A. Bauer, H. de The, and C. Rochette-Egly. 2009. A coordinated phosphorylation cascade initiated by p38MAPK/MSK1 directs RARalpha to target promoters. Embo J **28**:34-47.

Brunger, A. T., and M. Karplus. 1988. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins **4**:148-156.

Campo-Paysaa, F., F. Marletaz, V. Laudet, and M. Schubert. 2008. Retinoic acid signaling in development: tissue-specific functions and evolutionary origins. Genesis **46**:640-656.

Canestro, C., J. H. Postlethwait, R. Gonzalez-Duarte, and R. Albalat. 2006. Is retinoic acid genetic machinery a chordate innovation? Evol Dev **8**:394-406.

Canutescu, A. A., A. A. Shelenkov, and R. L. Dunbrack, Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci **12**:2001-2014.

Damagnez, V., T. P. Makela, and G. Cottarel. 1995. Schizosaccharomyces pombe Mop1-Mcs2 is related to mammalian CAK. Embo J **14**:6164-6172.

de StGroth, S. F., and D. Scheidegger. 1980. Production of monoclonal antibodies: strategy and tactics. J Immunol Methods **35**:1-21.

Dehal, P., and J. L. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3**:e314.

Dolle, P. 2009. Developmental expression of retinoic acid receptors (RARs). Nucl Recept Signal **7**:e006.

Dyson, H. J., and P. E. Wright. 2005. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol **6**:197-208.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.

Escriva, H., S. Bertrand, P. Germain, M. Robinson-Rechavi, M. Umbhauer, J. Cartry, M. Duffraisse, L. Holland, H. Gronemeyer, and V. Laudet. 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet **2**:e102.

Escriva, H., N. D. Holland, H. Gronemeyer, V. Laudet, and L. Z. Holland. 2002a. The retinoic acid signaling pathway regulates anterior/posterior patterning in the nerve cord and pharynx of amphioxus, a chordate lacking neural crest. Development **129**:2905-2916.

Escriva, H., L. Manzon, J. Youson, and V. Laudet. 2002b. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol **19**:1440-1450.

Fidelak, J., S. Ferrer, M. Oberlin, D. Moras, A. Dejaegere, and R. H. Stote. 2010. Dynamic correlation networks in human peroxisome proliferator-activated receptor-gamma nuclear receptor protein. Eur Biophys J **39**:1503-1512.

Fujiwara, S. 2006. Retinoids and nonvertebrate chordate development. J Neurobiol **66**:645-652.

Gaillard, E., N. Bruck, Y. Brelivet, G. Bour, S. Lalevee, A. Bauer, O. Poch, D. Moras, and C. Rochette-Egly. 2006. Phosphorylation by Protein Kinase A potentiates retinoic acid repeptor a activity by means of increasing interaction with and phosphorylation by cyclin H/cdk7. Proc Natl Acad Sci U S A **103**:9548-9553.

Gaillard, T., A. Dejaegere, and R. H. Stote. 2009. Dynamics of beta3 integrin I-like and hybrid domains: insight from simulations on the mechanism of transition between open and closed forms. Proteins **76**:977-994.

Germain, P., S. Kammerer, E. Perez, C. Peluso-Iltis, D. Tortolani, F. C. Zusi, J. Starrett, P. Lapointe, J. P. Daris, A. Marinier, A. R. de Lera, N. Rochel, and H. Gronemeyer. 2004. Rational design of RAR-selective ligands revealed by RARbeta crystal stucture. EMBO Rep **5**:877-882.

Germain, P., B. Staels, C. Dacquet, M. Spedding, and V. Laudet. 2006. Overview of nomenclature of nuclear receptors. Pharmacol Rev **58**:685-704.

Glineur, C., M. Zenke, H. Beug, and J. Ghysdael. 1990. Phosphorylation of the v-erbA protein is required for its function as an oncogene. Genes Dev **4**:1663-1676.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52**:696-704.

Hisata, K., S. Fujiwara, Y. Tsuchida, M. Ohashi, and K. Kawamura. 1998. Expression and function of a retinoic acid receptor in budding ascidians. Dev Genes Evol **208**:537-546.

Hoekstra, H. E., and J. A. Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution **61**:995-1016.

Holt, L. J., B. B. Tuch, J. Villen, A. D. Johnson, S. P. Gygi, and D. O. Morgan. 2009. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. Science **325**:1682-1686.

Jaillon, O., J. M. Aury, F. Brunet, J. L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biemont, Z. Skalli, L. Cattolico, J. Poulain, V. De Berardinis, C. Cruaud, S. Duprat, P. Brottier, J. P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J. N. Volff, R. Guigo, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quetier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. Roest Crollius. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature **431**:946-957.

Jin, Y., Y. Wang, D. L. Walker, H. Dong, C. Conley, J. Johansen, and K. M. Johansen. 1999. JIL-1: a novel chromosomal tandem kinase implicated in transcriptional regulation in Drosophila. Mol Cell **4**:129-135.

Kast, D., L. M. Espinoza-Fonseca, C. Yi, and D. D. Thomas. 2010. Phosphorylation-induced structural changes in smooth muscle myosin regulatory light chain. Proc Natl Acad Sci U S A **107**:8207-8212.

Keriel, A., A. Stary, A. Sarasin, C. Rochette-Egly, and J. M. Egly. 2002. XPD Mutations Prevent TFIIH-Dependent Transactivation by Nuclear Receptors and Phosphorylation of RARalpha. Cell **109**:125-135.

Klaholz, B. P., A. Mitschler, and D. Moras. 2000. Structural basis for isotype selectivity of the human retinoic acid nuclear receptor. J Mol Biol **302**:155-170.

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? Mol Biol Evol **26**:47-59.

Lalevee, S., G. Bour, M. Quinternet, E. Samarut, P. Kessler, M. Vitorino, N. Bruck, M. A. Delsuc, J. L. Vonesch, B. Kieffer, and C. Rochette-Egly. 2010. Vinexin{beta}, an

atypical "sensor" of retinoic acid receptor {gamma} signaling: union and sequestration, separation, and phosphorylation. FASEB J **24**:4523-4534.

Lalevee, S., C. Ferry, and C. Rochette-Egly. 2010. Phosphorylation control of nuclear receptors. Methods Mol Biol **647**:251-266.

Landry, C. R., E. D. Levy, and S. W. Michnick. 2009. Weak functional constraints on phosphoproteomes. Trends Genet **25**:193-197.

Lavery, D. N., and I. J. McEwan. 2005. Structure and function of steroid receptor AF1 transactivation domains: induction of active conformations. Biochem J **391**:449-464.

Linville, A., K. Radtke, J. S. Waxman, D. Yelon, and T. F. Schilling. 2009. Combinatorial roles for zebrafish retinoic acid receptors in the hindbrain, limbs and pharyngeal arches. Dev Biol **325**:60-70.

Liu, Q. Y., Z. L. Wu, W. J. Lv, Y. C. Yan, and Y. P. Li. 2007. Developmental expression of cyclin H and Cdk7 in zebrafish: the essential role of cyclin H during early embryo development. Cell Res **17**:163-173.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B **102**:3586-3616.

Mackerell, J., A. D., M. Feig, and r. Brooks, C. L. 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem **25**:1400-1415.

Mark, M., N. B. Ghyselinck, and P. Chambon. 2009. Function of retinoic acid receptors during embryonic development. Nucl Recept Signal **7**:e002.

Marletaz, F., L. Z. Holland, V. Laudet, and M. Schubert. 2006. Retinoic acid signaling and the evolution of chordates. Int J Biol Sci **2**:38-47.

Nguyen Ba, A. N., and A. M. Moses. 2010. Evolution of characterized phosphorylation sites in budding yeast. Mol Biol Evol **In press**.

Pogenberg, V., J. F. Guichou, V. Vivat-Hannah, S. Kammerer, E. Perez, P. Germain, A. R. de Lera, H. Gronemeyer, C. A. Royer, and W. Bourguet. 2005. Characterization of the

interaction between retinoic acid receptor/retinoid X receptor (RAR/RXR) heterodimers and transcriptional coactivators through structural and fluorescence anisotropy studies. J Biol Chem **280**:1625-1633.

Postlethwait, J. H., Y. L. Yan, M. A. Gates, S. Horne, A. Amores, A. Brownlie, A. Donovan, E. S. Egan, A. Force, Z. Gong, C. Goutel, A. Fritz, R. Kelsh, E. Knapik, E. Liao, B. Paw, D. Ransom, A. Singer, M. Thomson, T. S. Abduljabbar, P. Yelick, D. Beier, J. S. Joly, D. Larhammar, F. Rosa, M. Westerfield, L. I. Zon, S. L. Johnson, and W. S. Talbot. 1998. Vertebrate genome evolution and the zebrafish gene map. Nat Genet **18**:345-349.

Renaud, J. P., and D. Moras. 2000. Structural studies on nuclear receptors. Cell Mol Life Sci **57**:1748-1769.

Renaud, J. P., N. Rochel, M. Ruff, V. Vivat, P. Chambon, H. Gronemeyer, and D. Moras. 1995. Crystal structure of the RAR-gamma ligand-binding domain bound to all-trans retinoic acid. Nature **378**:681-689.

Robinson-Rechavi, M., B. Boussau, and V. Laudet. 2004. Phylogenetic dating and characterization of gene duplications in vertebrates: the cartilaginous fish reference. Mol Biol Evol **21**:580-586.

Rochette-Egly, C. 2003. Nuclear receptors: integration of multiple signalling pathways through phosphorylation. Cell Signal **15**:355-366.

Rochette-Egly, C., S. Adam, M. Rossignol, J. M. Egly, and P. Chambon. 1997. Stimulation of RAR alpha activation function AF-1 through binding to the general transcription factor TFIIH and phosphorylation by CDK7. Cell **90**:97-107.

Rochette-Egly, C., and P. Germain. 2009. Dynamic and combinatorial control of gene expression by nuclear retinoic acid receptors. Nuclear Receptor Signaling **7**:e005.

Rochette-Egly, C., M. Oulad-Abdelghani, A. Staub, V. Pfister, I. Scheuer, P. Chambon, and M. P. Gaub. 1995. Phosphorylation of the retinoic acid receptor-alpha by protein kinase A. Mol Endocrinol **9**:860-871.

Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. J Comp Phys **23**:327-341.

Sali, A., and T. L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol **234**:779-815.

Sato, Y., N. Ramalanjaona, T. Huet, N. Potier, J. Osz, P. Antony, C. Peluso-Iltis, P. Poussin-Courmontagne, E. Ennifar, Y. Mely, A. Dejaegere, D. Moras, and N. Rochel. 2010. The "phantom effect" of the rexinoid LG100754: Structural and functional insights. PloS One **5**:e15119.

Schaefer, C., A. Schlessinger, and B. Rost. 2010. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics **26**:625-631.

Schaefer, M., H. W. van Vlijmen, and M. Karplus. 1998. Electrostatic contributions to molecular free energies in solution. Adv Protein Chem **51**:1-57.

Theodosiou, M., V. Laudet, and M. Schubert. 2010. From carrot to clinic: an overview of the retinoic acid signaling pathway. Cell Mol Life Sci **67**:1423-1445.

Vucetic, Z., Z. Zhang, J. Zhao, F. Wang, K. J. Soprano, and D. R. Soprano. 2008. Acinus-S' represses retinoic acid receptor (RAR)-regulated gene expression through interaction with the B domains of RARs. Mol Cell Biol **28**:2549-2558.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-556.

Legends to Figures

Fig. 1. Schematic representation of RARα.

A. Human, mouse and zebrafish RARα modular structure with the phosphorylation sites (arrowheads) located in the N-terminal proline-rich domain [S(NTD)] and in L9-10 [S(LBD)]. B. Structure of the LBD in the presence of ligand (PDB2LBD). The serine located in loop L9-10 of RARα is shown as well as the cyclin H binding site, which encompasses loop L8-9.

Fig. 2. In hRARα, phosphorylation of S(LBD) increases the flexibility of loop L8-9: Molecular dynamic simulations.

A. Superposition of the average structures of hRARα LBD unphosphorylated (blue) and phosphorylated (red) in an holo conformation, i.e. complexed with RA and a coactivator peptide. The positions of phosphorylated S(LBD) and of L8-9 are indicated. The average structures were computed from the final 5ns of MD simulations.

B. Fluctuations of the backbone atoms of hRARα LBDs unphosphorylated (black) and phosphorylated at S(LBD) (red) in an holo conformation. Global motion was removed by reorienting trajectory frames onto H3, H5 and H10 of the initial LBD structure. Fluctuations were calculated from the final 5ns of molecular dynamics simulations and given in Å. Positions of the α-helices are schematized along the X axis and the cyclin H binding domain (L8-9) is highlighted in yellow.

C. Same as in B but with the LBDs of hRARα under the apo form, i.e. in the absence of RA and of a coactivator peptide and with H12 extended in solution. Note that in this apo form, the flexibility of the C-terminal end of the hRARα LBD is higher compared to that of the holo form shown in panel B, due to the extended conformation of H12. In contrast in the holo-form, H12 is stabilized against the core of the LBD.

Fig. 3. Comparison of h and zf RARα phosphorylation

A. In vitro phosphorylation of the hRARα and zfRARα LBDs fused to GST by MSK1 and analysis by autoradiography

B. *In vitro* phosphorylation of GST-hRARα and zfRARα (WT and ΔNTD) with cyclinH/cdk7, analyzed by immunoblotting with antibodies recognizing specifically the receptors phosphorylated at S(NTD).

C. In transfected ZF13 and COS-1 cells, B10-tagged zfRARα is phosphorylated at the N-terminal serine (S72). B10-zfRARα was immunoprecipitated with antibodies recognizing specifically the form phosphorylated at S(NTD) and immunoblotted with B10 antibodies. The lower panel corresponds to the inputs.

Fig. 4. Comparison of h and zf RARα interaction with cyclin H and flexibility of the the cyclin H binding domain

A. *In vitro*, zfRARα interacts more efficiently than hRARα with human cyclin H. The hRARαP345G/D346A mutant (hPD/GA) also interacts more efficiently.

B. Mean±SD of 5 individual experiments after quantification and normalization to cyclin H binding to hRARαWT. Significantly different data are shown by asterisks: * *p-value*<0,05, *** *p-value*<0,01

C. Fluctuations of the backbone atoms of the LBDs of hRARα (black), zfRARα-A (red) and zfRARα-B (blue) under the apo form, i.e. in the absence of RA and of a coactivator peptide and with H12 extended in solution. Fluctuations were calculated from the final 5ns of the molecular dynamics simulations and are given in Å. Global motion was removed by reorienting trajectory frames onto H3, H5 and H10 of the initial LBD structure.

D. Fluctuations of the backbone atoms of the LBDs of hRARα WT (black) and P345G-D346A (red) under the apo form and calculated as in C.

Fig. 5. Schematic phylogenetic tree of chordates showing the evolution of RAR phosphorylation sites.

The conservation of S(NTD) and S(LBD) is represented after multiple alignment of the RAR sequences from different species. ∅ Corresponds to gene loss. Hyphens correspond to the lack of validated sequences. Sequences are named with the nomenclature code used in the nuclear receptor database NUREXBASE (http://nurexbase.prabi.fr).

