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:

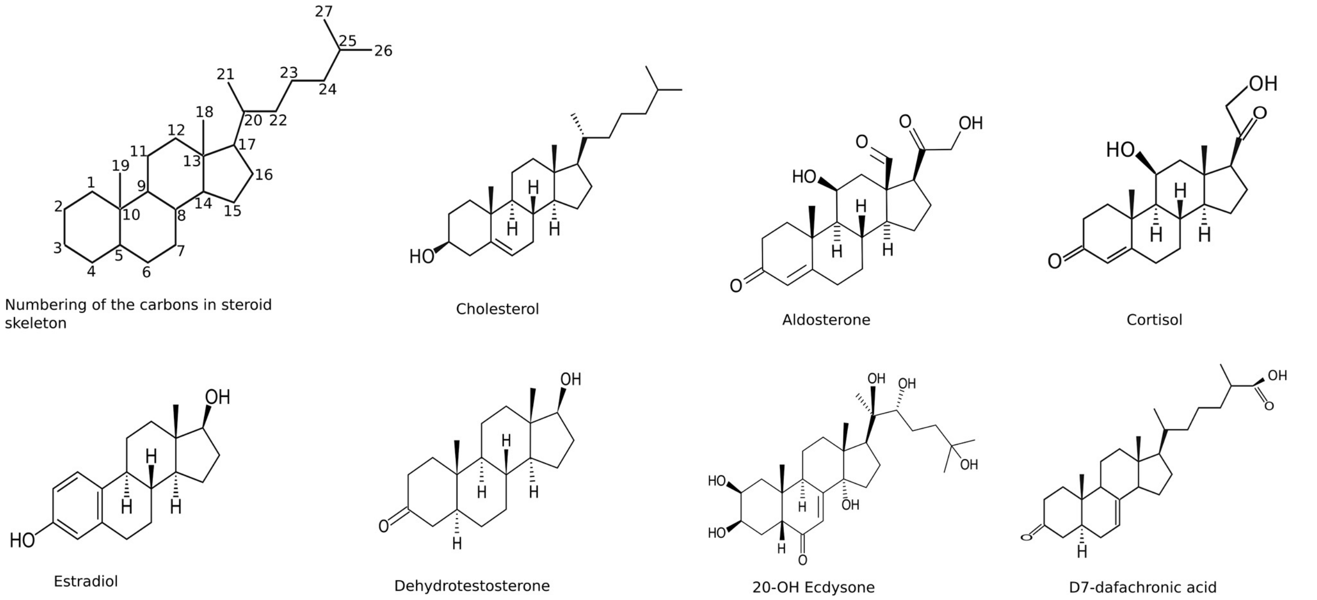
http://drnelson.utmem.edu/biblioC.html.

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.



Numbering of the carbons in steroid skeleton

Cholesterol

Aldosterone

Cortisol

Estradiol

Dehydrotestosterone

20-OH Ecdysone

D7-dafachronic acid

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*.

