Supplementary Material

METHODS

Orthology Assignment for Nuclear Receptors and Cytochrome p450s from Tiwary and Li

We determined the evolutionary relationships of the nuclear receptors and cytochrome p450s used in the study by Tiwary and Li (2009) using maximum-likelihood analyses. Because their dataset did not contain previously characterized sequences with which to root the analysis, we retrieved a set of protein sequences to represent the diversity within these two superfamilies. For the nuclear receptors, we retrieved sequences from a previous study for human, the fly *Drosophila melanogaster*, and the nematode *Caenorhabditis elegans* that represent the broad scale diversity of nuclear receptors for each of these species (Bertrand et al. 2004). For the cytochrome p450s, we selected human proteins from throughout the superfamily to represent the diversity of cytochrome p450s in animals (Supplemental Table 1). In addition, we included CYP19 from the cephalochordate *Branchiostoma floridae* (GenBank Accession ABA47317.1). We then aligned the sequences used by Tiwary and Li (2009) with the additional sequences using the multiple sequence alignment program Muscle (Edgar 2004). We trimmed the nuclear receptor alignment to the well-conserved DNA-binding and ligand-binding domains (~260 amino acids). For the cytochrome p450s, our alignment (~405 amino acids) included most of the protein with the exception of the more variable five- and three-prime end regions. We conducted maximum likelihood analyses using RAxML v. 7.0.4 (Stamatakis 2006) with the best-supported models (JTT+G for nuclear receptors, JTT for cytochrome p450s) determined by AIC criteria with ProtTest v1.4 (Abascal, Zardoya, and Posada 2005). Orthology assignment was inferred from the tree

topology and support for the relationships was determined with percentage of 1000 bootstraps.

Regressions of Protein Distances for Androgen Receptor and CYP19 from Tiwary and Li

From the phylogenetic trees generated above, we conducted three sets of regression analyses to test for significant correlations between protein distances for AR and CYP19: 1) all sequences, 2) only orthologous sequences identified with our phylogenetic analyses, and 3) the invertebrate sequences (none of which were supported as orthologs of either AR or CYP19). For each analysis we generated maximum-likelihood trees using the methods described above. From these trees, we generated matrices of pairwise protein distances with Patristic (Fourment and Gibbs 2006). We then regressed the distances for AR against CYP19 and tested for significant relationships for all three data sets (JMP 2002).

Testing for Coevolution of Androgen Receptor and CYP19 in Selected Vertebrates

To more fully test the hypothesis of coevolution between AR and CYP19, we employed multiple phylogenetic methods. We used these approaches on a reduced set of orthologs selected from eight vertebrate species distributed throughout the vertebrate subphylum (Tetraodon: *Tetraodon nigroviridis*, Fugu: *Takifugu rubripes*, Frog: *Xenopus tropicalis*, Chicken: *Gallus gallus*, Possum: *Monodelphis domestica*, Dog: *Canis familiaris*, Mouse: *Mus musculus*, and Human: *Homo sapiens*). Because our goal was to understand whether these two proteins have coevolved throughout the vertebrate lineage, we felt this was an appropriate scale to test for a relationship.

Within teleost fishes, both AR and CYP19 have duplicated at least once (Chiang et al. 2001; Douard et al. 2008). We conducted phylogenetic analyses containing both paralogs and selected the shorter branch representative from each species to avoid confounding effects of lineage-specific, rapid divergence in the other paralog (Douard et al. 2008). For both AR and CYP19, this represented a monophyletic group of genes. Lastly, we selected two sets of 'control' genes to separate true coevolutionary relationships from those due to shared evolutionary history and overall conservation in domains for particular gene families.

For the first set, we used two members of the same superfamily as AR and CYP19: the nuclear receptor germ cell nuclear factor (GCNF) and the cytochrome p450 CYP51 (see main text for justification, Supplemental Table 2 for sequences). Protein alignments were generated with similar methods to those used in the full data set discussed above and only included regions that could be aligned without ambiguity.