Fig. 6 Alignment of the proline-rich domain, the cyclin H binding domain (Loop L8-9 and the N-terminal tip of H9) and Loop L9-10 in RARαs from different species and for which complete full length sequences are available.

The strongly conserved positions are marked over the alignments, according to the Clustal software program. "*****" indicates positions which have a single, fully conserved residue. "**:**" indicates that one of the following 'strong' groups is fully conserved: STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW. "**.**" indicates that one of the following 'weaker' groups is fully conserved: CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HFY.

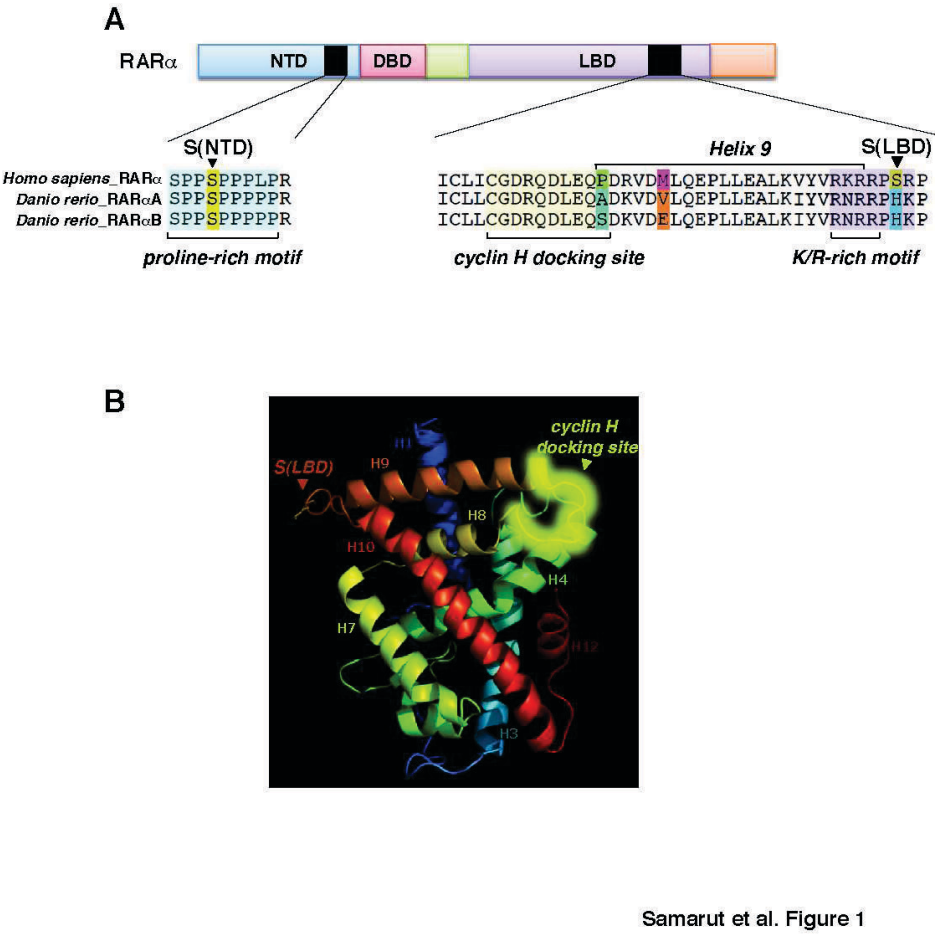
Fig. 7. Model for the evolution of RAR phosphorylation sites.

S(NTD): presence of a phosphorylatable serine residue in the N-terminal proline-rich motif of the common ancestor of RAR and TR. K/R: acquisitions of an arginine/lysine rich motif in H9 of the common ancestor of chordate RARs. N(LBD), acquisition of an asparagine in the LBD (L9-10). S(LBD): acquisition of a phosphorylatable serine in the the LBD (L9-10). A1 designate the duplication leading to the three RARs and A2 the diversification of RARα in gnathostomes.

Fig. 8. Model for the evolution of the fine-tuned phosphorylation of RARα.

A. In RARα from non-mammalian species, exemplified by zebrafish, L8-9 is naturally highly flexible as shown by the intense red halo, allowing cyclin H binding and S(NTD) phosphorylation by cdk7. RARα activity is switched on by ligand binding.

B. In mammalian RARα, L8-9 is naturally rigid making necessary a fine-tuned regulation of cyclin H recruitment by the phosphorylation of S(LBD). Subsequently RARα activity requires not only ligand binding, but also this evolved fine-tuning phosphorylation cascade.



Samarut et al. Figure 1

Figure 1
393x393mm (72 x 72 DPI)

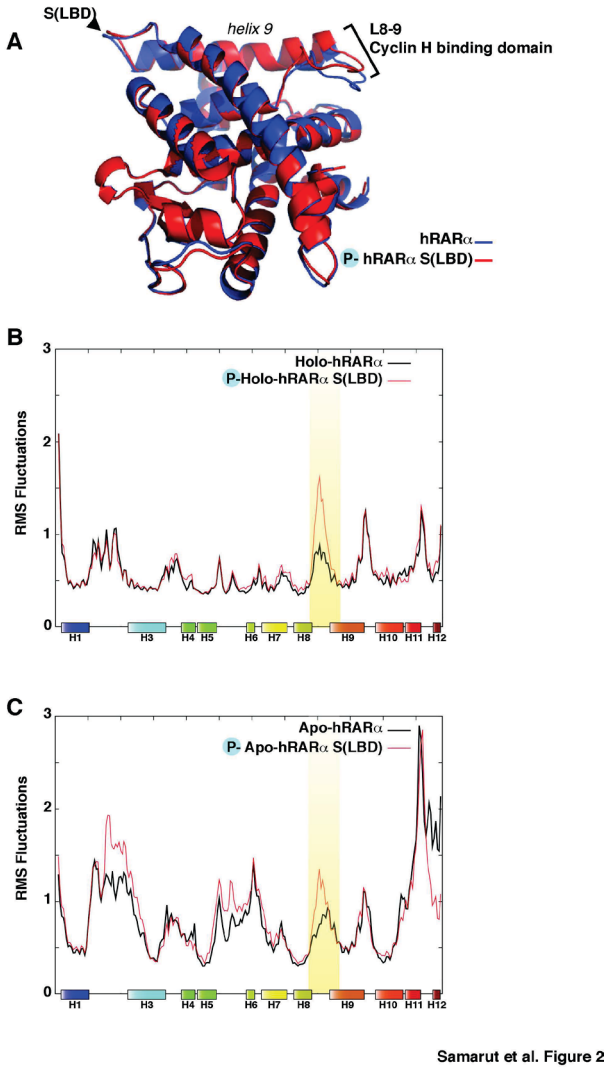
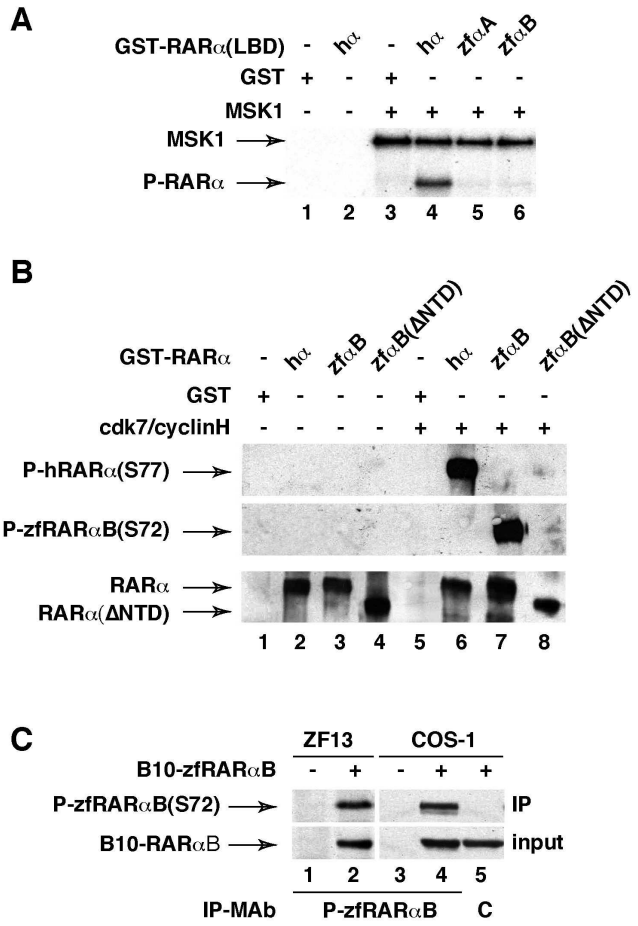


Figure 2
742x1264mm (72 x 72 DPI)



Samarut et al. Figure 3

Figure 3
603x919mm (72 x 72 DPI)

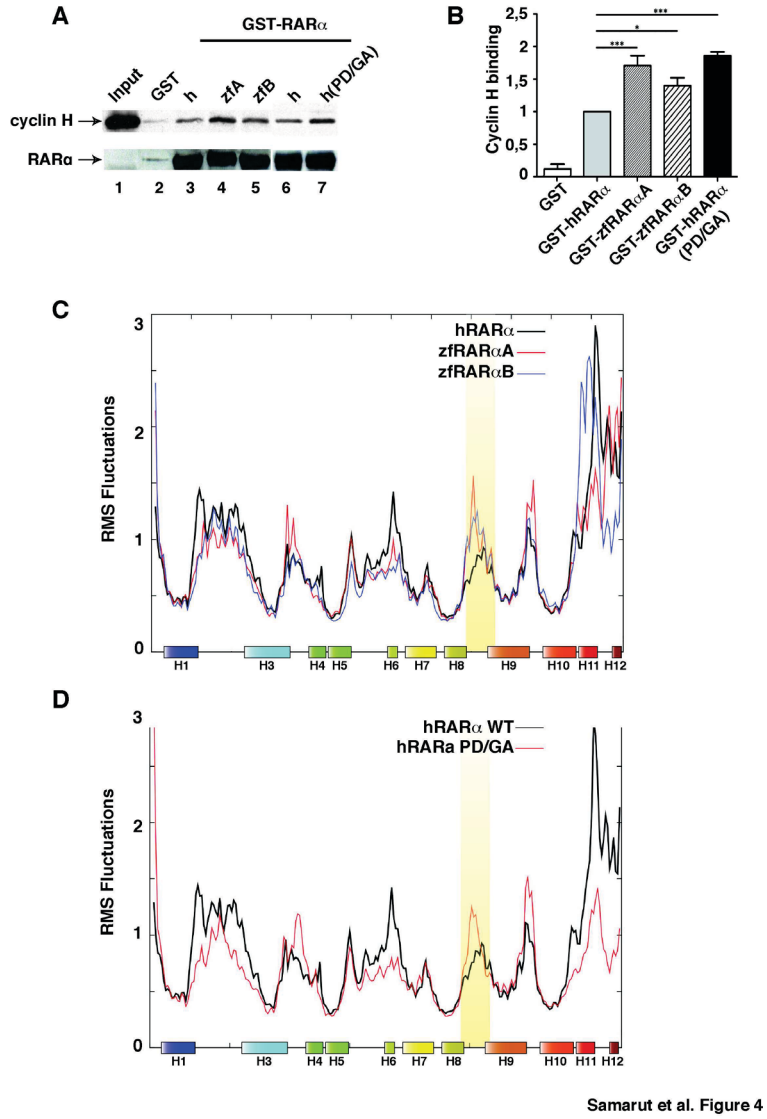
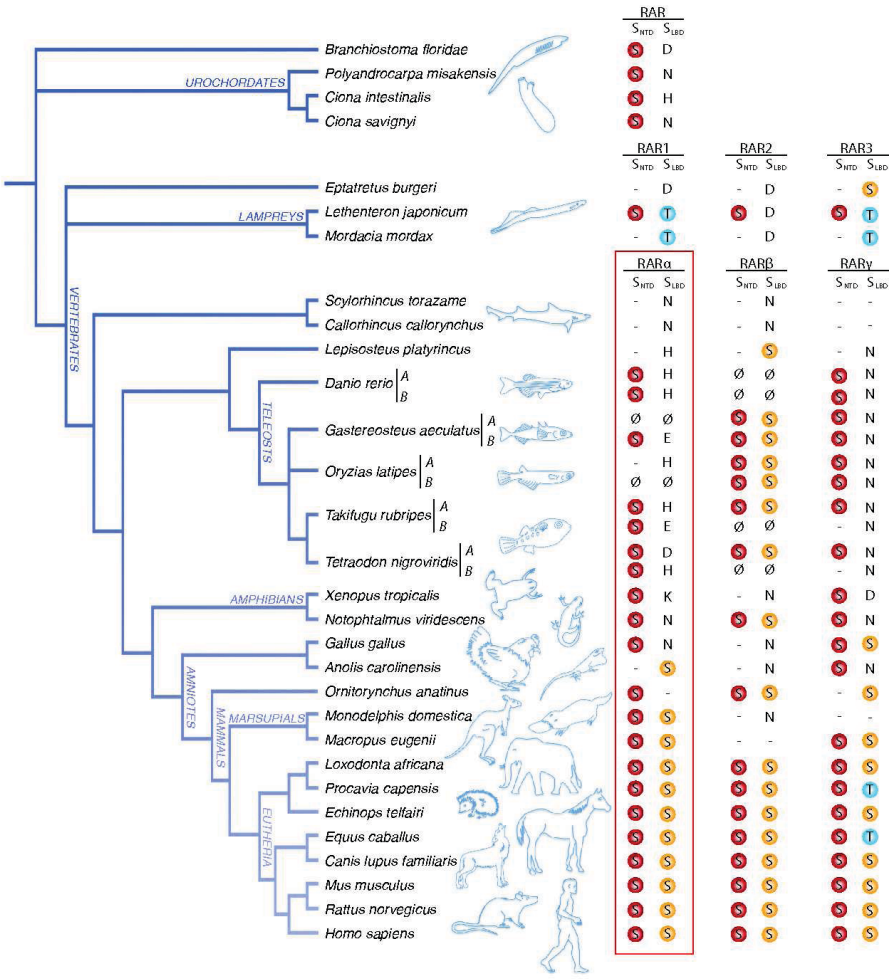
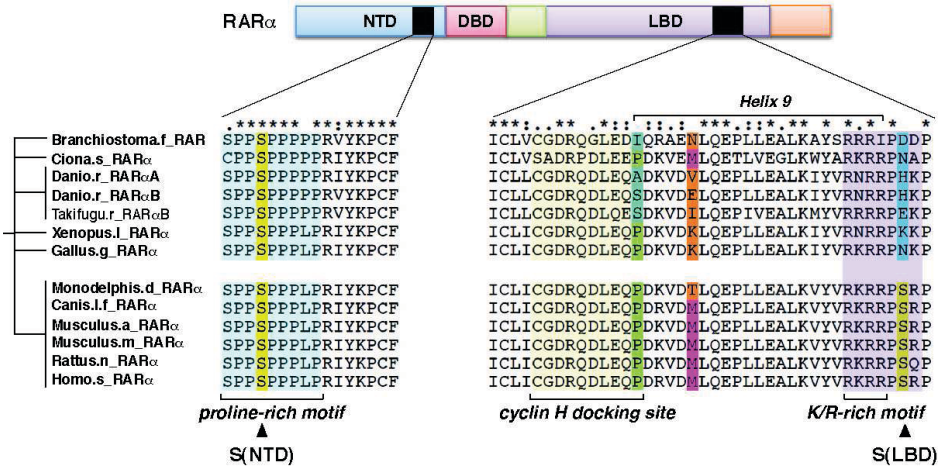