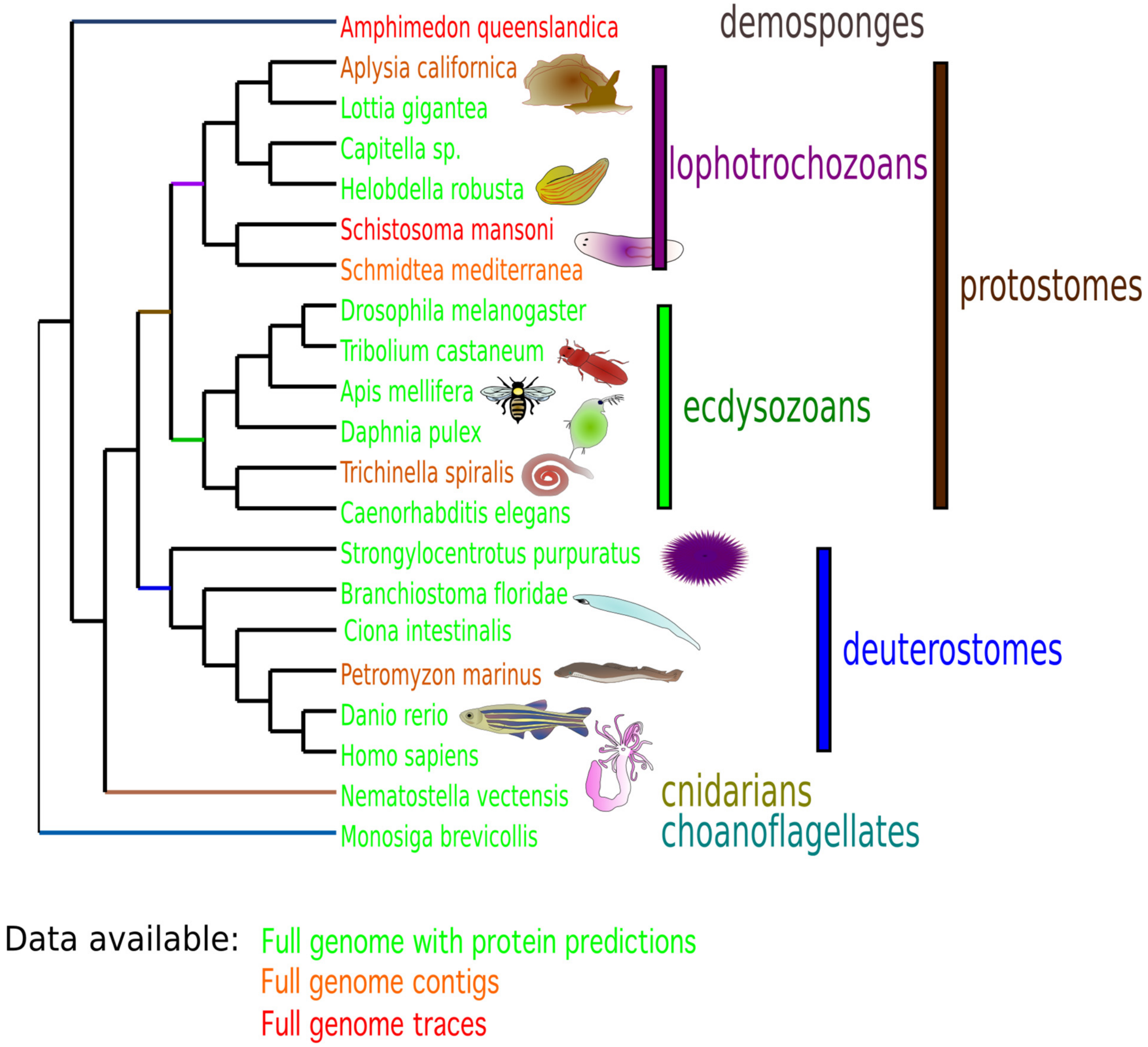


Fig. S2. Genomic data used in this study. The genomic model species that were screened in this study are indicated, with complementary information about their phylogenetic relationships and about the quality of their genome data. Species in green are those for which EST-based gene predictions are available.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