For the second set of control analyses, we retrieved protein sequences for glucokinase, glucagon, myoglobin, and erythropoietin from the same vertebrates as for the above analyses (see Supplemental Table 2 for sequences). We generated an alignment, maximum-likelihood trees, and patristic distance matrices using the same methods as for the other proteins. We compared molecular distances for these genes with distances from AR and CYP19 with regression analyses.

LITERATURE CITED

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

Bertrand, S., F. G. Brunet, H. Escriva, G. Parmentier, V. Laudet, and M. Robinson-Rechavi. 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. Molecular Biology and Evolution **21**:1923-1937.

Chiang, E. F.-L., Y.-L. Yan, Y. Guiguen, J. Postlethwait, and B.-c. Chung. 2001. Two cyp19 (p450 aromatase) genes on duplicated zebrafish chromosomes are expressed in ovary or brain. Molecular Biology and Evolution **18**:542-550.

Douard, V., F. Brunet, B. Boussau, I. Ahrens-Fath, V. Vlaeminck-Guillem, B. Haendler, V. Laudet, and Y. Guiguen. 2008. The fate of the duplicated androgen receptor in fishes: a late neofunctionalization event? BMC Evolutionary Biology **8**:336.

Edgar, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**:1792-1797.

Fourment, M., and M. Gibbs. 2006. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evolutionary Biology **6**:1.

JMP. 2002. SAS

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.

Tiwary, B. K., and W.-H. Li. 2009. Parallel evolution between aromatase and androgen receptor in the animal kingdom. Molecular Biology and Evolution **26**:123-129.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Maximum-likelihood analysis of the androgen receptor sequences used in the study by Tiwary and Li (2009). In our analysis we added additional sequences from a previous study by Bertrand et al. (2004) representing the full nuclear receptor (NR) superfamily to facilitate assessment of orthology. The tree was arbitrarily rooted with NR family 2, which forms a monophyletic clade and likely represents the ancestral NR family. Androgen receptors from a variety of vertebrates (indicated by gray box) form a strongly supported monophyletic clade (bootstrap = 95) and are positioned with other steroid-binding receptors in NR family 3. However, the invertebrate sequences used by Tiwary and Li (2009) represent diverse receptors distributed throughout the NR superfamily (arrows). Taxa names are abbreviated with the first three letters of the genus in capital letters and the first letter of the species in lower case (e.g., abbreviation for *Homo sapiens* is HOMs) for sequences from Tiwary and Li (2009) and in all capital letters for sequences from Bertrand et al. (2004).

Supplemental Figure 2. Maximum-likelihood analysis of the cytochrome p450 (CYP) sequences used by Tiwary and Li (2009). In our analysis we added additional sequences representing the diversity of the CYP superfamily to facilitate assessment of orthology (Supplemental Table 1). The tree was midpoint rooted. Aromatase (CYP19) sequences from a variety of vertebrates form a monophyletic clade with strong bootstrap support (bootstrap = 100). In addition, the inclusion of the CYP19 ortholog from the cephalochordate *Branchiostoma* forms an equally supported clade (indicated by gray box). However, the invertebrate sequences used by Tiwary and Li (2009) represent diverse CYP clans distributed throughout the CYP superfamily (arrows). Abbreviations

for taxa from Tiwary and Li (2009) are the same as in supp. fig. 1 and the CYPs from human are labeled ‘Hs.’

Supplemental Figure 3. Regressions of protein distances (A) AR and CYP51 and (B) GCNF and CYP19 from eight vertebrates using regressions of pairwise molecular distances (as in Figure 3).