Figure 4
808x1204mm (72 x 72 DPI)



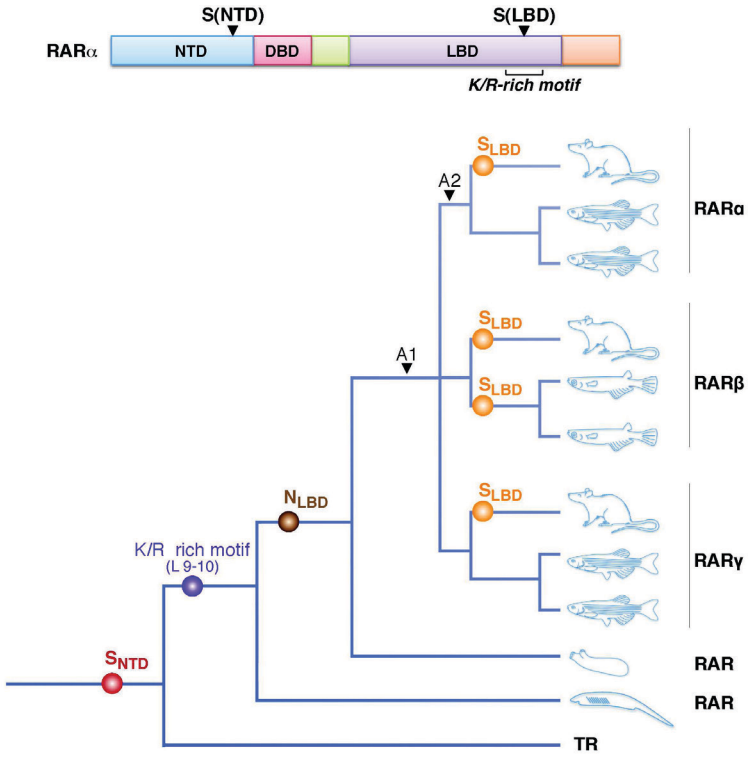
Samarut et al.
Figure 5

Figure 5
462x527mm (72 x 72 DPI)



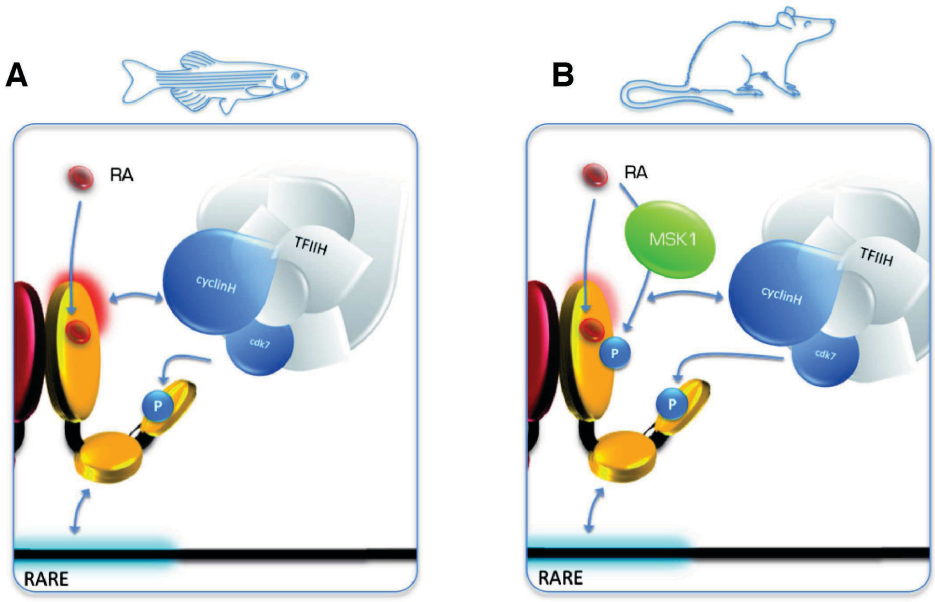
Samarut et al. Figure 6

Figure 6
425x279mm (72 x 72 DPI)



Samarut et al. Figure 7

Figure 7
894x900mm (72 x 72 DPI)



Samarut et al. Figure 8

Figure 8
742x628mm (72 x 72 DPI)

Safeguarding Developmental Potential: Convergent Functions of *Xenopus ventx* and Mammalian *Nanog*.

Pierluigi Scerbo[1,2,4], Fabrice Girardot[1,4,6], Gabriel V. Markov [1,3], Guillaume Luxardi[2], Céline Vivien[1], Barbara Demeneix[1], Laurent Kodjabachian[2,5] and Laurent Coen[1,5,6]

[1]Département Régulations, Développement et Diversité Moléculaire, UMR CNRS 7221 / USM MNHN 501, Évolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, Case 32, 75231 Paris Cedex 05, France.

[2]Institut de Biologie du Développement de Marseille Luminy, UMR CNRS 6216, Université de la Méditerranée, Case 907, 13288 Marseille Cedex 09, France.

[3]Institut de Génomique Fonctionnelle de Lyon, UMR CNRS 5242, Molecular Zoology Team, Université de Lyon, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France.

[4]These authors contributed equally to this work.

[5]Co-senior authors

[6]Corresponding authors

CONTACT

e-mail: girardot@mnhn.fr, coen@mnhn.fr; phone: +33 1 40 79 57 47; fax: +33 1 40 79 36 18

RUNNING HEAD

Convergent functions of *ventx* and *Nanog*

SUMMARY

Vertebrate development requires progressive commitment of embryonic cells into specific lineages through a continuum of signals playing off differentiation versus multipotency. In mammals, Nanog is a key transcription factor that maintains cellular pluripotency by controlling competence to respond to differentiation cues [1-6]. *Nanog* orthologs are known in most vertebrates examined to date, except in Anuran amphibians [1-2,7-12]. Here, we confirm the loss of *Nanog* in *Xenopus* and identify the *Xenopus laevis Ventral homeobox* factors (*xventxs*) [13-14] as *Nanog* counterparts in this taxon. *In silico* analyses and literature scanning reveal that *xventxs* and mammalian *Nanog* share extensive structural and functional properties. Overexpression of *xventxs* or mouse *Nanog* (*mNanog*) during *Xenopus* embryogenesis prevents multiple lineage commitment, leading to similar phenotypes. Finally, *mNanog* expression specifically rescues embryonic axis formation in *xventxs* deficient embryos. We conclude that *xventxs* play an unanticipated and evolutionary conserved role as guardians of high developmental potential and propose that *Nanog* functionally converged with *ventx* during tetrapod evolution. Indeed, human *VENTX* and *NANOG* are both expressed in pluripotent carcinoma cells [15-17] and maintain POU5F1 promoter activity in ES cells [18]. Our findings shed light on the composition and evolution of the gene regulatory network that controls pluripotency in vertebrates.

HIGHLIGHTS

- Absence of *Nanog* in *Xenopus* species is unique in vertebrates.
- *xventxs* and *Nanog* display extensive structural and functional similarities.
- Mouse *Nanog* or *xventx1/2* overexpression inhibits differentiation in *Xenopus* embryos.
- Mouse *Nanog* can specifically rescue *xventx1/2*-deficient embryos.

RESULTS

Orthologs of the NKL transcription factor Nanog have been identified in all vertebrate lineages, including non-anuran amphibians. Among vertebrates, *Nanog* is involved in the establishment of the germ-line as well as, in tetrapods but not telosteans, the maintenance of pluripotency [1-12]. No *Nanog* ortholog has been described so far in *Xenopus*; thus, either *Nanog* remains to be characterised in anurans or other(s) factor(s) must maintain the high developmental potential of uncommitted embryonic cells in this taxon [19-20]. To address the first possibility, we searched for a putative *Nanog* ortholog in *Xenopus*. *In silico* screening of sequence repositories confirmed the presence of *Nanog*-related sequences in all vertebrates, except *Xenopus*. Degenerate PCR-based approaches were also unsuccessful (data not shown). Syntheny analyses revealed that genomic regions containing *Nanog* in amniotes or teleosts are conserved in *Xenopus tropicalis* but do not contain any *Nanog*-related sequences (Figure 1A). Thus, the absence of *Nanog* from the *Xenopus* genus is probably due to secondary loss. Therefore, we tested the alternative hypothesis that other *Xenopus* transcription factors are capable of functionally replacing *Nanog*.

As Nanog belongs to the NKL subclass of homeodomain-containing proteins, we focused on this group to identify putative candidates. Phylogenetic reconstruction showed NKL families to be monophyletic, except NK4 and VENTX (Figure 1B). Surprisingly, the amphioxus Ventx orthologs [21] appear at the base of the NANOG group, suggesting that VENTX and NANOG families might be closely related (Figure 1D). Furthermore, these families share multiple features that are unique among NKLs. Notably, VENTX and NANOG are the only NKL families known to have been lost in specific vertebrate lineages: *Nanog* is absent in the *Xenopus* genus whereas rodents lack *Ventx*. Also, VENTX and NANOG are the only NKL to have numerous processed pseudogenes in the human genome (6 and 10 respectively) [22], which often correlates with expression in the germline or its embryonic precursors and is a proposed signature of genes involved in the maintenance of pluripotency

[23]. Finally, Ventx and Nanog groups have long branches when compared to other NKL families (e.g. NK1 or LBX, see Figures 1C and 1D), indicating that the HDs from these two families are less conserved among vertebrates than those from other NKLs (Table 1) and therefore might have evolved more rapidly. Testing for positive selection revealed no significant feature for *Xenopus* ventx paralogs (xventxs), while relaxation of purifying selection was detected in the Nanog group (Figure S1), suggesting that functional divergence has taken place during *Nanog* evolution after the tetrapod/teleost separation. In this respect, it is noteworthy that only the germ-line related function of *Nanog* is conserved in medaka [9-10]. These features make *Xenopus xventxs* good candidates for serving *Nanog*-like functions.

In line with this hypothesis, mammalian Nanog and *Xenopus* xventxs act as transcriptional repressors [24-26] and share striking functional similarities (summarised in Table S1). First, the orthologs of many genes regulated by *xventxs* in *Xenopus* are regulated by *Nanog* in mammals; second, *Nanog* and *xventxs* are regulated by the same signalling pathways and transcription factors; third, xventxs and Nanog interact with orthologous proteins; fourth, like Nanog, some xventx isoforms can form dimers and regulate their expression. Perhaps, the most significant parallel is that, in *Xenopus* and teleosts, endogenous ventx and pou5f1 (also known as Oct4) transcription factors interact physically and genetically during early development [27-28], as do mammalian Nanog and Pou5f1 [4]. In *Xenopus* and zebrafish these genes act in a BMP-activated pathway that is essential for proper establishment of the dorsoventral axis [27-30].

To assess if *xventxs* and *Nanog* have similar functional properties, we compared the effects of overexpression of mouse and medaka *Nanog* (*mNanog* and *OlNanog*, respectively) to combined *Xenopus ventx1.2* and *ventx2.1b* (referred to as *xventx1/2* from now on) overexpression on *Xenopus* embryonic development. The relevant mRNAs were dorsally injected at the 4-cell stage. As expected [14,25,31], *xventx1/2* overexpression led at tailbud

stage to severely ventralised phenotypes with truncated anterior structures. Remarkably, *mNanog* injection phenocopied *xventx1/2* overexpression and produced similar defects in comparable proportions. In contrast, in the same conditions, *OlNanog* overexpression only led to a modest reduction of size in a minority of injected embryos without loss of anterior structures (Figures 2A and 2B). Whole-mount *in situ* hybridization (WISH) confirmed differential activities of mouse and fish *Nanog*. As previously reported, *xventx1/2* overexpression strongly repressed the blood island marker *hba4* (also known as *alpha-T4 globin*) on the ventral side of injected embryos [32]. The same effect was observed in embryos injected with *mNanog* but not with *OlNanog* (Figure S2). In early gastrulae, WISH and RT-QPCR revealed that *xventx1/2* and *mNanog* but not *OlNanog* injection abolished expression of *gsc*, an early marker of committed dorsal mesoderm (Figures 2C, 2D and 2E). These results suggest that *mNanog* and *OlNanog* display distinct biochemical properties and consolidate our hypothesis of functional similarity for *mNanog* and *xventx1/2*.

To further evaluate this idea, the expression of developmental genes was examined by WISH and RT-QPCR at gastrula stage in embryos injected with *mNanog* or *xventx1/2*. *mNanog* strongly repressed expression of orthologs of its known targets *eomes*, *myf5*, *hhex*, *sox2*, *sox17a* and *gata6*. Furthermore, *xventx1/2* misexpression led to similar repression of all these targets, including genes not previously known to be regulated by *xventx1/2: eomes*, *gata6* and *sox17a* (Figure 2E). Conversely, *xbra*, which is known to be unaffected by *xventx1/2* overexpression [25] was also unaffected by *mNanog* overexpression (Figure 2E). Thus, *xventx1/2* and *mNanog* overexpression produce similar morphological and molecular defects, suggesting shared transcriptional properties.

These data suggest that similar to *mNanog*, *xventx1/2* could be involved in the maintenance of developmental potential. In line with this hypothesis, *xventxs* are known to inhibit differentiation in neurectoderm and mesoderm [25,27,31-32]. To further assess if

xventxs restrain commitment in *Xenopus*, we focused on the epidermis, a tissue derived from the gastrula ventral ectoderm, which expresses *xventx1/2* and harbours an elevated developmental potential. We injected *xventx1/2* or *mNanog* mRNAs at the 16-cell stage in one AB4 blastomere fated to give rise only to epidermis, then analysed gene expression at the onset of gastrulation. Both *mNanog* and *xventx1/2* strongly activated expression of *foxi1e*, an uncommitted ectoderm marker and repressed the committed epidermal marker *xk81* (Figure 2F). This result indicates that *mNanog* and *xventx1/2* do not favour, but rather impede, epidermal differentiation. Overall, *xventx1/2* and *mNanog* repress the expression of committed tissue markers from the three germ layers (*e.g. xk81*, *gsc*, *hba4*, *hhex*, *sox17a*), regardless of their ventral or dorsal character. In contrast, *xventx1/2* and *mNanog* overexpression has no effect or increases expression of uncommitted tissue markers such as *xbra* and *foxi1e* (Figures 2E and 2F). Taken together, the above data suggest that the earliest role of *xventx1/2* is not to ventralise the embryo, but rather to prevent premature differentiation, similar to *Xenopus pou5f1s* [33]. Thus, *xventx1/2* may act as guardians of high developmental potential in *Xenopus*, as does *Nanog* in amniotes [1-8].