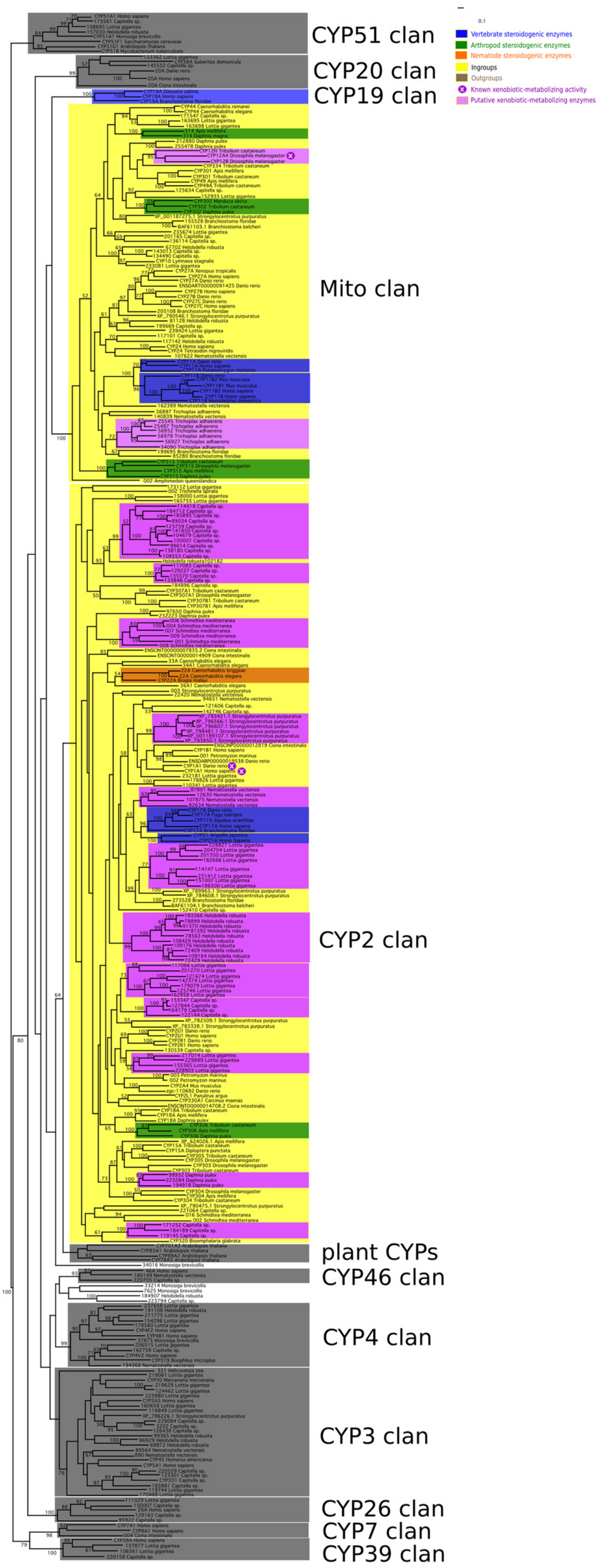

CYP51 clan
CYP20 clan
CYP19 clan
Mito clan
CYP2 clan
plant CYPs
CYP46 clan
CYP4 clan
CYP3 clan
CYP26 clan
CYP7 clan
CYP39 clan
0.1
Vertebrate steroidogenic enzymes
Arthropod steroidogenic enzymes
Nematode steroidogenic enzymes
Ingroups
Outgroups
Known xenobiotic-metabolizing activity
Putative xenobiotic-metabolizing enzymes

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.

