Supporting Online Information

For:

High-throughput mapping of he promoters of the mouse olfactory receptor genes reveals a new type of mammalian promoters

E. Josephine Clowney[1], Angeliki Magklara[2], Bradley M. Colquitt[3], Nidhi Pathak[4], Robert, P., Lane[4] and Stavros Lomvardas [1,2,3*]

[1]Program in Biomedical Sciences, University of California, San Francisco, CA 94158, USA.

[2]Department of Anatomy, University of California, San Francisco, CA 94158, USA.

[3]Program in Neurosciences, University of California, San Francisco, CA 94158, USA.

[4]Department of Molecular Biology and Biochemistry, Wesleyan University, Middletown CT, 06457, USA.

*Corresponding Author. Email: stavros.lomvardas@ucsf.edu: T (415)514-4811; F (415)476-1974

Supplemental Methods

PRIMER SEQUENCES FOR RLM-RACE

Primer Name	IUB Sequence
135R (TMIII)	CADATHGCHACRTAHWTRTCRTAHGCCATGGATCCG
214 (TMV)	GRTADATRAAIGGRTTIARCATNGG
b6 (TMVII)	GSWISWICCIACRAARAARTAIATRAAIGGRTT
P8 (TMVII)	RTTICKIARISWRTAIATRAAIGGRTT
P26R (TMIII)	CAIATIGCIACRTAICGRTCRTAIGC
P27 (TMVI)	ACIACIGAIAGRTGIGAISCRCAIGT

FIGURE 2 CATEGORIZATIONS

In figure 2C, promoters <=40%GC or >=65% GC were selected. 40% was chosen as the lower cutoff to include 75% of OR genes; 65% was chose for the upper cutoff so as to include similar numbers of genes (~2000) in each category. Duplicates were removed from each list and all annotations for each gene were collected from DAVID, UCSC, MGI, and Weitzmann GeneCards (Bult et al. 2010; Huang et al. 2009; Kent et al. 2002; Safran et al. 2002). Each gene was assigned exclusively to one category amongst those listed below; assignment was hierarchical such that proteins fitting more than one category were assigned preferentially to starred groups. Percentage of the AT- or GC-rich promoters with each function (except those marked by an ampersand below) was then plotted. To assess statistical significance, we added the AT and GC categories together, counted total occurrences of each function, calculated an expected number in AT and GC based on hypothetical random distribution around the murine AT/GC

composition midpoint, and compared these values to the actual findings. The “Cellular Metabolism” functions accounted for 8% of GC-rich and 1.5% of AT-rich promoters; the “Transport” functions accounted for 12% of GC-rich and 2.5% of AT-rich promoters.

*Chemosensation	GPCRs sensing foreign compounds
*Barriers and Immunity	Secreted and transmembrane proteins with immune function or forming barrier epithelia
*Xenobiotic Metabolism	Primary metabolism of foreign substances
*Nuclear Process	All nuclear proteins
Signal Transduction, Cell Cycle, Morphogens	All signaling compounds not contained in above categories
&Cellular Metabolism and Synthesis	Non-xenobiotic metabolism, energy production, and non-nuclear macromolecule synthesis
&Transport, Structure, and Motility	Transport organelles, cilia and flagella, cytoskeleton
&Miscellaneous	
&Unknown	

For Figure 2D and S1, existing KEGG (Xenobiotic Metabolism by Cytochrome p450) and GO (Transcription Factor Activity, Cell Cycle) categories were used when possible (Kanehisa and Goto 2000; Ashburner et al. 2000). When no existing category captured a function of interest (Chemosensation, Morphogen, Innate Defense and Barriers), a category was constructed by collecting all RefGene names with applicable prefixes, listed in supplemental methods (DeFranco et al. 2007). GC content distribution was plotted as percent of promoters of each functional category in 3%GC bins (26-28%, 29-31%, etc). For figure 2D, Chemosensory/Defense/Xenobiotic and Cell Cycle/Transcription Factor/Morphogen categories were combined by averaging the three

percentages at each GC content to give equal weight to categories with varying numbers of constituent genes. Fig. S1 shows each functional distribution individually.