We thus tested whether *mNanog* could functionally replace *xventx1/2* by evaluating whether *mNanog* could rescue morphological and molecular deficiencies in *xventx1/2* knockdowned embryos. Morpholino oligonucleotides (MOs) directed against *xventx1* and *xventx2* pseudoalleles [29] or control MO were injected radially at the 2-cell stage, followed by radial injections of *mNanog* mRNA at the 4-cell stage (Figure 3). Rescue controls were either mock radial injections or, to evaluate the specificity of *mNanog*, injections with mRNA coding for another ventralising NKL transcription factor, msx1 [34]. Radial injections of control MO+*mNanog* and control MO+*msx1* led to a high proportion of ventralised embryos (Figures 3A and 3B), as seen for dorsal injections of *mNanog* alone (Figures 2A and 2B). As described [29], *xventx1/2* MOs-mediated inactivation led to a high proportion of strongly

dorsalised embryos. Quite remarkably, almost 50% of embryos injected with *xvenxt1/2* MOs+*mNanog* were totally rescued at tailbud stage, showing normal elongated morphology (Figures 3A and 3B). WISH analysis revealed that axial structures such as the notochord (*shh*), ventral blood island (*hba4*), spinal cord (*hoxb9*) and brain (*six6*, *egr2*; also known as *optx2* and *krox-20*, respectively) were recovered in these embryos. In contrast, *msx1* overexpression did not permit recovery of normal morphology in *xventx1/2* morphants. Embryos either remained dorsalised or were ventralised by *msx1*, suggesting that msx1 does not control the same targets as xventx1/2. This possibility was further evaluated by RT-QPCR analyses in gastrulae. *sox17a* and *gsc*, activated by *xventx1/2* knockdown, returned to normal levels when *mNanog* was added and were further repressed by *msx1* (Figure 3C). However, *xk81* was repressed by *mNanog* but not by *msx1*, consistent with the activator role of msx1 on the epidermal programme [34]. These results demonstrate that *mNanog* is able to substitute for *xventx1/2* in *Xenopus* development and that this effect is specific, since it is not an attribute of all ventralising factors, as shown for *msx1*.

DISCUSSION

Nanog was first identified in mammals as essential for early embryonic development and germ-line establishment, being required to restrain premature differentiation of embryonic stem cells [1-3]. *Nanog* activity protects undifferentiated cells against the differentiation-inducing effects of extracellular signals and transcriptional noise [4-6]. *Nanog*, initially thought to be a mammalian-specific gene, has orthologs in most vertebrate species including birds, teleosts and non-anuran amphibians [7-12]. Among amniotes, *Nanog* is functionally conserved. The chick ortholog, *cNanog*, expressed in the developing germ-line and involved in maintaining undifferentiated pluripotent embryonic stem cells, can rescue *Nanog* loss-of-function in mouse ES cells [7-8]. In contrast, while the medaka ortholog, *OlNanog*, is also essential for early development and seems to share germ-line related

functions of its amniotes counterparts, it does not regulate expression of the orthologs of mammalian *Nanog* targets, nor does it control developmental potential and cell lineage decisions in early embryos [9-10]. In line with these data, we show here that overexpression of *mNanog* and *OlNanog* have different effect on *Xenopus laevis* development. While *mNanog* displays marked ventralising activity, *OlNanog* seems to have much more limited effects, notably it does not repress the expression of *gsc*. *Nanog* orthologs were recently identified in urodele amphians [11-12]. Axolotl *AxNanog* maintains pluripotency in *Nanog*-deficient mouse ES cells, suggesting that this factor controls developmental potential across tetrapods [12]. However, no *Nanog* ortholog has been identified in anuran amphibians so far. Nevertheless, in all amphibians, uncommitted embryonic cells maintain high developmental potential until the onset of gastrulation [19-20], similar to *Nanog*-expressing epiblastic cells in amniote embryos [35]. Interestingly, in anurans as in all tetrapods, these uncommitted cells express orthologs of *Pou5f1* [8,12,33], a key *Nanog* partner in the mammalian pluripotency network [4]. Thus, current evidence argues for the conservation of this network in all vertebrates. Based on phylogenetic, structural and functional data, we propose that this network mobilises *ventxs* factors in the place of *Nanog* in the *Xenopus* taxon and perhaps also in teleosts.

Our data imply that *xventx1/2* differ from other ventral regulators, as they seem to control the rate of differentiation of early embryonic cells and not strictly their positional identity. Indeed, in early embryos, *xventx1/2* do not directly promote ventrocaudal fates, as shown here for gastrula epidermis (Figure 2E) and elsewhere for the tailbud blood lineage [32]. Rather, they prevent precocious commitment and differentiation, indirectly affecting embryonic axes establishment and patterning. *xventx1/2* may maintain early embryonic cells in an undetermined state and limit their competence to respond to differentiation-inducing signals, as for maintenance of pluripotent cells by *Nanog* in mammalian embryos [1-3]. It is

important to note that active clearance of xventx proteins coincides with loss of multipotency at mid-gastrula stages [36]. Furthermore, numerous common *xventxs* and *Nanog* regulators, transcriptional targets and interactors are known to be involved in dorsoventral patterning during *Xenopus* and teleost embryogenesis and in cell lineage decisions in mammals (Table S1). Our work thus provides novel insights linking control of cell commitment and embryonic axis patterning and is coherent with recent experimental and theoretical works, in which the dorso-ventral (or rostro-caudal in the revised model) genetic system is reinterpreted as a regulator of timing of cell commitment throughout the embryo, thus indirectly affecting axis patterning [37].

In our re-interpretation of their role, *xventxs* may control the progressive allocation of embryonic cells to the developing body axis. Loss of *xventx1/2* abrogates inhibition of differentiation and most cells precociously adopt the same positional identities as *xventxs* negative cells in the embryo: dorsal and anterior [29] (see also Figures 3A-C). Consequently, the pool of cells available to build posterior territories is depleted, resulting in minute trunk-tail structures. Consistently, correct development can be recovered by the concomitant depletion of commitment factors normally repressed by *xventxs*, like *gsc* [29]. Conversely, ectopic activity of *xventx1/2* represses early commitment factors and causes the depletion of dorso-anterior territories [14,25,31] (see also Figures 2A-D). This proposed role of *xventxs* in the maintenance of developmental potential arose from their functional similarity with *mNanog*. In turn, *mNanog* is primarily characterised as a central factor in the mammalian pluripotency network [4] but not known to participate in the dorsal-ventral network. We surmise that these networks largely overlap, which may have marked consequences in stem cell biology.

Other amphibians possess *Nanog* orthologs [11-12], raising the question of whether the role of *ventxs* in developmental potential maintenance is ancestral or is an innovation

specific to *Xenopus*. We did not detect any positive selection or relaxation of purifying selection in the *Xenopus xventxs* branches, arguing against a functional shift during *ventx* evolution. Conversely, the relaxation observed in the *Nanog* branch suggests that this gene functionally diverged since the tetrapod/teleostean separation. Our results thus support an ancestral role of *ventxs* role in maintenance of developmental potential. However, we cannot rule out that this role is specific to *Xenopus*, since saturation can obliterate signals concerning ancient events, especially if only few sequences are available. Nevertheless, functional data strongly support the ancestrality of *ventxs* involvement in this process, since teleost and *Xenopus ventxs* serve the same function [28,30,38]. Furthermore, medaka and amniote *Nanog* orthologs have different functions during early development [9-10] and we show here that their ectopic expression has different effects on *Xenopus* development. Therefore, it seems that during evolution, *Nanog* functionally converged with the *ventxs* in the amniote lineage, rather than *ventxs* with *Nanog* in the *Xenopus* lineage. This leads us to hypothesize that a single gene regulatory network (GRN), comprising *ventx* and *pou5f1* was present in the common ancestor of vertebrates. In this scenario, the ancestral GRN restrained differentiation and secondarily co-opted *Nanog,* when it acquired functional redundancy with *ventxs*. Functional redundancy would explain the loss of *Nanog* in *Xenopus* and of *ventx* in rodents. Functional data concerning amniotes *ventx* genes is scarce, probably because they are absent from the genome of the mouse, the main experimental model in this taxon. However, the human *ventx* ortholog (*VENTX*) located next to the stem-cell marker *UTF-1*, shares features with its counterparts in *Xenopus* and fish [15]. Both human and *Xenopus ventx* orthologs block WNT signalling [39] and human *VENTX* display ventralising activity in zebrafish embryos [15]. As mentioned earlier, *VENTX* retropseudogenes are unusually frequent in the human genome [22], a feature that is proposed to be a specific signature of genes involved in pluripotency maintenance such as *POU5F1* and *NANOG* [23]. In line with this idea, *VENTX*

is co-expressed with *NANOG* and *POU5F1* in pluripotent-embryonal carcinomas [17], a subtype of human male germ cell tumours constituted of cells highly similar to early zygotic and ES cells [16]. Furthermore, all these genes are strongly down-regulated when tumour cell differentiation is forced *in vitro* [17]. Finally, a genome-wide RNA interference screen has recently shown that in human ES cells, *VENTX* or *NANOG* knockdown results in reduced expression of a POU5F1-GFP reporter construct in a modest but comparable way (see Supplemental Information in [18]). Intriguingly, knockdown of other major pluripotency regulators such as *SOX2* and *KLF4* show even less pronounced effects. Functional redundancy with *SOX2* and *KLF4* relatives might explain these weak effects [40,41]. Wether such redundancy also exists between human *NANOG* and *VENTX* is an important issue for future work.

In summary, we confirm that *Nanog* has been lost secondarily in the *Xenopus* lineage and we identify the *Xenopus ventral homeobox* genes (*xventxs*) as *Nanog* counterparts in this taxon. Overexpression of *mNanog* or *xventx1/2* in frog embryos prevents lineage decisions and causes similar phenotypes. Furthermore, overexpressed *mNanog* can specifically rescue *xventx1/2* deficient embryos. In conclusion, we propose that *ventx* genes are novel guardians of high developmental potential, functionally conserved among vertebrates, including human. Our re-interpretation of *ventxs* function offers new insights linking cell commitment and axis patterning and improves our understanding of the composition and evolution of the gene regulatory network that controls pluripotency in vertebrates.

ACKNOWLEDGEMENTS

We are grateful to V. Thomé for excellent technical assistance, to J.L. Mullor for providing the pCS2+MT-Ol-Nanog plasmid as well as to Pr P.W. Holland and T. Butts for kindly sharing sequence data that proved instrumental in the initial stages of this work. This work was funded by grants to B.D. (EU grant Crescendo and PNR 2010) and L.K. (ANR) and PhD grants to P.S. (French Ministry of Research), G.V.M. (French Ministry of Research and FRM) and G.L. (French Ministry of Research and ARC). The authors declare no competing financial interests.

REFERENCES

1. Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631-642.
2. Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643-655.
3. Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230-1234.
4. Chambers, I., and Tomlinson, S.R. (2009). The transcriptional foundation of pluripotency. Development *136*, 2311-2322.
5. Silva, J., Nichols, J., Theunissen, T.W., Guo, G., van Oosten, A.L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. Cell *138*, 722-737.
6. Kalmar, T., Lim, C., Hayward, P., Munoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol *7*, e1000149.
7. Canon, S., Herranz, C., and Manzanares, M. (2006). Germ cell restricted expression of chick Nanog. Dev Dyn *235*, 2889-2894.
8. Lavial, F., Acloque, H., Bertocchini, F., Macleod, D.J., Boast, S., Bachelard, E., Montillet, G., Thenot, S., Sang, H.M., Stern, C.D., Samarut, J., and Pain, B. (2007). The Oct4 homologue PouV and Nanog regulate pluripotency in chicken embryonic stem cells. Development *134*, 3549-3563.

9. Camp, E., Sanchez-Sanchez, A.V., Garcia-Espana, A., Desalle, R., Odqvist, L., Enrique O'Connor, J., and Mullor, J.L. (2009). Nanog regulates proliferation during early fish development. Stem Cells *27*, 2081-2091.

10. Sanchez-Sanchez, A.V., Camp, E., Leal-Tassias, A., Atkinson, S.P., Armstrong, L., Diaz-Llopis, M., and Mullor, J.L. (2010). Nanog regulates primordial germ cell migration through Cxcr4b. Stem Cells *28*, 1457-1464.

11. Maki, N., Suetsugu-Maki, R., Tarui, H., Agata, K., Del Rio-Tsonis, K., and Tsonis, P.A. (2009). Expression of stem cell pluripotency factors during regeneration in newts. Dev Dyn *238*, 1613-1616.

12. Dixon, J.E., Allegrucci, C., Redwood, C., Kump, K., Bian, Y., Chatfield, J., Chen, Y.H., Sottile, V., Voss, S.R., Alberio, R., and Johnson, A.D. (2010). Axolotl Nanog activity in mouse embryonic stem cells demonstrates that ground state pluripotency is conserved from urodele amphibians to mammals. Development *137*, 2973-2980.

13. Papalopulu, N., and Kintner, C. (1996). A Xenopus gene, Xbr-1, defines a novel class of homeobox genes and is expressed in the dorsal ciliary margin of the eye. Dev Biol *174*, 104-114.

14. Gawantka, V., Delius, H., Hirschfeld, K., Blumenstock, C., and Niehrs, C. (1995). Antagonizing the Spemann organizer: role of the homeobox gene Xvent-1. Embo J *14*, 6268-6279.

15. Moretti, P.A., Davidson, A.J., Baker, E., Lilley, B., Zon, L.I., and D'Andrea, R.J. (2001). Molecular cloning of a human Vent-like homeobox gene. Genomics *76*, 21-29.

16. Clark, A.T. (2007). The stem cell identity of testicular cancer. Stem Cell Rev *3*, 49-59.

17. Korkola, J.E., Houldsworth, J., Chadalavada, R.S., Olshen, A.B., Dobrzynski, D., Reuter, V.E., Bosl, G.J., and Chaganti, R.S. (2006). Down-regulation of stem cell

genes, including those in a 200-kb gene cluster at 12p13.31, is associated with in vivo differentiation of human male germ cell tumors. Cancer Res *66*, 820-827.

18. Chia, N.Y., Chan, Y.S., Feng, B., Lu, X., Orlov, Y.L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.S., Huss, M., Soh, B.S., Kraus, P., Li, P., Lufkin, T., Lim, B., Clarke, N.D., Bard, F., and Ng, H.H. (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature.

19. Snape, A., Wylie, C.C., Smith, J.C., and Heasman, J. (1987). Changes in states of commitment of single animal pole blastomeres of Xenopus laevis. Dev Biol *119*, 503-510.

20. Kaneko, K., Sato, K., Michiue, T., Okabayashi, K., Ohnuma, K., Danno, H., and Asashima, M. (2008). Developmental potential for morphogenesis in vivo and in vitro. J Exp Zool B Mol Dev Evol *310*, 492-503.

21. Kozmik, Z., Holland, L.Z., Schubert, M., Lacalli, T.C., Kreslova, J., Vlcek, C., and Holland, N.D. (2001). Characterization of Amphioxus AmphiVent, an evolutionarily conserved marker for chordate ventral mesoderm. Genesis *29*, 172-179.

22. Holland, P.W., Booth, H.A., and Bruford, E.A. (2007). Classification and nomenclature of all human homeobox genes. BMC Biol *5*, 47.

23. Pain, D., Chirn, G.W., Strassel, C., and Kemp, D.M. (2005). Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. J Biol Chem *280*, 6265-6268.

24. Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S.Y., Qin, J., Wong, J., Cooney, A.J., Liu, D., and Songyang, Z. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. Nat Cell Biol *10*, 731-739.

25. Onichtchouk, D., Glinka, A., and Niehrs, C. (1998). Requirement for Xvent-1 and

Xvent-2 gene function in dorsoventral patterning of Xenopus mesoderm. Development *125*, 1447-1456.

26. Friedle, H., Rastegar, S., Paul, H., Kaufmann, E., and Knochel, W. (1998). Xvent-1 mediates BMP-4-induced suppression of the dorsal-lip-specific early response gene XFD-1' in Xenopus embryos. Embo J *17*, 2298-2307.

27. Cao, Y., Knochel, S., Donow, C., Miethe, J., Kaufmann, E., and Knochel, W. (2004). The POU factor Oct-25 regulates the Xvent-2B gene and counteracts terminal differentiation in Xenopus embryos. J Biol Chem *279*, 43735-43743.

28. Reim, G., and Brand, M. (2006). Maternal control of vertebrate dorsoventral axis formation and epiboly by the POU domain protein Spg/Pou2/Oct4. Development *133*, 2757-2770.

29. Sander, V., Reversade, B., and De Robertis, E.M. (2007). The opposing homeobox genes Goosecoid and Vent1/2 self-regulate Xenopus patterning. Embo J *26*, 2955-2965.

30. Flores, M.V., Lam, E.Y., Crosier, K.E., and Crosier, P.S. (2008). Osteogenic transcription factor Runx2 is a maternal determinant of dorsoventral patterning in zebrafish. Nat Cell Biol *10*, 346-352.

31. Ladher, R., Mohun, T.J., Smith, J.C., and Snape, A.M. (1996). Xom: a Xenopus homeobox gene that mediates the early effects of BMP-4. Development *122*, 2385-2394.

32. Kumano, G., Belluzzi, L., and Smith, W.C. (1999). Spatial and temporal properties of ventral blood island induction in Xenopus laevis. Development *126*, 5327-5337.

33. Morrison, G.M., and Brickman, J.M. (2006). Conserved roles for Oct4 homologues in maintaining multipotency during early vertebrate development. Development *133*, 2011-2022.

34. Suzuki, A., Ueno, N., and Hemmati-Brivanlou, A. (1997). Xenopus msx1 mediates epidermal induction and neural inhibition by BMP4. Development *124*, 3037-3044.

35. O'Farrell, P.H., Stumpff, J., and Su, T.T. (2004). Embryonic cleavage cycles: how is a mouse like a fly? Curr Biol *14*, R35-45.

36. Zhu, Z., and Kirschner, M. (2002). Regulated proteolysis of Xom mediates dorsoventral pattern formation during early Xenopus development. Dev Cell *3*, 557-568.

37. Lane, M.C., and Sheets, M.D. (2006). Heading in a new direction: implications of the revised fate map for understanding Xenopus laevis development. Dev Biol *296*, 12-28.

38. Imai, Y., Gates, M.A., Melby, A.E., Kimelman, D., Schier, A.F., and Talbot, W.S. (2001). The homeobox genes vox and vent are redundant repressors of dorsal fates in zebrafish. Development *128*, 2407-2420.

39. Gao, H., Le, Y., Wu, X., Silberstein, L.E., Giese, R.W., and Zhu, Z. (2010). VentX, a novel lymphoid-enhancing factor/T-cell factor-associated transcription repressor, is a putative tumor suppressor. Cancer Res *70*, 202-211.

40. Guth, S.I., and Wegner, M. (2008). Having it both ways: Sox protein function between conservation and innovation. Cell Mol Life Sci *65*, 3000-3018.

41. Jiang, J., Chan, Y.S., Loh, Y.H., Cai, J., Tong, G.Q., Lim, C.A., Robson, P., Zhong, S., and Ng, H.H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat Cell Biol *10*, 353-360.

FIGURE LEGENDS

Figure 1: Syntheny of *Nanog* loci and phylogenic reconstruction of the NKL group homeodomains relationships using Maximum Likelihood. (A) Syntheny analysis suggests that *Nanog* has been lost in the *Xenopus* lineage. In all amniotes (represented here by the mouse) and teleosts (represented here by the medaka), *Nanog* orthologs (in red) are found in different but conserved regions of the genome (top and bottom panels, respectively). While these regions can be readily identified in the *Xenopus tropicalis* genome scaffolds, none of these contain any *Nanog*-related sequence (a more detailed representation including more species is available upon request). Note that comparative mapping indicates that, in axoltl (not shown), the *AxNanog* locus lies in the conserved amniote synthenic region [12]. (B) Global view of an unrooted neighbour-joining tree obtained with the homeodomain sequences of all known NKL members found in the genomes of the fly, amphioxus and a representative selection of vertebrates. NKL families are highlighted in different shades of grey except for NANOG (red) and VENTX (blue). Relationships between NKL families remain elusive; however all are monophyletic and well supported by bootstrap analysis with three exceptions: the NK4 (paraphyletic) VENTX (polyphyletic) and NANOG (monophyletic, but poorly supported, boostrap: 53,1%). (C) Close-up of the region of the tree where most VENTX orthologs are found. (D) Close-up of the region of the tree containing the monophyletic NANOG group. Note that amphioxus VENTX homeodomains (VENT1 *Branchiostoma floridae* and VENT2 *Branchiostoma floridae*) are found at the root of the NANOG subtree, the resulting topology fitting with chordate phylogeny. However, this association is not supported by bootstrap analysis (bootstrap: 21,5%) and the interpretation of amphioxus VENTXs as NANOG orthologs is at odds with the literature [21]. Both NANOG and VENTX groups have longer branches than typical NKL-class members (e.g. NK1 and LBX groups on panels C and D, see also Table 1 and Figure S2).

Figure 2: Overexpression of *mNanog* but not *OlNanog* phenocopies *xventx1/2* gain-of-function. (A) Embryos were injected dorsally at 2-cell stage (NF2) with water as a control (Ctrl), or with mRNAs coding for *xventx1/2* (1:3 mix of *xventx1* and *xventx2*; 1.0 ng final), *mNanog* (0.3 ng final) and *OlNanog* (0.3 ng final). At tailbud stage (NF30), *mNanog* and *xventx1/2* overexpression led to similar ventralised phenotypes, while *OlNanog*-injected embryos were considerably less affected, notably in their anterior structures (lateral views, anterior to the left, dorsal to the top). (B) Percentages of observed phenotypes in three independent experiments for mock (n=30), *mNanog* (n=38), *xventx1/2* (n=31) or *OlNanog* (n=68) mRNAs injections. (C) RT-QPCR experiments and (D) whole-mount *in situ* hybridizations (WISH) in early gastrulae (NF10.5) showed that overexpression of *mNanog* and *xventx1/2* but not *OlNanog* affects *gsc* expression (left panels: whole embryos, ventral view, dorsal to the top; right panels: hemisected embryos, lateral view, dorsal to the left, vegetal pole to the bottom). In panels D and F, the number of embryos showing staining similar to the one photographed over the total number of embryos assayed is indicated. (E) Embryos injected dorsally at 2-cell stage (NF2) with *mNanog* (0.3 ng) or *xventx1/2* (1.0 ng) mRNAs, and water for controls, were processed for RT-QPCR at gastrula stage (NF10.5) using markers of mesoderm (*xbra*, *eomes*, *myf5*), organizer (*gsc*, *hhex*), endoderm (*sox17a*, *gata6*) and neuroectoderm (*sox2*) tissues. All genes studied were affected in a similar fashion by *mNanog* and *xventx1/2* overexpression. (F) Animal view of NF10.5 embryos injected unilaterally at 16-cell stage (NF5) in one AB4 blastomere with 1:3 mix *xventx1/2* (0.5 ng final) or *mNanog* (0.15 ng final) mRNAs, or water for controls (white arrowheads and black dotted lines indicate the injected side). WISH for *xk81* (epidermis) and *foxi1e* (uncommitted ectoderm) indicate that *xventx1/2* and *mNanog* repress commitment during epidermal differentiation.

Figure 3: Heterologous *mNanog* expression rescues *xventx1/2* knockdown specifically.

Embryos were first injected radially at 2-cell stage (NF2) with control MO or a 1:1 mix of *xventx1* and *xventx2* MOs (total 60 ng) and subsequently at 4-cell stage (NF3) with water, *mNanog* or *msx1* capped mRNAs (600 pg). (A) WISH of tailbud-stage embryos (anterior to the left, dorsal to the top). Controls were morphologically normal and displayed wild-type expression profiles for all analysed genes (*six6*, *hba4*, *shh*, *otx2*, *myod*). *mNanog* and *msx1* overexpression in control MO-injected embryos led to abnormal phenotypes with anterior truncations, while injection of *xventx1/2* MOs+water strongly dorsalised the embryos. Coinjecting *mNanog* but not *msx1* restored normal phenotypes. (B) Percentages of observed phenotypes in three independent experiments for Ctrl MO+water (n=36), Ctrl MO+*mNanog* (n=15), Ctrl MO+*msx1* (n=16), *xventx1/2* MO+water (n=29), *xventx1/2* MO+*mNanog* (n=103), *xventx1/2* MO+*msx1* (n=60). (C) In gastrulae (NF10.5) RT-QPCR for the genes *gsc, sox17a* and *xk81* confirmed the specific rescue of *xventx1/2* knockdown by *mNanog*.

TABLES

NKL family	Paralog used	Consensus sequence (Human, Xenopus, Zebrafish and Fugu*)	% seq. id.	# processed pseudogenes
LBX	LBX1	RRKSRTAFTNHQIYELEKRFLYQKYLSPADRDQIAQQLGLTNAQVITWFQNRRAKLKRDL	100	0
NK2.1	NKX2.1	RRKRRVLFSQAQVYELERRFKQQKYLSAPEREHLASMIHLTPTQVKIWFQNHRYKMKRQA	100	0
NK3	NKX3.2	KKRSRAAFSHAQVFELERRFNHQRYLSGPERADLAASLKLTETQVKIWFQNRRYKTKRRQ	100	0
BSX	BSX	RRKARTVFSDSQLSGLEKRFE-QRYLSTPERVELATALSLSETQVKTWFQNRRMKHKKQL	98,3	0
EMX	EMX1 (£)	PKRIRTAFSPSQLLRLERAFEKNHYVVGAERKQLA-SLSLSETQVKVWFQNRRTKYKRQK	98,3	0
HLX	HLX	RSWSRAVFSNLQRKGLEKRFE-QKYVTKPDRKQLAAMLGLTDAQVKVWFQNRRMKWRHSK	98,3	0
BARX	BARX2 (€,$)	PRRSRTIFTE-QLMGLEKKFQKQKYLSTPDRLDLAQSLGLTQLQVKTWYQNRRMKWKK-V	96,7	0
MSX	MSX1	NRKPRTPFTT-QLLALERKFRQKQYLSIAERAEFSSSL-LTETQVKIWFQNRRAKAKRLQ	96,7	1
VAX	VAX2	PKRTRTSFTAEQLYRLE-EFQRCQYVVGRERTELARQLNLSETQVKVWFQNRRTKQKKD-	96,7	0
HHEX	HHEX	RKGGQVRFSNDQT-ELEK-FETQKYLSPPERKRLAK-LQLSERQVKTWFQNRRAKWRRLK	95	0
NK5	HMX1	KKKTRTVFSRSQVFQLESTFD-KRYLSS-ERAGLAA-L-LTETQVKIWFQNRRNKWKRQL	93,3	0
NK6	NKX6.3	KKHTRPTF-GHQIF-LEKTFEQTKYLAGPERARLA-SLGM-ESQVKVWFQNRRTKWRKKS	93,3	0
EN	EN2	DKRPRTAFTA-QLQRLK-EFQTNRYLTEQRRQ-LAQEL-LNESQIKIWFQNKRAKIKKA-	91,7	0
DLX	DLX4	-RKPRTIYSSLQLQ-L-QRFQ-TQYLALPERA-LAA-LGLTQTQVKIWFQNKRSKYKK--	88,3	0
NK1	NKX1.2	PRRARTAFTYEQLVALE--FR--RYLSVCERL-LAL-L-LTETQVKIWFQNRRTKWKKQ-	88,3	0
TLX	TLX2	RKKPRTSFSR-Q--ELE-RF-RQKYLASAERA-LAKAL-M-D-QVKTWFQNRRTKWRRQT	85	0
NK2.2	NKX2.8	-KKRRVLFSKAQT-ELERRFRQQRYLS-PER-QLA--L-LTPTQVKIWFQNHRYK-KR--	83,3	0
BARHL	BARHL1 (€)	-RKARTAF---QL--LERSF--QKYLSVQDRMELAASL-L-DTQVKTWYQNRRTKWKRQ-	81,7	0
NK4	NKX2.6 (£)	RR-PRVLFSQ-QV--LERRFKQQRYLSAPER--LA--L-LTS-QVKIWFQNRRYKCKRQ-	81,7	0
DBX	DBX2	-ILRRAVFSE-QR--LE--F--QKYISK--R--LA--L-LKE-QVKIWFQNRRMKWRN--	70	0
NOTO	NOTO	-KR-RT-F---QL--LEK-F--Q---VG--R--LA--L-L-E-QV-VWFQNRR-K--KQ-	53,3	0
VENTX	VENTX2	--R-RT-FT--Q---LE--F--H-YL---ER---A----L-E-Q--TWFQNRRMK-KR--	48,3	6
NANOG	NANOG (&)	----R--FS--Q---L---F--Q-Y--------L-----L-YKQVK-WFQN-RMK-----	36,7	10

* For some families a different set of species was used, see legend for details

Table 1: Nanog and Ventx homeodomains are less conserved than other NKL families.

For each NKL family present in all vertebrates (1st column) the homeodomains (HDs) of all *Homo sapiens*, *Xenopus tropicalis*, *Danio rerio* and *Takifugu rubripes* paralogs were retrieved. When a given paralog was unknown in a given species but present in a closely related one, this alternate sequence was used instead. More specifically: (€) BARX2 and

BARHL1 being unknown in *D.rerio*, the *Gasterostus aculeatus* sequences were used; (£) EMX1 and NKX2.6 being unknown in *T.rubripes*, the *Takifugu nigroviridis* sequences were used; ($) BARX2 being unknown in *G.gallus*, the *Taeniopygia guttata* sequence was used; (&) NANOG being unknown in *Xenopus* species the *Ambystoma mexicanum* sequence was used. For each group of orthologs, the percentage of identity along the HD of the four relevant sequences was computed. For families with multiple paralogs, only the least conserved are shown here (2nd column). The consensus sequence and percentage of identity thus obtained are indicated (3rd and 4th column). The VENTX and NANOG families (in bold) present the lowest sequence identity in the HD, and are the only NKL families for which numerous processed pseudogenes are found in the human genome (5thcolumn)[22]. This similarity extends to functional properties (see Supplemental Data, Table S1).