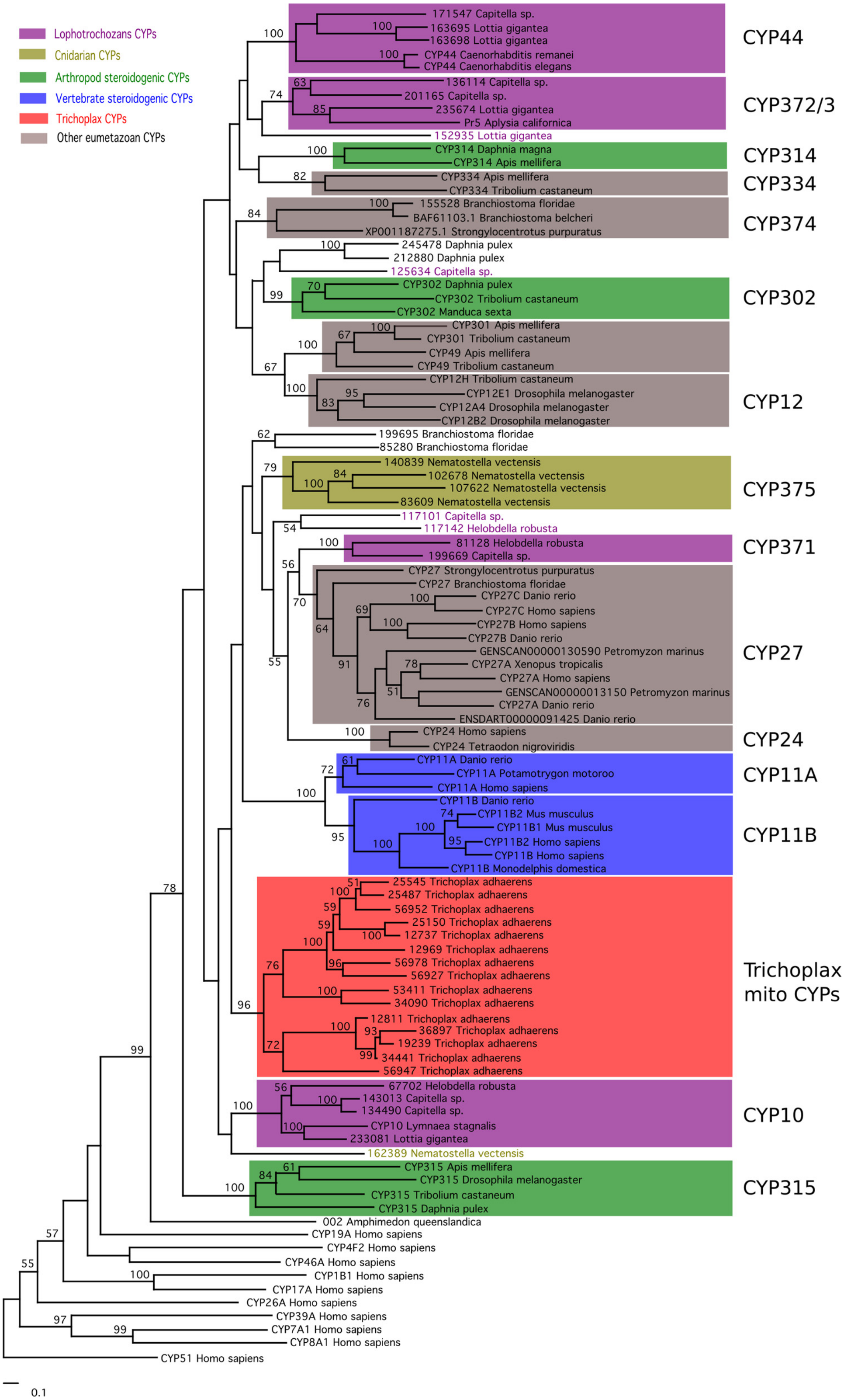

Lophotrochozans CYPs
Cnidarian CYPs
Arthropod steroidogenic CYPs
Vertebrate steroidogenic CYPs
Trichoplax CYPs
Other eumetazoan CYPs
171547 Capitella sp.
163695 Lottia gigantea
163698 Lottia gigantea
CYP44 Caenorhabditis remanei
CYP44 Caenorhabditis elegans
CYP44
136114 Capitella sp.
201165 Capitella sp.
235674 Lottia gigantea
Pr5 Aplysia californica
CYP372/3
152935 Lottia gigantea
CYP314 Daphnia magna
CYP314 Apis mellifera
CYP314
CYP334 Apis mellifera
CYP334 Tribolium castaneum
CYP334
155528 Branchiostoma floridae
BAF61103.1 Branchiostoma belcheri
XP001187275.1 Strongylocentrotus purpuratus
CYP374
245478 Daphnia pulex
212880 Daphnia pulex
125634 Capitella sp.
CYP302 Daphnia pulex
CYP302 Tribolium castaneum
CYP302 Manduca sexta
CYP302
CYP301 Apis mellifera
CYP301 Tribolium castaneum
CYP49 Apis mellifera
CYP49 Tribolium castaneum
CYP12H Tribolium castaneum
CYP12E1 Drosophila melanogaster
CYP12A4 Drosophila melanogaster
CYP12B2 Drosophila melanogaster
CYP12
199695 Branchiostoma floridae
85280 Branchiostoma floridae
140839 Nematostella vectensis
102678 Nematostella vectensis
107622 Nematostella vectensis
83609 Nematostella vectensis
CYP375
117101 Capitella sp.
117142 Helobdella robusta
81128 Helobdella robusta
199669 Capitella sp.
CYP371
CYP27 Strongylocentrotus purpuratus
CYP27 Branchiostoma floridae
CYP27C Danio rerio
CYP27C Homo sapiens
CYP27B Homo sapiens
CYP27B Danio rerio
GENSCAN00000130590 Petromyzon marinus
CYP27A Xenopus tropicalis
CYP27A Homo sapiens
GENSCAN00000013150 Petromyzon marinus
CYP27A Danio rerio
ENSDART00000091425 Danio rerio
CYP27