Supplemental Figure 4. (A) Maximum likelihood analysis of glucokinase sequences from eight selected vertebrate species. The tree shows that the glucokinase sequence from *Xenopus* is a relatively long branch that is explained by a disproportionately high number of unique replacements for this species. (B) Regressions of protein distances for glucokinase and androgen receptor. The fit of this correlation increases considerably with the removal of the *Xenopus* sequence from the analysis. However, the relationship is significant regardless of whether *Xenopus* is included or not, consistent with the results from the other control genes we analyzed (see main text).

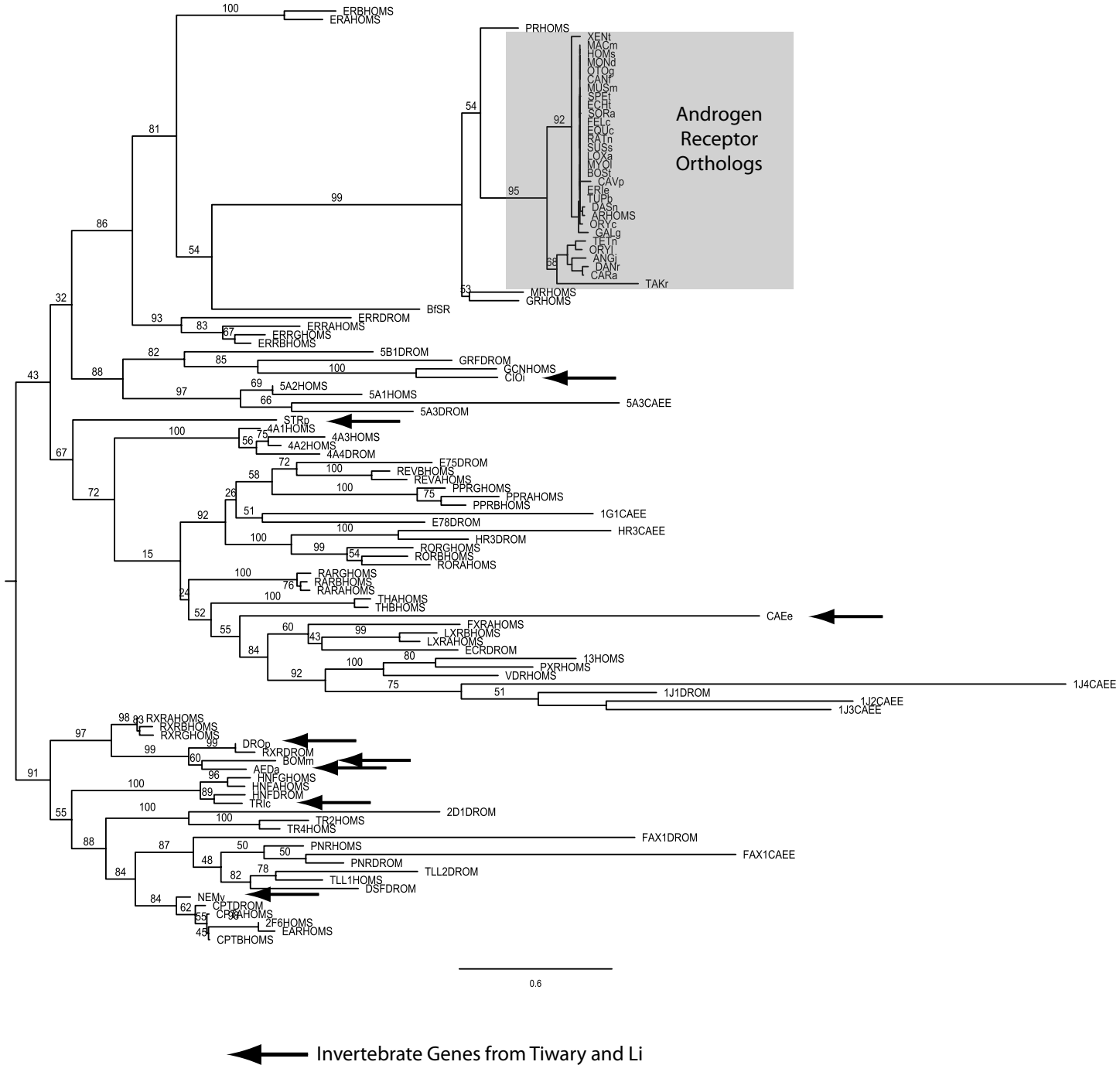
Supplemental Table 1. Human cytochrome p450 (CYP) proteins used for phylogenetic analysis of the CYPs from the study by Tiwary and Li (2009). We selected proteins that represent the broad-scale diversity of CYPs.