Figure 1
Click here to download high resolution image

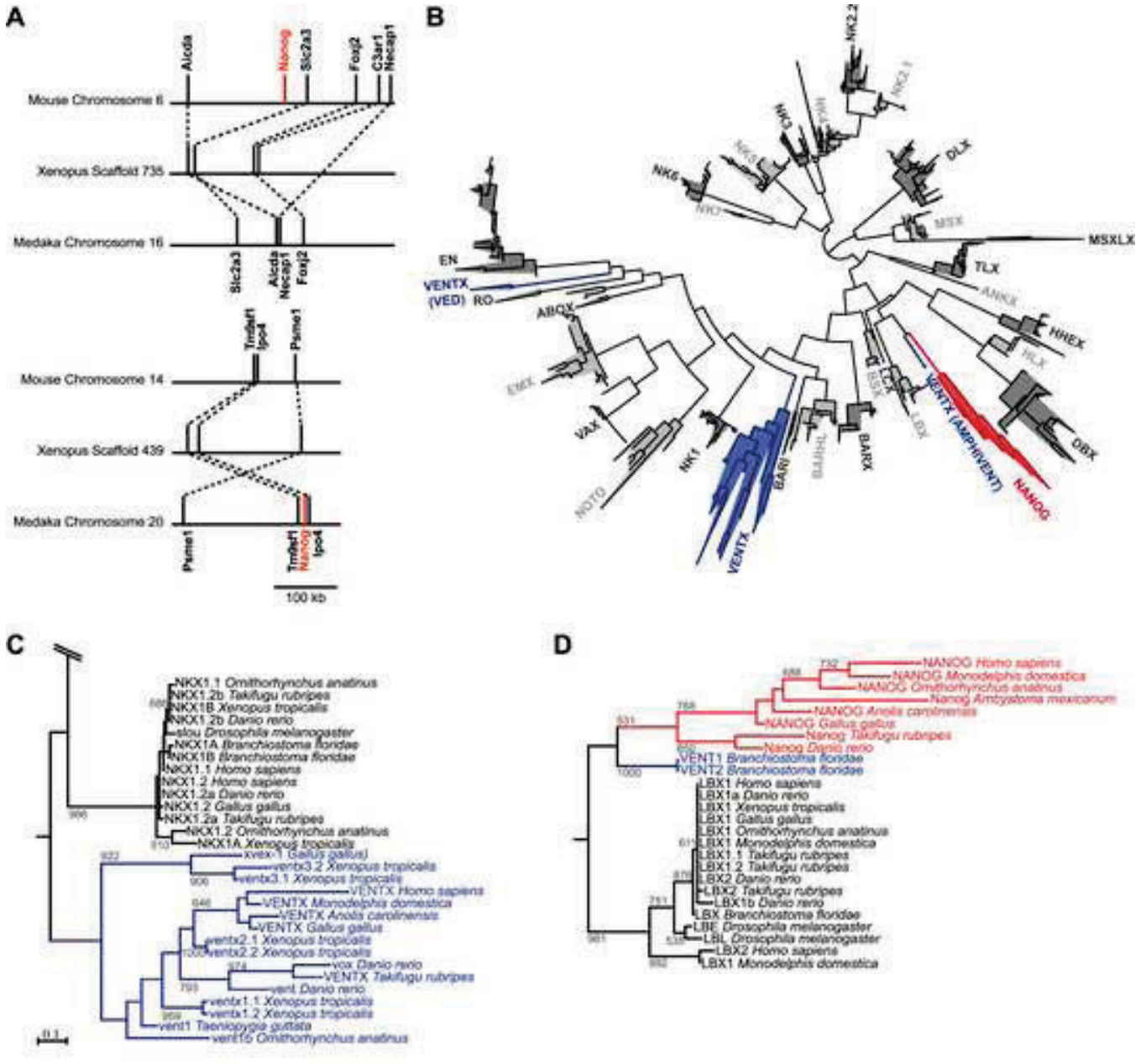


Figure 1: Syntheny of *Nanog* loci and phylogenic reconstruction of the NKL homeodomains relationships using Maximum Likelihood.

Figure 2
Click here to download high resolution image

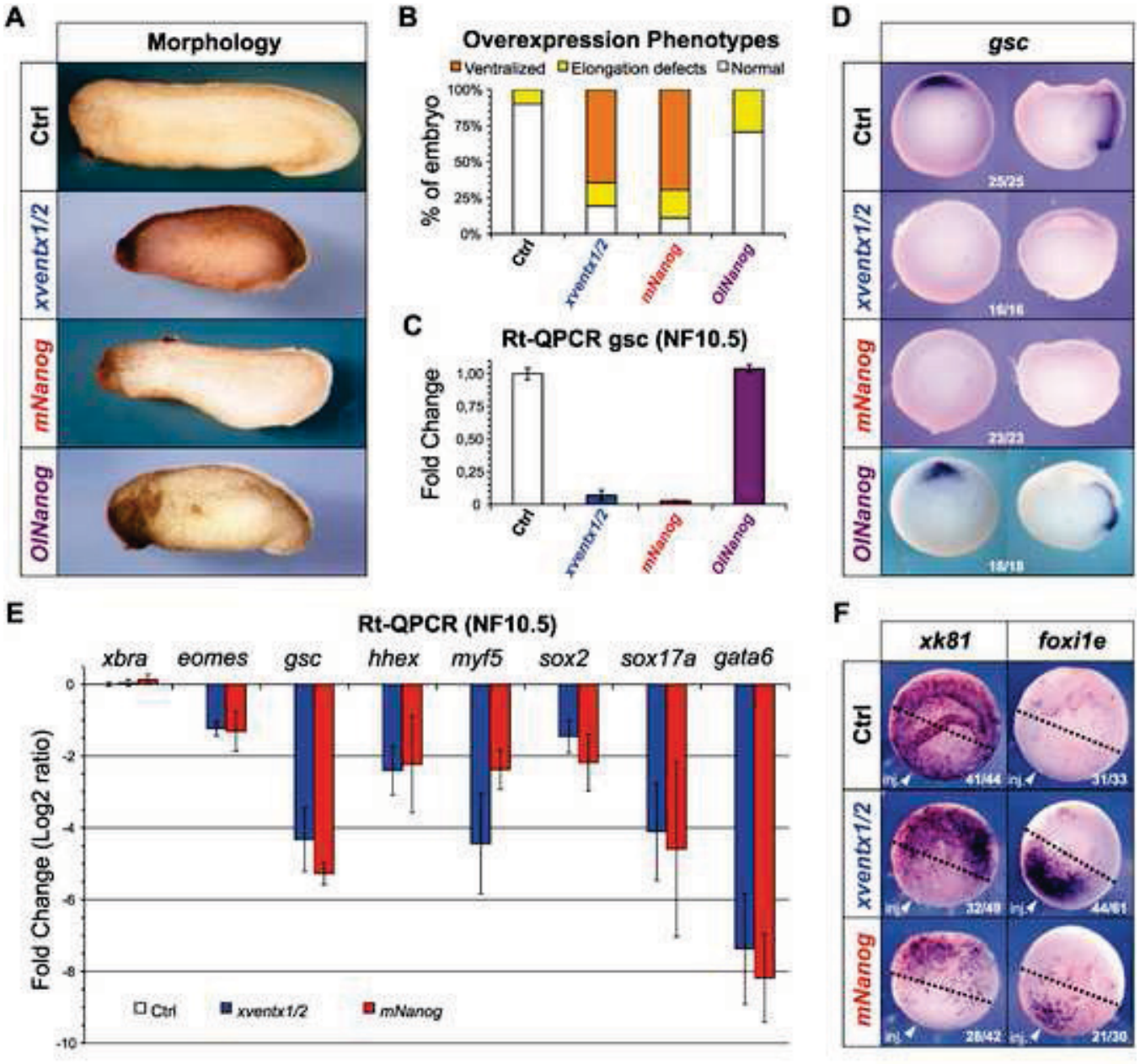


Figure 2: Overexpression of *mNanog* but not *OlNanog* phenocopies *xventx1/2* gain-of-function.

Figure 3
Click here to download high resolution image

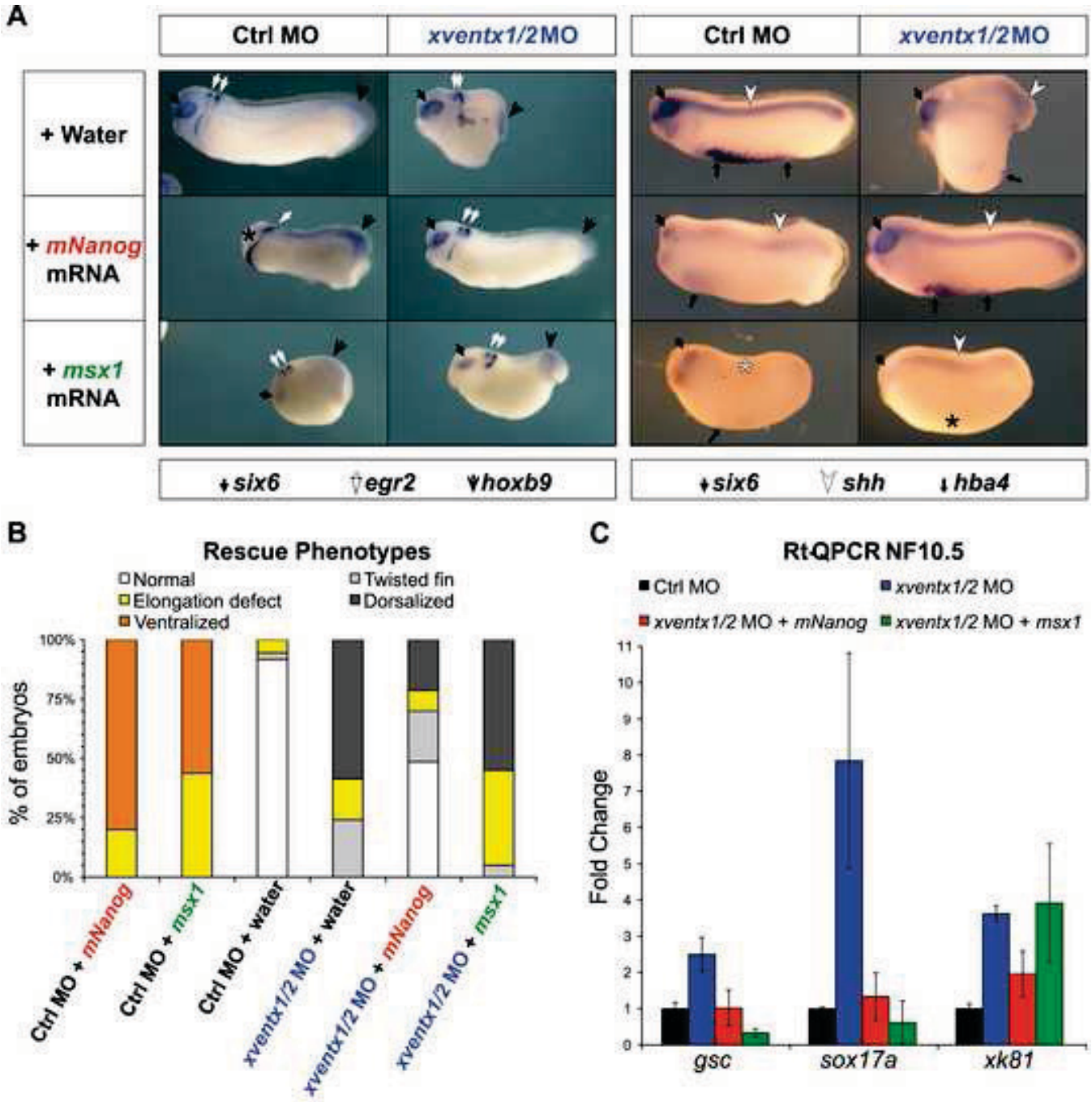


Figure 3: Heterologous *mNanog* expression rescues *xventx1/2* knockdown specifically.

INVENTORY OF SUPPLEMENTAL INFORMATION

The Supplemental Data file provided herewith contains the following items:

- Figure S1: Relaxation of purifying selection is detected during *Nanog* evolution, related to Figure 1.

The observation that NANOG and VENTX groups have long branches in the phylogeny shown in Figure 1 suggests that these genes evolved more rapidly than other NKLs. We therefore checked if any modification of selection was detected in these groups. This supplemental figure displays the results of testing for positive selection / relaxation of purifying selection in the *Nanog* and *Ventx* homeoboxes. The data are coherent with the occurrence of an event of functional shift for *Nanog* after the tetrapod/teleost separation and with the conservation of an ancestral function for *Xenopus xventxs* genes.

- Figure S2: *mNanog* and *xventx1/2* overexpression leads to comparable molecular regulation of common targets, related to Figure 2.

In Figure 2 we compare the effect of mouse *Nanog* and *xventx* overexpressions on *Xenopus laevis* development. We show that the expression of all genes assayed is affected in the same way in both conditions at gastrula stage, as is the morphology of tailbud stage embryos. In this supplementary figure, we pursue our comparison by analysing gene expression at tailbud stage: whole mount *in situ* hybridisations with *hba4* and *egr2* probes of *Xenopus* tailbud stage embryos injected with mouse *Nanog* or *xventx1/2* mRNAs are shown.

- Table S1: Mammalian *Nanog* and *Xenopus xventxs* share striking functional similarities, related to Table 1.

Table 1 shows that *ventx* and *Nanog* share structural features: they are the least conserved members of the NKL subclass of homeobox genes and both have numerous pseudogenes in the human genome. This prompted us to search for cues of functional similarities in published data. The results of this review are summarised in this supplementary table, along with the relevant references.