Chemosensation	**Innate Defense / Barrier**		**Morphogen**
V1r	Clec	Skint	Wnt
V2r	Def	Sprr	Fgf
Vmn1r	Ifn	Tlr	Bmp
Vmn2r	Klr	Lman	Tgf
Fpr	Krt	Marco	Shh
Olfr	Krtap	Mbl	Dhh
Tas1r	Lce	Apobec	Ihh
Tas2r	Ms4	C1	Egf
Taar	Muc	Colec	Egfr
	Nirp	Masp	Fgfr
	Reg	Cd14	Bmpr
	Serpin	Fc	
	Cd36		

CLOWNEY_Figure S1

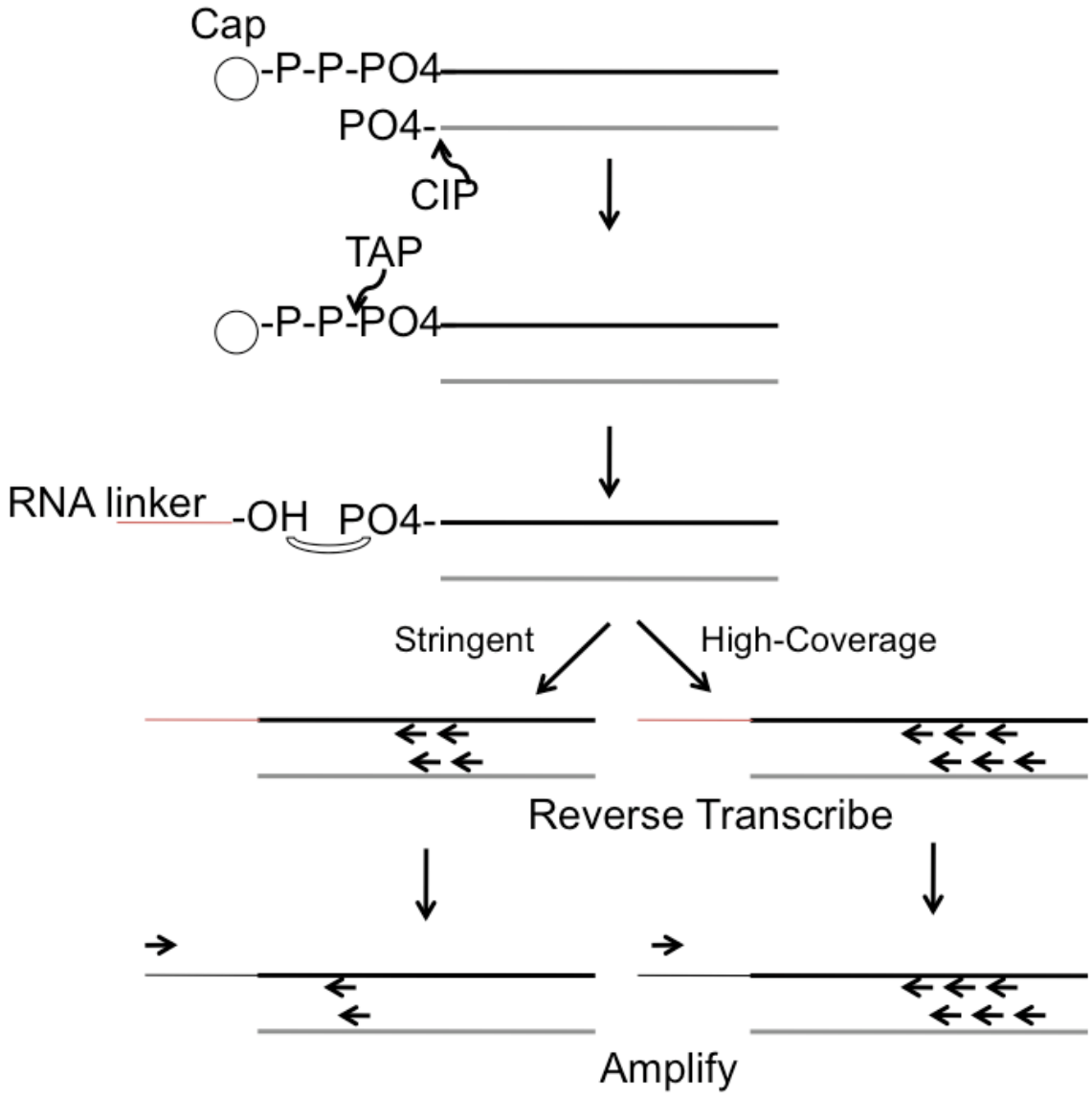


Figure S1. Schematic of RLM-RACE method used for promoter mapping

Adapted from Michaeloski et al., 2006.