CYP24 Homo sapiens
CYP24 Tetraodon nigroviridis
CYP24
CYP11A Danio rerio
CYP11A Potamotrygon motoroo
CYP11A Homo sapiens
CYP11A
CYP11B Danio rerio
CYP11B2 Mus musculus
CYP11B1 Mus musculus
CYP11B2 Homo sapiens
CYP11B Homo sapiens
CYP11B Monodelphis domestica
CYP11B
25545 Trichoplax adhaerens
25487 Trichoplax adhaerens
56952 Trichoplax adhaerens
25150 Trichoplax adhaerens
12737 Trichoplax adhaerens
12969 Trichoplax adhaerens
56978 Trichoplax adhaerens
56927 Trichoplax adhaerens
53411 Trichoplax adhaerens
34090 Trichoplax adhaerens
12811 Trichoplax adhaerens
36897 Trichoplax adhaerens
19239 Trichoplax adhaerens
34441 Trichoplax adhaerens
56947 Trichoplax adhaerens
Trichoplax mito CYPs
67702 Helobdella robusta
143013 Capitella sp.
134490 Capitella sp.
CYP10 Lymnaea stagnalis
233081 Lottia gigantea
CYP10
162389 Nematostella vectensis
CYP315 Apis mellifera
CYP315 Drosophila melanogaster
CYP315 Tribolium castaneum
CYP315 Daphnia pulex
CYP315
002 Amphimedon queenslandica
CYP19A Homo sapiens
CYP4F2 Homo sapiens
CYP46A Homo sapiens
CYP1B1 Homo sapiens
CYP17A Homo sapiens
CYP26A Homo sapiens
CYP39A Homo sapiens
CYP7A1 Homo sapiens
CYP8A1 Homo sapiens
CYP51 Homo sapiens
0.1

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.

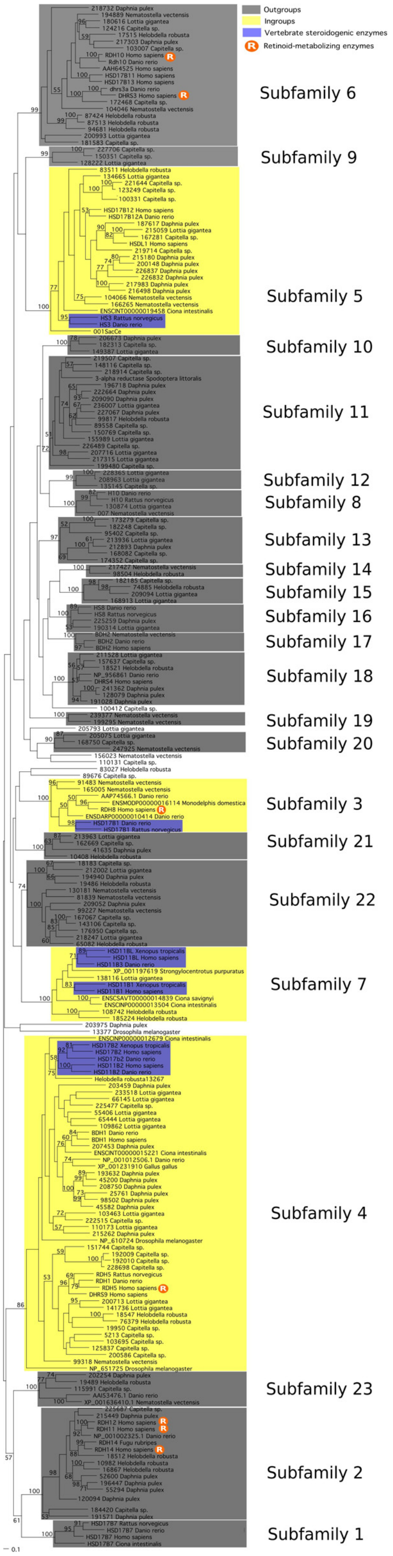


Fig. S5. Phylogeny of the SDR family. Our analysis reveals the existence of 23 strongly supported families that were arbitrarily named from 1 to 23, and are highlighted in gray. Groups of vertebrate steroidogenic enzymes are in blue, and other member of the same subfamily are in yellow. Enzymes with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