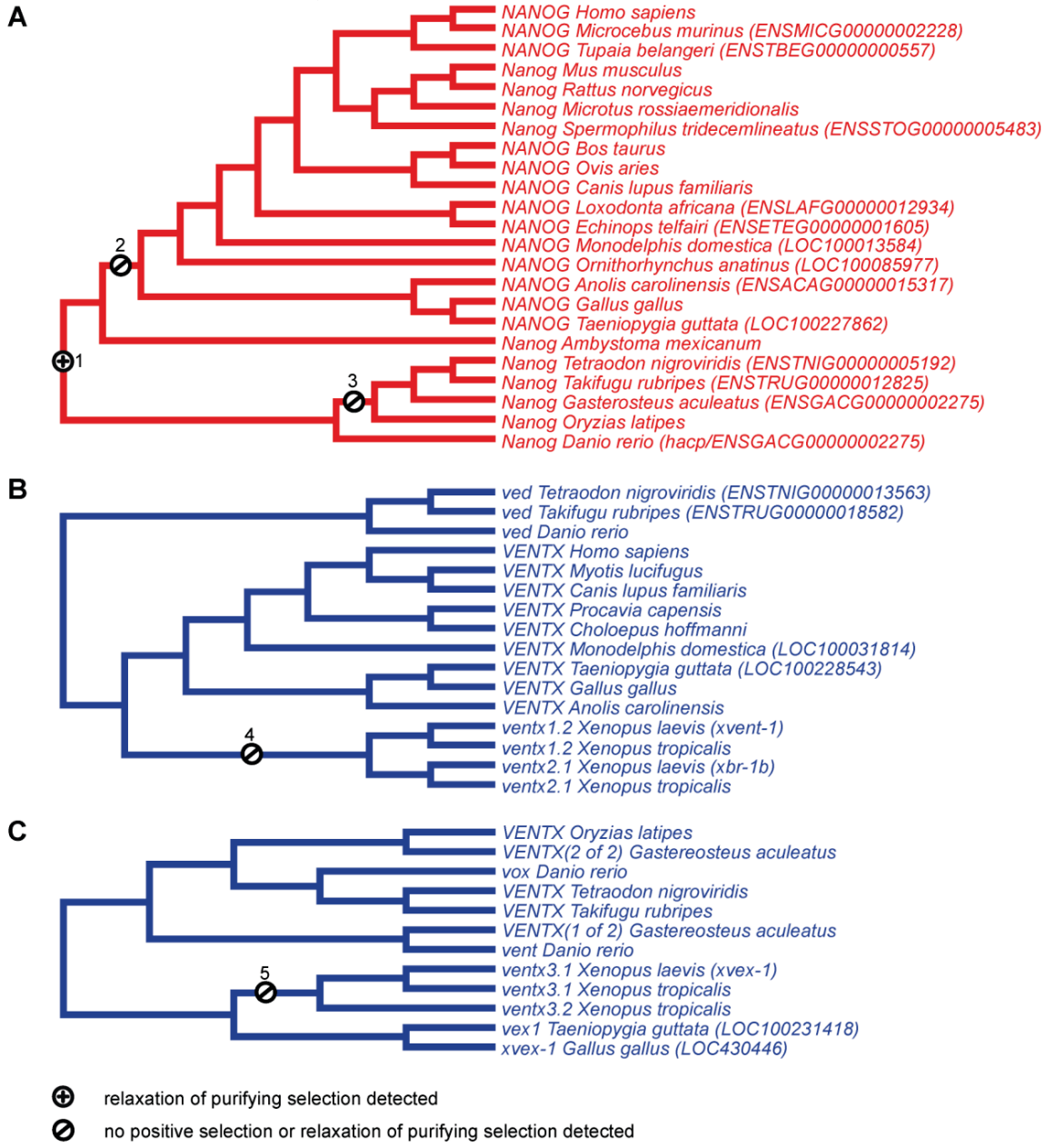
- Supplemental Experimental Procedures

The methods used in this study are described concisely in the figure legends and text. Therefore we felt that a detailed methods section was not necessary in the main body of our manuscript and that placing it in the Supplemental Information was more appropriate. The sequences of the oligonucleotides used for PCR experiments in this study are listed in a table, together with the original references when appropriate.

- **Supplemental References**

This section contains literature citations that are unique to the Supplemental Information and are cited in the Supplemental Experimental Procedures (references [42-53]) and Supplemental Table S3 (references [53-87]).

SUPPLEMENTAL DATA, FIGURES



D

				Models			Likelihood Ratio Test	
				No shift	Branch relaxation	Positive selection	Branch relaxation	Positive selection
Gene	Tested branch (#)	Species	Sites	lnL	lnL	lnL	p-value	p-value
Nanog	Tetrapods vs Teleosts (1)	23	180	-2104.21	-2095.38	-2095.14	0.000115 ***	0.49 (NS)
	Amniotes (2)	23	180	-2104.21	-2103.75	-2103.92	0.631 (NS)	0.375 (NS)
	Percomorphs (3)	23	180	-2104.21	-2104.21	-2104.21	1 (NS)	1 (NS)
xventx1/2	Xenopus (4)	16	180	-1664.53	-1664.53	-1664.53	1 (NS)	1 (NS)
xventx3	Xenopus (5)	12	180	-1375.19	-1374.83	-1374.26	0.698 (NS)	0.286 (NS)

Figure S1: Relaxation of purifying selection is detected during *Nanog* evolution, related to Figure 1

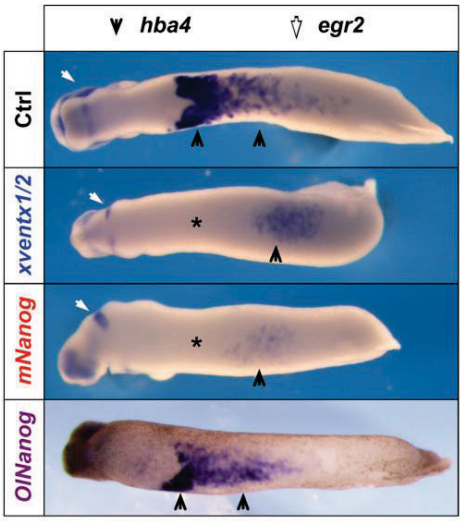


Figure S2: *mNanog* and *xventx1/2* overexpression leads to comparable molecular regulation of common targets, related to Figure 2.

SUPPLEMENTAL DATA, FIGURE LEGENDS

Figure S1: Relaxation of purifying selection is detected during *Nanog* evolution, related to Figure 1. (A) Testing for positive selection revealed that relaxation of purifying selection has occurred in the *Nanog* family prior to the teleostean and tetrapod diversifications (indicated by a circled "+" on the tree, highly significant (***): $p < 0{,}0001$). (B&C) No significant change in evolutionary rates was detected in any of the *Xenopus ventx* paralogs (indicated by circled "/" on the trees). (D) Results obtained for each tested branch. *Nanog* and *Ventx* groups are displayed in red and blue respectively.

Figure S2: *mNanog* and *xventx1/2* overexpression leads to comparable molecular regulation of common targets, related to Figure 2. WISH of tailbud-stage embryos (NF30). Overexpression of *mNanog* (0.3 ng) or *xvent1/2* (1.0 ng) resulted in reduced ventral mesoderm formation as indicated by the restricted expression domains of *hba4* (black arrows).

SUPPLEMENTAL DATA, TABLES

	mammalian Nanog			Xenopus xventx1/2		
Regulated by:						
Signalling pathways:						
	BMP4	P/C	[54]	BMP4	D/V	[14,31,65-68]
	NODAL	P/C	[55]	NODAL/ACTIVIN	D/V	[69-70]
	FGF	P/C	[56]	FGF	-	[31,71]
	WNT	P/C	[57]	WNT	-	[31,72-74]
Transcription Factors:						
	POU5F1	P/C	[58]	Pou5f1.1/Pou5f1.3	D/V	[27,75]
	SMAD1	P/C	[54,59]	Smad1	D/V	[25,68,76-77]
	SOX2	P/C	[58]	Sox2	D/V	[78]
	STAT3	P/C	[60]	Stat3	D/V	[79]
	TCF3	P/C	[61]	Tcf3	D/V	[80]
	NANOG	P/C	[58]	Unknown	n.a.	n.a.
	Unknown	n.a.	n.a.	Xventx1/2	D/V	[25,65,81-82]
Regulates and/or binds to promoter region of:						
	Bambi	-	[62]	*bambi*	D/V	[80]
	Bmp4	P/C	[59]	*bmp4*	D/V	[14,53,66-67,80-81]
	FoxA2	P/C	[62]	*foxa2/foxa4*	D/V	[26,73,83-84]
	Gsc	P/C	[55,62]	*gsc*	D/V	[14,25,29,53,67,81,84-87]
	Hesx1	P/C	[62]	*xanf1*	-	[88]
	Hhex	P/C	[62]	*hhex*	-	[73]
	Myf5	P/C	[62]	*myf5*	-	[25,89]
	Nodal	P/C	[55,59,62]	*xnr1*	D/V	[84]
	Gata2	P/C	[59]	*gata2*	D/V	[83,86-87]
	Nanog	P/C	[55,59,62]	Unknown	n.a.	n.a.
	Unknown	n.a.	n.a.	*xventx1/2*	D/V	[25,53,76-77,81-82]
	Gata6	P/C	[55,62,63]	*gata6*	-	this paper
	Eomes	P/C	[55,59,62]	*eomes*	-	this paper
	Sox2	P/C	[55,59,62]	*sox2*	-	this paper
	Sox17	P/C	[55]	*sox17a*	D/V	this paper
Interacts with:						
	POU5F1	P/C	[24]	Pou5f1.1	D/V	[27]
	SMAD1	P/C	[60]	Smad1	D/V	[76]
	NANOG	P/C	[64]	Unknown	n.a.	n.a.
	Unknown	n.a.	n.a	Xventx1/2	D/V	[25]

Table S1: Mammalian *Nanog* and *Xenopus xventxs* share striking functional similarities, related to Table 1. Mammalian *Nanog* (left) and *Xenopus xventxs* (right) are "regulated by" (top panel), "regulate" (middle panel) and "interact" (bottom panel) with orthologous pathways, transcription factors, genes and proteins, respectively. Most of these factors are known to be involved in the regulation of pluripotency and/or cell commitment and differentiation in mammals (indicated by P/C next to the gene names), while their

counterparts in frog and/or fish are known to be involved in dorsoventral patterning during embryogenesis (indicated by D/V next to the gene names). To our knowledge, *gata6*, *eomes*, *sox2* and *sox17a* were known to be regulated by *Nanog* in mammals but not by *xventxs* in *Xenopus* before this study. Relevant publications are indicated in each case (references [53-87] are listed as Supplemental References).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

In silico screening

Homeodomain sequences from all reported *Nanog* genes were retrieved from public repositories (http://www.ncbi.nlm.gov/; http://www.ensembl.org/) and used as queries to perform several rounds of TBLASTN screening on the *Xenopus tropicalis* genome assembly (http://genome.jgi-psf.org/Xentr4/Xentr4.home.html), as well as on available expressed sequence tags and cDNA sequences from *Xenopus tropicalis* and *Xenopus laevis* (http://www.ncbi.nlm.gov/).

RT-PCR screening

Total RNAs were extracted from *Xenopus laevis* ovaries, unfertilized and fertilized eggs (NF1), blastulae (NF8) and early gastrulae (NF10.5) using the RNeasy mini Kit (Qiagen), then reverse transcribed using superscript II reverse transcriptase (Invitrogen). The resulting cDNAs were screened by PCR using degenerate primers (MWG-Biotech). Primer design was based on the strict consensus sequence extracted from the *Nanog* homeoboxes found in amniotes (forward: 5'-TTYCARNNNCARAARTAYYTNWSNCC-3' or 5'-TTYGTNNNNCARAARTAYYTNWSNCC-3'), in teleosts (forward: 5'-CNGCNTTYWSNGARWSNCARATG-3' or 5'- CNGCNTTYWSNGARGARCARATG-3', reverse: 5'-ACYTGYTTRTANGTNARNCCNG-3') or in all vertebrates (reverse: 5'-TTYTGRAACCANGTYTTNAC-3'); deoxyinosine was used to reduce degeneracy. PCR reactions (94°C, 2'; 40x[94°C, 30''; 45-55°C, 1'; 72°C, 30'']; 72°C, 2', with hot-start and annealing temperature gradient) were done in the presence of ExTaq (Takara) on a MyCycler thermocycler (Biorad). No significant amplification was detected upon gel electrophoresis in presence of SybrSafe DNA Gel Stain fluorescent dye (Invitrogen). Additional PCR conditions did not yield amplification products either (primer concentrations from 0.2 to 1.0μM, Mg2+ concentrations from 1 to 4 mM, alternate cycling conditions: touch-down, bottom-up).

Specific primers for *xventx2.1* (also known as *xom*) and *xventx1* (also known as *vent-1*) were used as positive controls (see below), leading to single band amplicons of the expected size in all experiments.

Syntheny analysis

Annotated genes present in the vicinity of *Nanog* orthologs in *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Gallus gallus*, *Ornithorhynchus anatinus*, *Monodelphis domestica*, *Mus musculus* and *Homo sapiens* were retrieved from the ensembl website. This revealed that *Nanog* orthologs are found in two distinct synthenic regions: one in amniotes and another one in teleosts. These regions were compared to scaffolds of the *Xenopus tropicalis* genome containing the orthologs of the genes thus identified. A simplified representation showing the absence of *Nanog* in *Xenopus* is presented as Figure 1A, a representation including all analysed species is available upon request.

Sequences retrieval

We retrieved protein sequences of all referenced NKL factors from *Homo sapiens*, *Branchiostoma floridae* and *Drosophila melanogaster* (referenced on the homeoDB website [42]: http://homeodb.cbi.pku.edu.cn/), as well as those from *Danio rerio*, *Takifugu rubripes*, *Xenopus tropicalis*, *Anolis carolinensis*, *Gallus gallus*, *Ornithorhynchus anatinus* and *Monodelphis domestica* (referenced on the ncbi, ensembl or Joint Genome Institute websites: http://www.ncbi.nlm.gov/, http://www.ensembl.org/, http://www.jgi.doe.gov/). When a given paralog was unknown in one of these species but present in a closely related one, the relevant sequences were retained (*Gasterosteus aculeatus* for *Danio rerio*, *Tetraodon nigroviridis* for *Takifugu rubripes*, *Taeniopygia guttata* for *Gallus gallus and Xenopus laevis* or *Ambystoma mexicanum* for *Xenopus tropicalis*). All these sequences were compiled and aligned using the Seaview software [43]. This dataset as well as others used in this study are available upon

request to the corresponding author.

Phylogenetic analyses

Molecular phylogenetic analyses were performed on the 60 amino acids of the aligned NKL homeodomains using Maximum likelihood (JTT model of amino-acids substitution) and neighbour-joining methods, as implemented in the PHYML software [44]. Branch support was assessed using boostrap replication (1000 replicates).

Conservation analysis

For each NKL family conserved among vertebrates (Lbx, NK2.1, NK3, Bsx, Emx, Hlx, Barx, Msx, Vax, Hhex, NK5, NK6, En, Dlx, NK1, Tlx, Nk2.2, Dbx, Noto, Ventx and Nanog), the 60 amino acids of the homeodomains of *Homo sapiens*, *Xenopus tropicalis*, *Danio rerio* and *Takifugu rubripes* representatives were separately aligned. In order to always compare two tetrapod and two teleost sequences, when a given paralog was unknown in one of these species but present in a closely related one, the relevant sequences were retained (e.g. for Barx2, the *Gasterosteus aculeatus* sequence was used instead of *Danio rerio*; for Nanog, the *Ambystoma mexicanum* sequence was used instead of *Xenopus laevis*). For each set of orthologs, strict consensus sequences were obtained and the percentage of sequence identity computed using the Seaview software [43]. In the case of families represented by multiple paralogs, only the least conserved subfamily was retained to generate Table 1.

Test for positive selection

Positive selection was tested using the branch-site model A, as implemented in codeml from the PAML package version 4b [45]. Positive selection is detected if there is a category of sites with dN/dS ratio omega > 1 on the tested branch. Importantly, the test contrasts positive selection on the branch of interest to the possibility of relaxed purifying selection, which avoids a major source of false positive results. The test is done by comparing the difference of log-likelihood (lnl) values to a chi2 distribution of 1 degree of freedom and

corrected for multiple testing [46]. The test was carried on the whole homeobox (180 nucleotides) on a representative set of vertebrates.