CLOWNEY_Figure S2

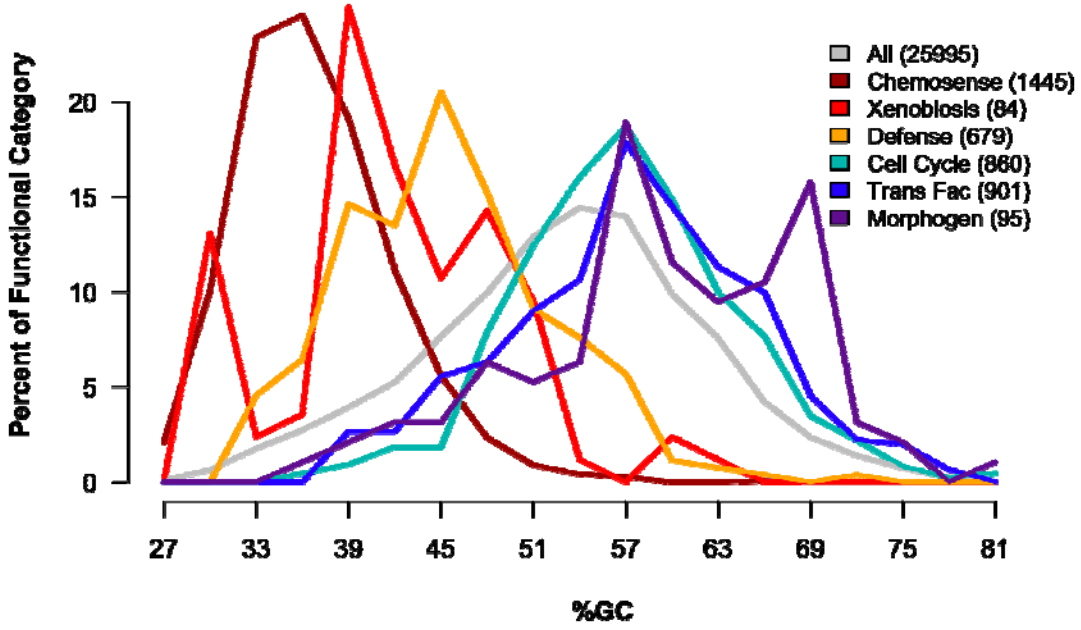


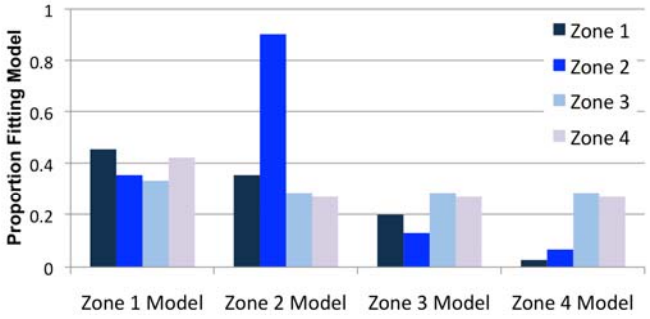
Figure S2 Percent GC distribution of functional gene categories. %GC distribution (in 3% bins) of promoters is shown as percentage of total promoters in the functional category. For category definitions, see methods. Number of promoters in each category is shown in the legend. “All” (from which functional categories are sampled) does not violate normality hypothesis (Anderson-Darling test, p-value of non-normality hypothesis >.05); means of pairs of categories are all significantly different ($p<.0001$ except Morphogen vs Cell Cycle=.004, Morphogen vs Transcription Factor=.025; 2-tailed, unpaired t-test) except Defense vs. Xenobiosis and Cell Cycle vs. Transcription Factor.

CLOWNEY_Figure S3

A

	TF Family	# Zone Proms with Site	# Sites	# OR Proms with Site	Enrichment: Zone vs All ORs	Z Score: Zone vs All ORs
Zone1	PTBP	65(/110)	165	826	1.21	2.42
	HOXH	69	134	659	1.2	9
Zone2	RORA	24 (/31)	43	589	1.48	2.49
	TSF	28	75	820	1.27	2
	PDX1	28	96	822	1.22	1.92
	RXRF	25	61	740	1.21	1.43
Zone3	RXRF	16(/21)	55	740	1.76	4.17
	STAT	16	62	751	1.48	3.01
	NR2F	15	51	796	1.42	2.42
	BCDF	15	46	824	1.3	1.69
	IKRS	13	15	473	1.24	0.68
	DLXF	15	68	818	1.2	1.44
Zone4	YTBP	17(/26)	53	606	2.03	5.15
	PAX8	15	24	475	1.51	1.9
	IRFF	19	55	790	1.36	2.21
	CEBP	21	49	776	1.33	1.94
	PAX6	17	38	691	1.29	1.47
	HOXH	15	33	659	1.22	1.06
	PTBP	19	40	826	1.21	1.13
	NFAT	18	52	758	1.2	1.25