Species	Gene Name	NCBI Accession
Homo sapiens	CYP1A1	NP_000490.1
	CYP2A6	NP_000753.3
	CYP2C8	NP_000761.3
	CYP2D6	NP_000097.2
	CYP3A4	NP_059488.2
	CYP4B1	NP_000770.2
	CYP4F3	NP_000887.2
	CYP4V2	NP_997235.3
	CYP7A1	NP_000771.2
	CYP8B1	NP_004382.2
	CYP11A1	NP_000772.2
	CYP17A1	NP_000093.1
	CYP20A1	NP_803882.1
	CYP27A1	NP_000775.1

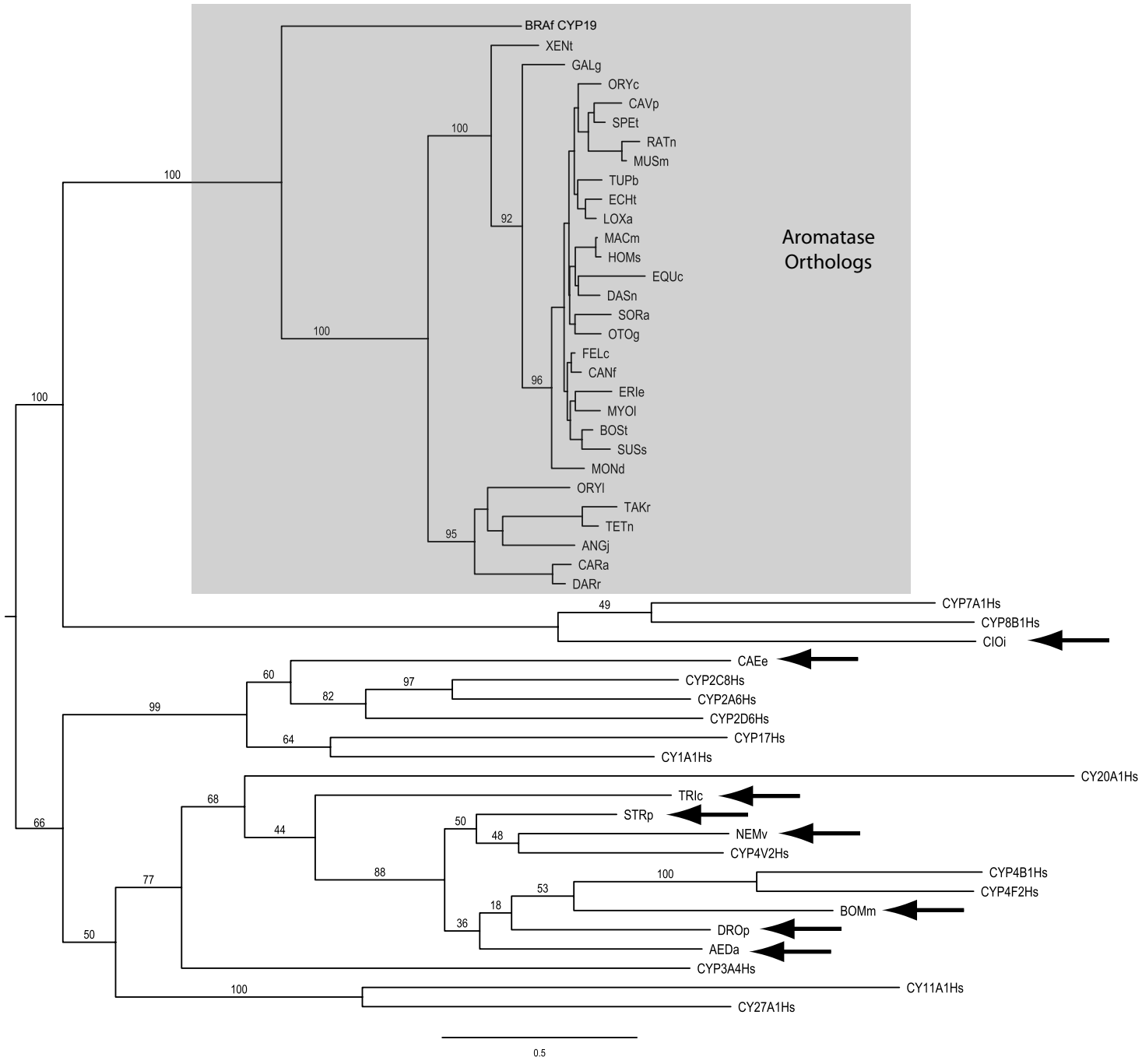
Supplemental Table 2. Accession numbers of GCNF, CYP51, and control genes and length of alignment for eight vertebrate species used in regression analyses.