Embryo manipulation and injection

Xenopus laevis were obtained from NASCO. All animal studies were conducted according to the principles and procedures described in the Guidelines for care and Use of Experimental animals. Oocytes obtained from females were fertilized *in vitro*, de-jellied before injection and the developing embryos cultured until the appropriate stages using standard procedures [14,25,27,29]. Synthetic capped mRNAs were transcribed with the mMessage mMachine SP6 kit (Ambion) using the following templates: pCS2+-Vent1 and pCS2+-Xbr1b, both linearised with NotI (gifts of N. Papalopulu, University of Manchester, UK and respectively referred to as *xventx1* and *xventx2* in this work); pSP64T-xMsx1, linearised with EcoRI [13,14]; pCS2+MT-OlNanog, linearised with Sac II (a gift of J.L. Mullor, Hospital La Fe, Universidad de Valencia, Spain) [9]. To express *mNanog*, the ORF of a commercial clone (Geneservice) was PCR amplified and cloned into pCS2+, linearised with NotI and transcribed with SP6. Previously described morpholino oligonucleotides (MOs) directed against *xventx1* and *xventx2* pseudoalleles [29] were obtained from GeneTools. In order to rule out possible interference of MOs mixed with mRNAs before injection, we performed rescue assays through injections of MOs at the 2-cell stage, followed by mRNAs injections at the 4-cell stage. All injections were performed at least in three times to assess reproducibility.

In situ hybridization and Real-Time Quantitative RT-PCR

Injected embryos were processed for whole-mount *in situ* hybridization (WISH) with digoxygenin-labelled probes (Roche) using standard procedures [14] and staining was done with BM purple (Roche). Embryos were bleached with hydrogen peroxide 4% (Carlo Erba Reagenti) and photographed with a SMZ800 binocular (Nikon) coupled to a DS-Fi1/DS-L2

acquisition system (Nikon).

For Real-Time Quantitative RT-PCR (RT-QPCR) total RNAs were extracted from 10 embryos, as described above. Three independent biological replicates were collected and RT-QPCR reactions were performed in duplicate for each sample using Power SYBR® master mix (Applied Biosystems) on a 7300 Real-Time PCR System (Applied Biosystems) following manufacturer recommendations. Primers (MWG Biotech) were described in previous publications or designed using Primer Express Software (Applied Biosystems); the relevant sequences and references are listed below. Ct data were collected using 7300 system software (Applied Biosystem) and analysed using Microsoft Excel. First, the Ct for each technical duplicate were averaged and normalised against the *DNA elongation factor type 1 α* (*eef1a1*) as a loading control. Variations of expression were quantified using the ΔΔCts method, using the mean of the control condition as reference for each gene tested. Finally, data from independent experiments were averaged and presented in histograms as fold change (in log scale on Figure 2E), with SEM as error bars.

Gene	Original reference	Forward primer sequences	Reverse primer sequences
eef1a1	Fini J.B. et al., unpublished	5'-TGG ATA GCC CCT GTG TTG GAT T-3'	5'-TCC ACG CAC ATT GGC TTT CCT-3'
eomes	[47]	5'-TGG TCC TCA AGG TCA AGT CC-3'	5'-GGG GAG TTT TCA TTG CTT GA-3'
gata6	[48]	5'-CCA ACC GGG AGC CCC GAT A-3'	5'-GCT GCT GTA GCC TGT ATC C-3'
gsc	[49]	5'-TTC ACC GAT GAA CAA CTG GA-3'	5'-TTC CAC TTT TGG GCA TTT TC-3'
hhex	[50]	5'-AAC AGC GCA TCT AAT GGG AC-3'	5'-CCT TTC CGC TTG TGC AGA GG-3'
xk81	XMMR	5'-CAC CAG AAC ACA GAG TAC-3'	5'-CAA CCT TCC CAT CAA CCA-3'
myf5	This paper	5'-TAG CTG TTC AGA TGG CAT GTC T-3'	5'-CGG AAG GGA GTC AGT GCT AC-3'
sox17a	[48]	5'-GCA AGA TGC TTG GCA AGT CG-3'	5'-GCT GAA GTT CTC TAG ACA CA-3'
sox2	[51]	5'-CCA GTC CAC CTG TAG TCA CCT CT-3'	5'-CAC TTC TGC CCC AGG TAG GTA C-3'
xbra	[52]	5'-TTC TGA AGG TGA GCA TGT CG-3'	5'-GTT TGA CTT TGC TAA AAG AGA CAG G-3'
xom	[53]	5'-TTT CAG ATG CTC TAC CTG C-3'	5'-CAA ATG GCC TTT CTT CCT G-3'
vent-1	[29]	5'-TTC CCT TCA GCA TGG TTC AAC-3'	5'-GCA TCT CCT TGG CAT ATT TGG-3'

XMMR: *Xenopus* Molecular Marker Ressource (http://www.xenbase.org/xmmr/Marker_pages/primers.html)

Table: Primer pairs used for RT-PCR experiments in this study. For each primer pair, the forward and reverse sequences are listed, as well as the original publications (references [47-53] are listed in Supplemental References).

SUPPLEMENTAL REFERENCES

42. Zhong, Y.F., Butts, T., and Holland, P.W. (2008). HomeoDB: a database of homeobox gene diversity. Evol Dev *10*, 516-518.

43. Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol *27*, 221-224.

44. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52*, 696-704.

45. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586-1591.

46. Anisimova, M., and Yang, Z. (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol *24*, 1219-1228.

47. Kofron, M., Wylie, C., and Heasman, J. (2004). The role of Mixer in patterning the early Xenopus embryo. Development *131*, 2431-2441.

48. Xanthos, J.B., Kofron, M., Wylie, C., and Heasman, J. (2001). Maternal VegT is the initiator of a molecular network specifying endoderm in Xenopus laevis. Development *128*, 167-180.

49. Xanthos, J.B., Kofron, M., Tao, Q., Schaible, K., Wylie, C., and Heasman, J. (2002). The roles of three signaling pathways in the formation and function of the Spemann Organizer. Development *129*, 4027-4043.

50. Chang, C., and Hemmati-Brivanlou, A. (2000). A post-mid-blastula transition requirement for TGFbeta signaling in early endodermal specification. Mech Dev *90*, 227-235.

51. Cao, Y., Siegel, D., and Knochel, W. (2006). Xenopus POU factors of subclass V inhibit activin/nodal signaling during gastrulation. Mech Dev *123*, 614-625.

52. Sun, B.I., Bush, S.M., Collins-Racie, L.A., LaVallie, E.R., DiBlasio-Smith, E.A.,

Wolfman, N.M., McCoy, J.M., and Sive, H.L. (1999). derriere: a TGF-beta family member required for posterior development in Xenopus. Development *126*, 1467-1482.

53. Gao, H., Wu, B., Giese, R., and Zhu, Z. (2007). Xom interacts with and stimulates transcriptional activity of LEF1/TCFs: implications for ventral cell fate determination during vertebrate embryogenesis. Cell Res *17*, 345-356.

54. Xu, R.H., Sampsell-Barron, T.L., Gu, F., Root, S., Peck, R.M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., Stewart, R., and Thomson, J.A. (2008). NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell *3*, 196-206.

55. Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L.E., Trotter, M.W., Cho, C.H., Martinez, A., Rugg-Gunn, P., Brons, G., and Pedersen, R.A. (2009). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. Development *136*, 1339-1349.

56. Nichols, J., Silva, J., Roode, M., and Smith, A. (2009). Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. Development *136*, 3215-3222.

57. Sumi, T., Tsuneyoshi, N., Nakatsuji, N., and Suemori, H. (2008). Defining early lineage specification of human embryonic stem cells by the orchestrated balance of canonical Wnt/beta-catenin, Activin/Nodal and BMP signaling. Development *135*, 2969-2979.

58. Boer, B., Kopp, J., Mallanna, S., Desler, M., Chakravarthy, H., Wilder, P.J., Bernadt, C., and Rizzino, A. (2007). Elevating the levels of Sox2 in embryonal carcinoma cells and embryonic stem cells inhibits the expression of Sox2:Oct-3/4 target genes. Nucleic Acids Res *35*, 1773-1786.

59. Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.Y., Sung, K.W., Lee, C.W., Zhao, X.D., Chiu, K.P., Lipovich, L., Kuznetsov, V.A., Robson, P., Stanton, L.W., Wei, C.L., Ruan, Y., Lim, B., and Ng, H.H. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet *38*, 431-440.

60. Suzuki, A., Raya, A., Kawakami, Y., Morita, M., Matsui, T., Nakashima, K., Gage, F.H., Rodriguez-Esteban, C., and Izpisua Belmonte, J.C. (2006). Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. Proc Natl Acad Sci U S A *103*, 10294-10299.

61. Pereira, L., Yi, F., and Merrill, B.J. (2006). Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal. Mol Cell Biol *26*, 7479-7491.

62. Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford, D.K., Melton, D.A., Jaenisch, R., and Young, R.A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

63. Fujikura, J., Yamato, E., Yonemura, S., Hosoda, K., Masui, S., Nakao, K., Miyazaki Ji, J., and Niwa, H. (2002). Differentiation of embryonic stem cells is induced by GATA factors. Genes Dev *16*, 784-789.

64. Wang, J., Levasseur, D.N., and Orkin, S.H. (2008). Requirement of Nanog dimerization for stem cell self-renewal and pluripotency. Proc Natl Acad Sci U S A *105*, 6326-6331.

65. Rastegar, S., Friedle, H., Frommer, G., and Knochel, W. (1999). Transcriptional regulation of Xvent homeobox genes. Mech Dev *81*, 139-149.

66. Schuler-Metz, A., Knochel, S., Kaufmann, E., and Knochel, W. (2000). The homeodomain transcription factor Xvent-2 mediates autocatalytic regulation of BMP-

4 expression in Xenopus embryos. J Biol Chem *275*, 34365-34374.

67. Schmidt, J.E., von Dassow, G., and Kimelman, D. (1996). Regulation of dorsal-ventral patterning: the ventralizing effects of the novel Xenopus homeobox gene Vox. Development *122*, 1711-1721.

68. Lee, H.S., Park, M.J., Lee, S.Y., Hwang, Y.S., Lee, H., Roh, D.H., Kim, J.I., Park, J.B., Lee, J.Y., Kung, H.F., and Kim, J. (2002). Transcriptional regulation of Xbr-1a/Xvent-2 homeobox gene: analysis of its promoter region. Biochem Biophys Res Commun *298*, 815-823.

69. Piepenburg, O., Grimmer, D., Williams, P.H., and Smith, J.C. (2004). Activin redux: specification of mesodermal pattern in Xenopus by graded concentrations of endogenous activin B. Development *131*, 4977-4986.

70. Wessely, O., Kim, J.I., Geissert, D., Tran, U., and De Robertis, E.M. (2004). Analysis of Spemann organizer formation in Xenopus embryos by cDNA macroarrays. Dev Biol *269*, 552-566.

71. Keren, A., Keren-Politansky, A., and Bengal, E. (2008). A p38 MAPK-CREB pathway functions to pattern mesoderm in Xenopus. Dev Biol *322*, 86-94.

72. Marom, K., Fainsod, A., and Steinbeisser, H. (1999). Patterning of the mesoderm involves several threshold responses to BMP-4 and Xwnt-8. Mech Dev *87*, 33-44.

73. McLin, V.A., Rankin, S.A., and Zorn, A.M. (2007). Repression of Wnt/beta-catenin signaling in the anterior endoderm is essential for liver and pancreas development. Development *134*, 2207-2217.

74. Hikasa, H., Ezan, J., Itoh, K., Li, X., Klymkowsky, M.W., and Sokol, S.Y. (2010). Regulation of TCF3 by Wnt-dependent phosphorylation during vertebrate axis specification. Dev Cell *19*, 521-532.

75. Snir, M., Ofir, R., Elias, S., and Frank, D. (2006). Xenopus laevis POU91 protein, an

Oct3/4 homologue, regulates competence transitions from mesoderm to neural cell fates. Embo J *25*, 3664-3674.

76. Henningfeld, K.A., Friedle, H., Rastegar, S., and Knochel, W. (2002). Autoregulation of Xvent-2B; direct interaction and functional cooperation of Xvent-2 and Smad1. J Biol Chem *277*, 2097-2103.

77. Henningfeld, K.A., Rastegar, S., Adler, G., and Knochel, W. (2000). Smad1 and Smad4 are components of the bone morphogenetic protein-4 (BMP-4)-induced transcription complex of the Xvent-2B promoter. J Biol Chem *275*, 21827-21835.

78. Rogers, C.D., Harafuji, N., Archer, T., Cunningham, D.D., and Casey, E.S. (2009). Xenopus Sox3 activates sox2 and geminin and indirectly represses Xvent2 expression to induce neural progenitor formation at the expense of non-neural ectodermal derivatives. Mech Dev *126*, 42-55.

79. Nishinakamura, R., Matsumoto, Y., Matsuda, T., Ariizumi, T., Heike, T., Asashima, M., and Yokota, T. (1999). Activation of Stat3 by cytokine receptor gp130 ventralizes Xenopus embryos independent of BMP-4. Dev Biol *216*, 481-490.

80. Karaulanov, E., Knochel, W., and Niehrs, C. (2004). Transcriptional regulation of BMP4 synexpression in transgenic Xenopus. Embo J *23*, 844-856.

81. Onichtchouk, D., Gawantka, V., Dosch, R., Delius, H., Hirschfeld, K., Blumenstock, C., and Niehrs, C. (1996). The Xvent-2 homeobox gene is part of the BMP-4 signalling pathway controlling [correction of controling] dorsoventral patterning of Xenopus mesoderm. Development *122*, 3045-3053.

82. Friedle, H., and Knochel, W. (2002). Cooperative interaction of Xvent-2 and GATA-2 in the activation of the ventral homeobox gene Xvent-1B. J Biol Chem *277*, 23872-23881.

83. Xu, R.H., Ault, K.T., Kim, J., Park, M.J., Hwang, Y.S., Peng, Y., Sredni, D., and

Kung, H. (1999). Opposite effects of FGF and BMP-4 on embryonic blood formation: roles of PV.1 and GATA-2. Dev Biol *208*, 352-361.29.

84. Melby, A.E., Clements, W.K., and Kimelman, D. (1999). Regulation of dorsal gene expression in Xenopus by the ventralizing homeodomain gene Vox. Dev Biol *211*, 293-305.

85. Trindade, M., Tada, M., and Smith, J.C. (1999). DNA-binding specificity and embryological function of Xom (Xvent-2). Dev Biol *216*, 442-456.

86. Hwang, Y.S., Seo, J.J., Cha, S.W., Lee, H.S., Lee, S.Y., Roh, D.H., Kung Hf, H.F., Kim, J., and Ja Park, M. (2002). Antimorphic PV.1 causes secondary axis by inducing ectopic organizer. Biochem Biophys Res Commun *292*, 1081-1086.

87. Hwang, Y.S., Lee, H.S., Roh, D.H., Cha, S., Lee, S.Y., Seo, J.J., Kim, J., and Park, M.J. (2003). Active repression of organizer genes by C-terminal domain of PV.1. Biochem Biophys Res Commun *308*, 79-86.

86. Martynova, N., Eroshkin, F., Ermakova, G., Bayramov, A., Gray, J., Grainger, R., and Zaraisky, A. (2004). Patterning the forebrain: FoxA4a/Pintallavis and Xvent2 determine the posterior limit of Xanf1 expression in the neural plate. Development *131*, 2329-2338.

87. Polli, M., and Amaya, E. (2002). A study of mesoderm patterning through the analysis of the regulation of Xmyf-5 expression. Development *129*, 2917-2927.