B



C

	P-Value (Distribution of Known in Model Predictions)	Number Uniquely Predicted in Whole OR Family
Zone 1 Model	8.E-01	161
Zone 2 Model	1.E-04	111
Zone 3 Model	6.E-01	57
Zone 4 Model	7.E-05	0

D

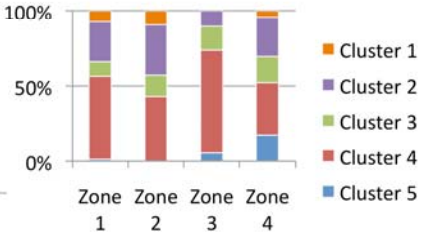


Figure S3 Weak zonal promoter models based on TFBS. 98 class 2 and 100 class 1 ORs with known zone were searched for TFBS enriched relative to all OR promoters after repeat-masking (*A*). Motifs listed are present in greater than half the promoters for a particular zone and enriched at least 20% over all OR promoters. We created zonal models by selecting promoters from the whole 1085 set that contained every enriched site for a particular zone in (*A*) and then checked the distribution of promoters with known zone in these model-predicted groups (*B*). Only models for zones 2 and 4 deviated significantly from random distribution (*C*), and only the zone 2 model can correctly re-identify the promoters used to build it. (*D*) Promoters with known zone were clustered by this method using various parameters. No parameter set was found under which promoters of a particular zone clustered coherently. One representative trial is shown.

CLOWNEY_Figure S4

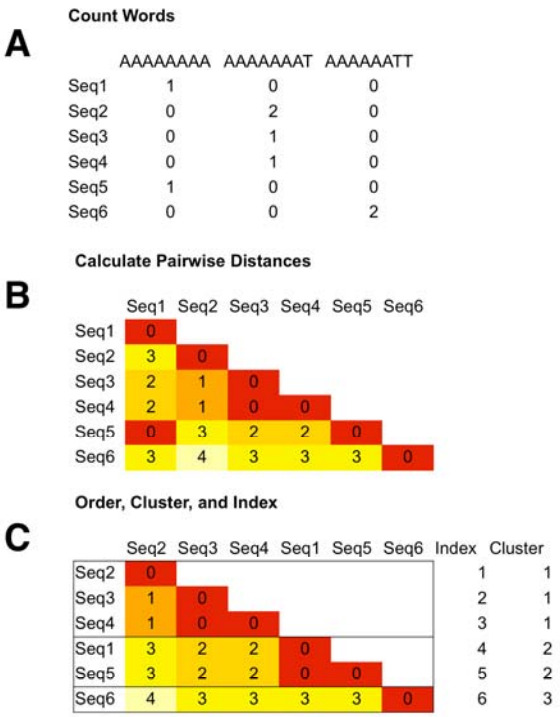


Figure S4: Word-based clustering model of promoter relatedness. (*A*) First, the procedure builds a dictionary of k-mers present in each input sequence and the number of times they occur. We worked with 8-mers. Reverse complements can be collapsed in this stage. (*B*) Next, the wordcounts for each sequence are compared to each other sequence to generate a pairwise distance matrix. Here, higher numbers (and lighter shades) correspond to greater differences between sequences. For speed, we calculated distances using only the most common words in the whole set of sequences. (*C*) Finally, sequences are ordered and clustered according to the distance matrix, so that more similar sequences are close together. Distance measures and ordering/clustering were performed using the R implementation of the Hopach package. The wordcount function was written in Perl.