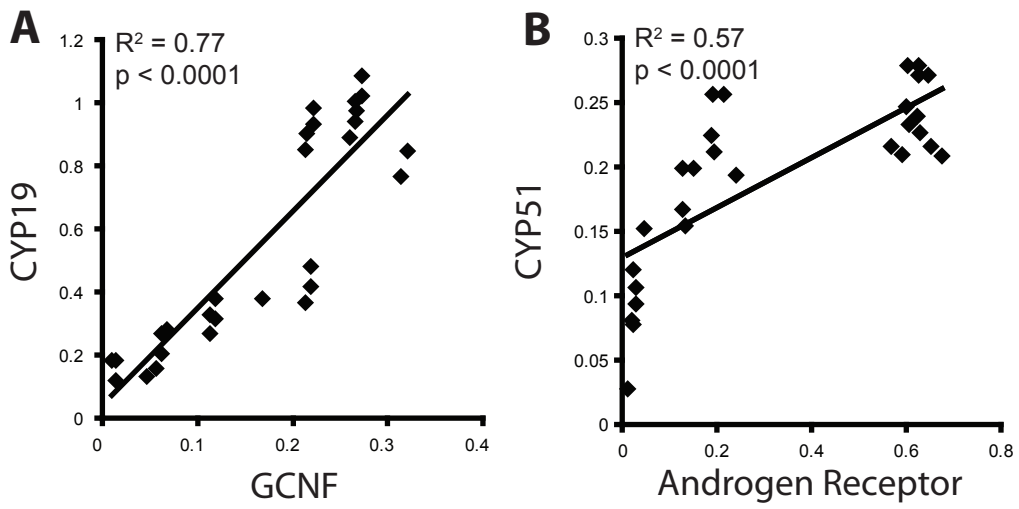
Species	GCNF	CYP51	glucokinase
Homo	ENSP00000420267	ENSP00000003100	ENSP00000384247
Mus	ENSMUSG00000063972	ENSMUSG00000001467	ENSMUSP00000105448
Canis	ENSCAFG00000020216	ENSCAFG00000001945	ENSCAFP00000004364
Monodelphis	ENSMODG00000019748	ENSMODG00000010610	ENSMODP00000020206
Gallus	ENSGALG00000001073	ENSGALG00000009365	ENSGALP00000001780 / AAM83106
Xenopus	ENSXETG00000008578	ENSXETG00000003432	ENSXETP00000041217
Takifugu	ENSTRUG00000006306	ENSTRUG00000013309	ENSTRUP00000042306
Tetraodon	ENSTNIG00000010587	ENSTNIG00000017535	ENSTNIP00000019018
Alignment length	423	441	450