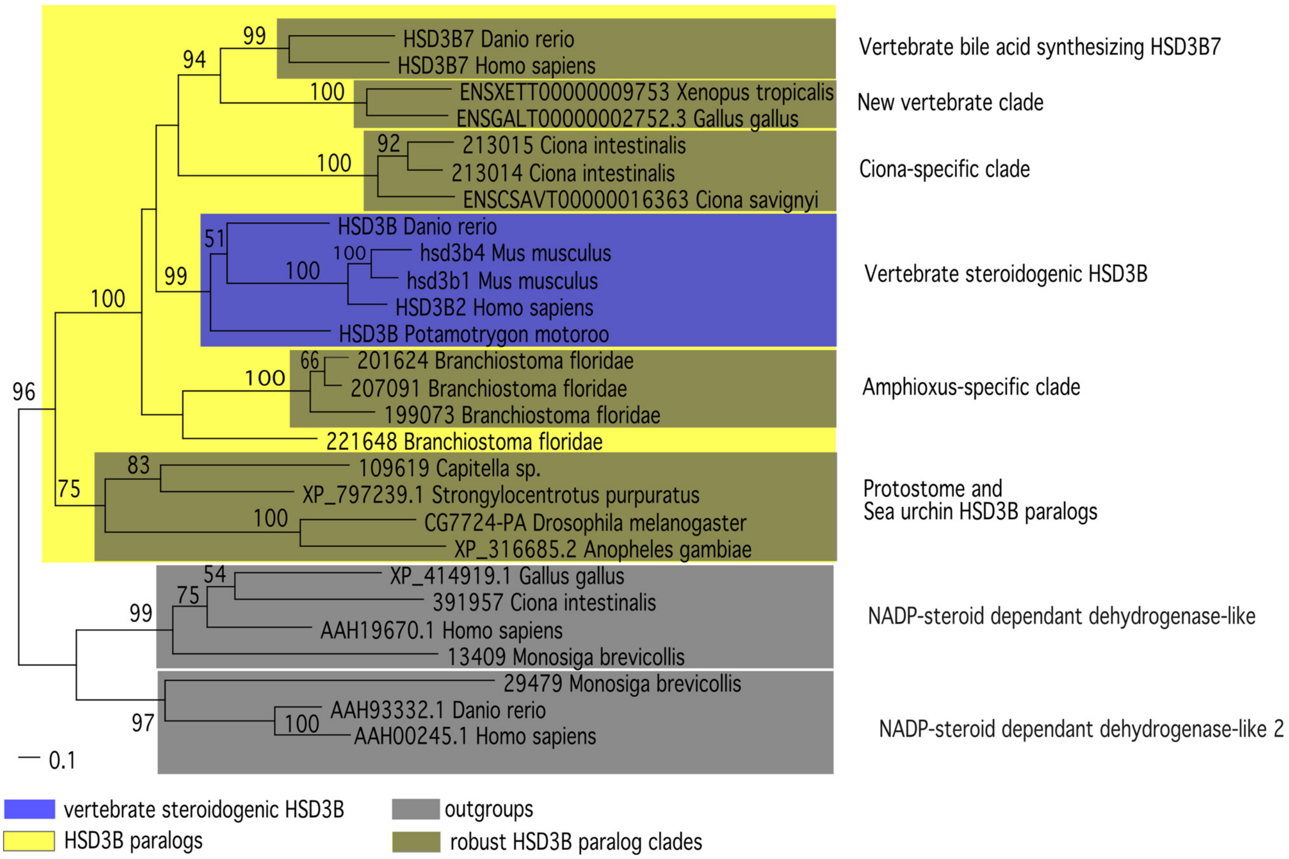

99
HSD3B7 Danio rerio
94
HSD3B7 Homo sapiens
100
ENSXETT00000009753 Xenopus tropicalis
ENSGALT00000002752.3 Gallus gallus
92
213015 Ciona intestinalis
100
213014 Ciona intestinalis
ENSCSAVT00000016363 Ciona savignyi
HSD3B Danio rerio
51
100
hsd3b4 Mus musculus
99
100
hsd3b1 Mus musculus
100
HSD3B2 Homo sapiens
HSD3B Potamotrygon motoroo
66
201624 Branchiostoma floridae
100
207091 Branchiostoma floridae
96
199073 Branchiostoma floridae
221648 Branchiostoma floridae
83
109619 Capitella sp.
75
XP_797239.1 Strongylocentrotus purpuratus
100
CG7724-PA Drosophila melanogaster
XP_316685.2 Anopheles gambiae
54
XP_414919.1 Gallus gallus
75
391957 Ciona intestinalis
99
AAH19670.1 Homo sapiens
13409 Monosiga brevicollis
29479 Monosiga brevicollis
AAH93332.1 Danio rerio
97
100
AAH00245.1 Homo sapiens
0.1
Vertebrate bile acid synthesizing HSD3B7
New vertebrate clade
Ciona-specific clade
Vertebrate steroidogenic HSD3B
Amphioxus-specific clade
Protostome and
Sea urchin HSD3B paralogs
NADP-steroid dependant dehydrogenase-like
NADP-steroid dependant dehydrogenase-like 2
vertebrate steroidogenic HSD3B
HSD3B paralogs
outgroups
robust HSD3B paralog clades

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.