CLOWNEY_Table S1

Gene Family	Description	# Observed	# Total	Enrich-ment	P Value	Mean %GC
Olfr	Olfactory receptor	846	1085	7.5	0.E+00	37
V1r/Vmn1r	Vomernasal receptor	115	115	9.6	3.E-218	36
V2r/Vmn2r	Vomernasal receptor	117	140	8.1	1.E-177	36
Prl	Prolactin	27	27	9.6	1.E-52	37
Klr	Killer cell receptor	24	30	7.7	6.E-36	38
Tas2r	Taste receptor	25	37	6.5	4.E-30	36
Skint	Selection and upkeep of intraepithelial T-cells	11	11	9.6	2.E-22	35
Pcdh	Protocadherin	27	66	3.9	4.E-16	44
Csn	Casein	5	5	9.6	5.E-11	34
Amy	Amylase	5	5	9.6	5.E-11	36
Sprr	Small proline rich region	9	15	5.8	3.E-10	41
Fpr	Formyl peptide receptor	5	6	8.0	5.E-09	38
Ifn	Interferon	10	21	4.6	2.E-08	45
Reg	Regenerating islet derived	5	7	6.9	1.E-07	40
Adam	A disintegrin and metalloprotease domain	11	30	3.5	2.E-06	47
Crisp	Cysteine-rich secretory protein	3	4	7.2	2.E-05	34
Hbb	Hemoglobin	3	4	7.2	2.E-05	41
Taar	Trace amine receptor	6	15	3.9	2.E-04	40
Akr1	Aldo-keto reductase	6	15	3.9	2.E-04	44
Tmprss	Transmembrane serine protease	6	15	3.9	2.E-04	47
Sult	Sulfotransferase	5	12	4.0	4.E-04	45
Ugt	UDP glucuronosyltransferase	7	21	3.2	6.E-04	43
Mrg	MAS-related GPR	6	18	3.2	1.E-03	44
Cyp2	Cytochrome p450 family 2	12	50	2.3	2.E-03	46

Mup	Major urinary protein	5	14	3.4	2.E-03	41
Krtap	Keratin-associated protein	9	40	2.2	1.E-02	45
Serpin	Serine protease inhibitor	12	60	1.9	1.E-02	46
Nlrp	NACHT, LRR and PYD containing protein	5	20	2.4	3.E-02	44
Def	Defensin	6	53			45
Clec	C-type lectin	5	31			46
Tas1r	Taste receptor	0	3			54
All	All families above	1338	1975			41.5
Total	**Unique Murine Gene Names (RefGene)**	**2206**	**21281**			**53**

Table S1. Gene families enriched in AT-rich promoters. Families enriched (p<.05, Chi-square test) in the <=40%GC group as compared to expected random representation of 10.4% of each family (2206/21281 unique RefGene names). “Enrichment” is (# Observed)/(# Expected). Some categories (e.g. Def, Clec) were not significantly enriched in the <=40% category but still have a skewed mean distribution. Chemosensory families are shaded blue, defense and barriers shaded pink, xenobiotic metabolism shaded green, misc unshaded. Genes without mapped transcription start sites were not excluded.

CLOWNEY_Table S2

Gene Family	Mouse	Human	Zebrafish	Reference
OR	1391 (328)	802 (414)	176 (21)	Nei et al. 2008
V1R	308 (121)	120 (115)	2 (0)	Nei et al. 2008
V2R	279 (158)	20 (20)	52 (8)	Nei et al. 2008
Krtap	188 (13)	122 (21)	0	Wu et al. 2008
Defensin	94 (22)	66 (13)	3	Zou et al. 2007
Cyp2-4 (Unstable)	92 (55)	165 (87)	63 (14)	Nelson et al. 2009, Thomas et al. 2007
Pcdh	65 (2)	70 (3)	58	Wu et al. 2005
Serpin	61 (3)	38 (1)	15	
Mup	43 (22)	1 (1)	0	Logan et al. 2008
T2R	41 (6)	36 (11)	4 (0)	Nei et al. 2008
Klr	35	0	0	DeFranco et al. 2007, Yoder et al. 2009
Ifn	34 (6)	34 (11)	4	Zou et al. 2007
Cyp1, 5+ (stable)	30	24	39	Nelson et al. 2009, Thomas et al. 2007
Prl	27 (2)	1	2	Simmons et al. 2008
TAAR	16 (1)	9 (3)	119 (10)	Nei et al. 2008

Ugt	26 (4)	30 (8)	45 (5)	Huang et al. 2010
Csn	5	2	0	
T1R	3	3	1	Nei et al. 2008

Table S2. Size of gene families of interest in mouse, human and zebrafish. Vega and UCSC annotations and papers of interest were searched for family members (number of which are pseudogenes in parentheses) (Aherst et al. 2005; Hinrichs et al. 2006). Pcdh and Ugt families produce diverse gene products by alternative splicing and have undergone differential expansion of subfamilies in mammal vs. fish lineage (Wu 2005; Huang and Wu 2010). Klr (aka Ly49) genes in mice are functionally similar to KLR genes in human and NITR genes in fish but do not share evolutionary ancestry (DeFranco et al. 2007; Yoder 2009). Numbers differ from counts of RefGene prefixes used in other analyses.