Species	Glucagon	Myoglobin	Erythropoietin
Homo			
Mus			
Canis			
Monodelphis			
Gallus			
Xenopus			
Takifugu			
Tetraodon			
Alignment length			

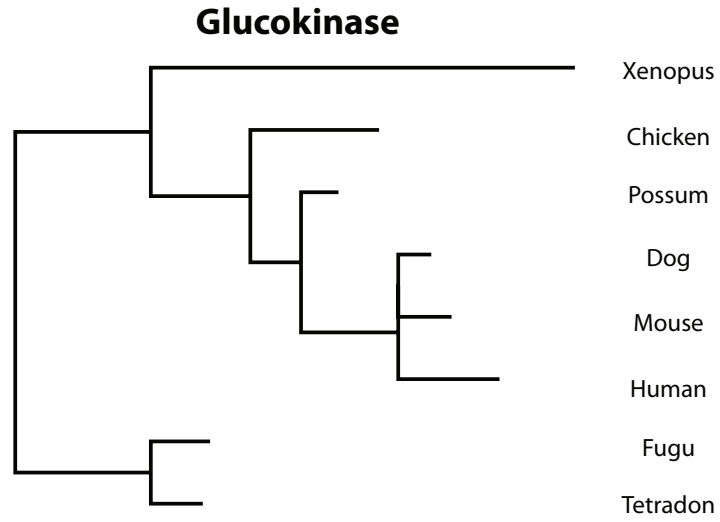


Invertebrate Genes from Tiwary and Li

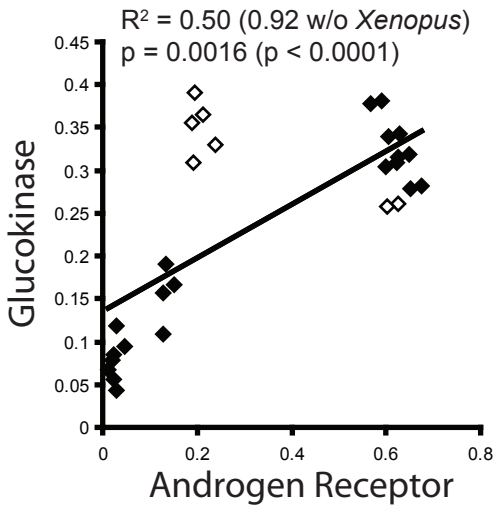

BRAf CYP19
XENt
GALg
ORYc
CAVp
SPEt
RATn
MUSm
TUPb
ECHt
LOXa
MACm
HOMs
EQUc
DASn
SORa
OTOg
FELc
CANf
ERIe
MYOl
BOSt
SUSs
MONd
ORYl
TAKr
TETn
ANGj
CARa
DARr
Aromatase Orthologs
CYP7A1Hs
CYP8B1Hs
ClOi
CAEe
CYP2C8Hs
CYP2A6Hs
CYP2D6Hs
CYP17Hs
CY1A1Hs
CY20A1Hs
TRIc
STRp
NEMv
CYP4V2Hs
CYP4B1Hs
CYP4F2Hs
BOMm
DROp
AEDa
CYP3A4Hs
CY11A1Hs
CY27A1Hs
0.5
Invertebrate Genes from Tiwary and Li


A
R² = 0.77
p < 0.0001
CYP19
GCNF
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
B
R² = 0.57
p < 0.0001
CYP51
Androgen Receptor
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8

A


Glucokinase
Xenopus
Chicken
Possum
Dog
Mouse
Human
Fugu
Tetradon

B


R² = 0.50 (0.92 w/o Xenopus)
p = 0.0016 (p < 0.0001)
Glucokinase
Androgen Receptor
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0
0.2
0.4
0.6
0.8