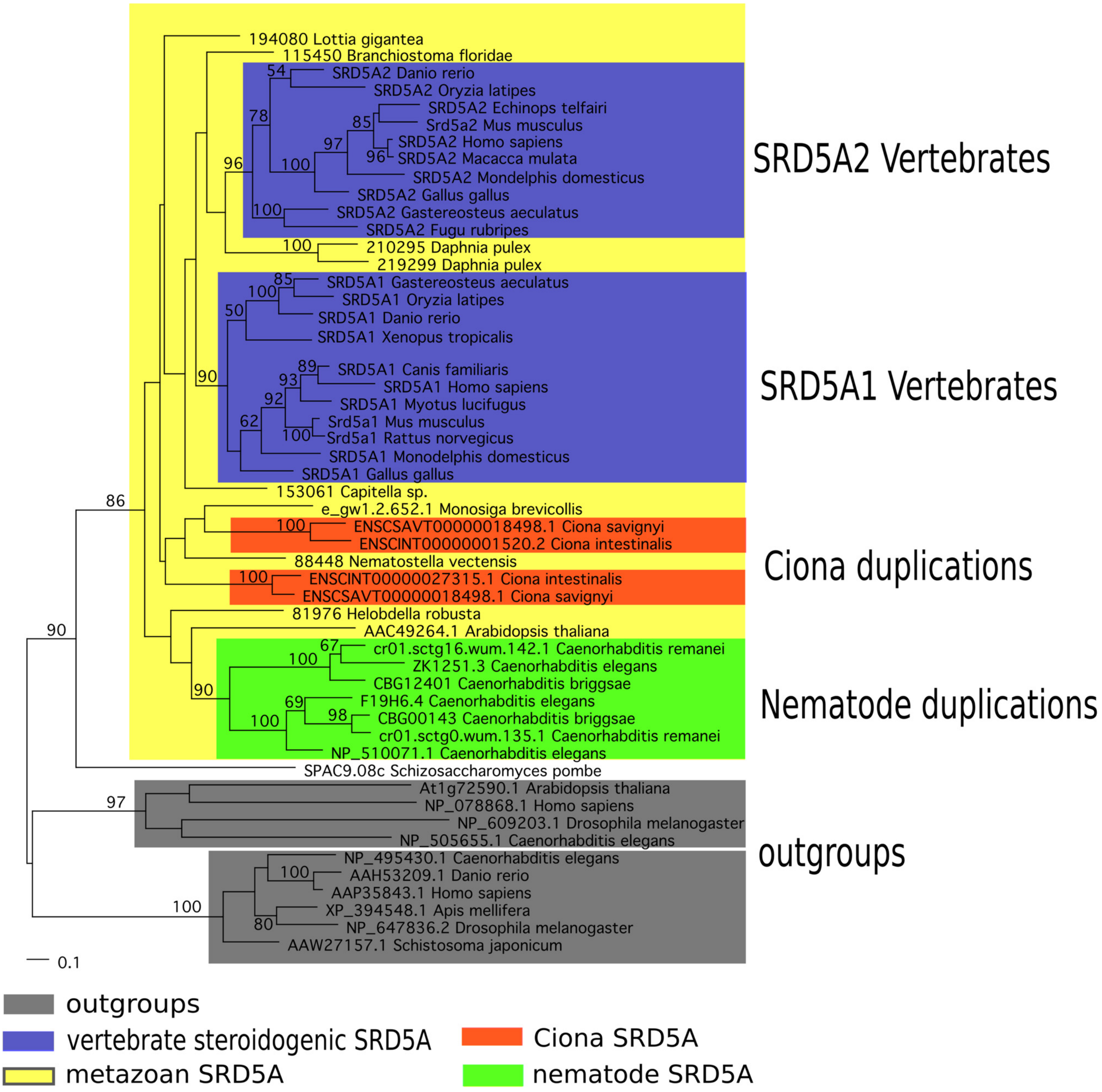

194080 Lottia gigantea
115450 Branchiostoma floridae
54
SRD5A2 Danio rerio
SRD5A2 Oryzia latipes
78
85
SRD5A2 Echinops telfairi
Srd5a2 Mus musculus
97
SRD5A2 Homo sapiens
96
SRD5A2 Macacca mulata
96
100
SRD5A2 Mondelphis domesticus
SRD5A2 Gallus gallus
100
SRD5A2 Gastereosteus aeculatus
SRD5A2 Fugu rubripes
SRD5A2 Vertebrates
100
210295 Daphnia pulex
219299 Daphnia pulex
85
SRD5A1 Gastereosteus aeculatus
100
SRD5A1 Oryzia latipes
50
SRD5A1 Danio rerio
SRD5A1 Xenopus tropicalis
89
SRD5A1 Canis familiaris
93
SRD5A1 Homo sapiens
92
SRD5A1 Myotus lucifugus
90
62
Srd5a1 Mus musculus
100
Srd5a1 Rattus norvegicus
SRD5A1 Monodelphis domesticus
SRD5A1 Gallus gallus
SRD5A1 Vertebrates
153061 Capitella sp.
86
e_gw1.2.652.1 Monosiga brevicollis
100
ENSCSAVT00000018498.1 Ciona savignyi
ENSCINT00000001520.2 Ciona intestinalis
88448 Nematostella vectensis
100
ENSCINT00000027315.1 Ciona intestinalis
ENSCSAVT00000018498.1 Ciona savignyi
Ciona duplications
81976 Helobdella robusta
90
AAC49264.1 Arabidopsis thaliana
67
cr01.sctg16.wum.142.1 Caenorhabditis remanei
100
ZK1251.3 Caenorhabditis elegans
CBG12401 Caenorhabditis briggsae
90
69
F19H6.4 Caenorhabditis elegans
98
CBG00143 Caenorhabditis briggsae
100
cr01.sctg0.wum.135.1 Caenorhabditis remanei
NP_510071.1 Caenorhabditis elegans
Nematode duplications
SPAC9.08c Schizosaccharomyces pombe
At1g72590.1 Arabidopsis thaliana
97
NP_078868.1 Homo sapiens
NP_609203.1 Drosophila melanogaster
NP_505655.1 Caenorhabditis elegans
NP_495430.1 Caenorhabditis elegans
100
AAH53209.1 Danio rerio
AAP35843.1 Homo sapiens
100
XP_394548.1 Apis mellifera
80
NP_647836.2 Drosophila melanogaster
AAW27157.1 Schistosoma japonicum
outgroups
0.1
outgroups
vertebrate steroidogenic SRD5A
metazoan SRD5A
Ciona SRD5A
nematode SRD5A

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)