CLOWNEY Table S3

	Olfr	Non-Olfr	P value of improvement over previous filter
Unfiltered	1040	~21000	
TK20	996	1048	~0
TK20 + AT-rich	682	253	1x10^-49
TK20 + AT-rich + O/E	392	110	1x10^-2

Table S3. **Numbers of Olfr and non-Olfr promoters fitting model.** Total genes passing each filter described in Figure 3C-D. P values are observed proportion of Olfr and non-Olfr genes passing the filter vs expected random distribution; ATrich and O/E filter P values are calculated based on improvement of the previous filter, not based on enrichment vs. no filtering.

CLOWNEY_Table S4

Family	Type I Enrichment	Type II Enrichment	All OR promoter enrichment
PAX1	1.16	.76	.81
NBRE	1.13	.84	.89
GABF	.85	1.27	1.19

Table S4. Genomatix TFBS differentially enriched in Type I and Type II OR promoters

CLOWNEY_Table S4

Matrix Family	Most Expressed # Seq with Site (of 51)	Most Expressed # Site	Most Expressed Promoter Enrichment	Least Expressed # Seq with Site (of 53)	Least Expressed # Site	Least Expressed Promoter Enrichment	Most/ Least
BRNF	50	601	2.47	53	570	2.26	1.09
ARID	43	141	2.52	45	115	1.98	1.27
DLXF	44	202	2.60	44	159	1.97	1.32
LHXF	49	560	2.53	52	493	2.14	1.18
NKX6	47	220	2.31	52	206	2.08	1.11
BRN5	49	376	2.59	51	329	2.18	1.19
CART	49	427	2.36	52	388	2.07	1.14
PDX1	43	183	2.63	46	145	2.01	1.31
HBOX	51	425	2.26	53	383	1.96	1.15
CDXF	50	167	2.13	48	166	2.04	1.04
OCT1	51	622	2.25	53	623	2.17	1.04
HOXF	50	496	2.02	52	503	1.97	1.03
MEF2	42	128	1.86	47	148	2.07	0.90
VTBP	51	391	2.13	53	322	1.69	1.26
SATB	36	117	3.75	41	100	3.09	1.21
PIT1	45	181	2.93	38	136	2.12	1.38
ATBF	38	123	3.09	43	89	2.15	1.44
PAXH	33	113	2.56	33	100	2.18	1.17
NKX1	37	105	2.72	27	70	1.75	1.55
HNF1	36	75	1.35	44	137	2.37	0.57

Table S4. Genomatix TFBS enrichment in promoters of 5% most expressed and 5% least expressed OR genes by RNASeq in OMP+ neurons

Supplemental References:

Ashurst JL et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33**, D459-65.

DeFranco AL, Locksely R, Roberston M. 2007. *Immunity: The Immune Respose in Infectious and Inflammatory Disease*, 350 (Oxford University Press, Oxford, UK).

Hinrichs, AS et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-8.

Huang H, Wu Q. 2010. Cloning and comparative analyses of the zebrafish Ugt repertoire reveal its evolutionary diversity. *PLoS One* **5**, e9144.

Logan DW, Marton TF, Stowers L. 2008. Species specificity in major urinary proteins by parallel evolution. *PLoS One* **3**, e3280.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet* **9**, 951-63.

Nelson DR. 2009. The cytochrome p450 homepage. *Hum Genomics* **4**, 59-65.

Simmons DG, Rawn S, Davies A, Hughes M, Cross JC. 2008. Spatial and temporal expression of the 23 murine Prolactin/Placental Lactogen-related genes is not associated with their position in the locus. *BMC Genomics* **9**, 352.

Thomas JH. 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet* **3**, e67.

Wu DD, Irwin DM, Zhang YP. 2008. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. *BMC Evol Biol* **8**, 241.

Wu Q. 2005. Comparative genomics and diversifying selection of the clustered vertebrate protocadherin genes. *Genetics* **169**, 2179-88.

Zou J, Mercier C, Koussounadis A, Secombes C. 2007. Discovery of multiple beta-defensin like homologues in teleost fish. *Mol Immunol* **44**, 638-47.

Yoder JA. 2009. Form, function and phylogenetics of NITRs in bony fish. *Dev Comp Immunol* **33**, 135-44.

Zou J, Tafalla C, Truckle J, Secombes CJ. 2007. Identification of a second group of type I IFNs in fish sheds light on IFN evolution in vertebrates. *J Immunol* **179**, 3859-71.