Supplementary material for "**Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system**"

Katherine Belov[1*], Claire E. Sanderson[1], Janine E. Deakin[2], Emily S.W. Wong[1], Daniel Assange[3], Kaighin A. McColl[3], Alex Gout[3,4], Bernard de Bono[5], Terence P. Speed[3], John Trowsdale[5], Anthony T. Papenfuss[3]

1. Faculty of Veterinary Science, University of Sydney, Sydney, Australia
2. ARC Centre for Kangaroo Genomics, Research School of Biological Sciences, The Australian National University, Canberra, Australia
3. Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Australia
4. Department of Medical Biology, The University of Melbourne, Parkville, Australia
5. Immunology Division, University of Cambridge, Cambridge, UK

*Corresponding author: K. Belov, Faculty of Veterinary Science, University of Sydney, NSW 2006, Australia ph 61 2 9351 3454, fx 61 2 9351 3957, email kbelov@vetsci.usyd.edu.au

MHC paralogous regions

Only 36 of the 114 genes in the opossum MHC have paralogs in one of the three paralogous regions (Supplementary Table 1). Genes represented in at least three of the four paralogous regions (13 genes) were used to compare gene order, revealing rearrangements between the four regions in opossum.

Table 1: MHC genes with paralogs on opossum chromosomes 1, 2 and 3, corresponding to MHC paralogous regions on human chromosomes 9, 1 and 19 respectively.

MHC	Chromosome 1 (Human Chr 9)	Chromosome 2 (Human Chr 1)	Chromosome 3 (Human Chr 19)
AGPAT1	*AGPAT2*		
AIF1	*C9orf58*		
ATP6V1G2	*ATP6V1G1*	*ATP6V1G3*	
B3GALT4		*B3GALT2*	
BAT1			*DDX39*
BAT2	*KIAA0515*	*BAT2D1*	
BRD2	*BRD3*	*BRDT*	*BRD4*
C4	*C5*		*C3*
SLC44A4		*SLC44A5*	*SLC44A2*
CLIC1	*CLIC3*	*CLIC4*	
COL11A2	*COL5A1*	*COL11A1*	*COL5A3*
CREBL1		*ATF6*	
DDAH2		*DDAH1*	
DDR1		*DDR2*	
EGFL8	*EGFL7*		
EHMT2	*EHMT1*		
GPX5			*GPX4*
MHC Class I		*CD1*	
HSPA1A	*HSPA5*		
MDC1		*PRG4*	
NOTCH4	*NOTCH1*	*NOTCH2*	*NOTCH3*
PBX2	*PBX3*	*PBX1*	*PBX4*
PHF1		*MTF2*	
PRSS16	*DPP7*		
PSMB9	*PSMB7*		
RGL2	*RALGDS*	*RGL1*	*RGL3*
RING1		*RNF2*	
RXRB	*RXRA*	*RXRG*	
SYNGAP1		*RASAL2*	
TAP	*ABCA2*		
TNF/LTA/LTB	*TNFSF8/TNFSF15*	*TNFSF4*	*CD70/TNFSF9/ TNFSF14/*
TNXB	*TNC*	*TNR*	

Table 2. Summary of opossum antimicrobial peptides.

Gene Name	Chromosomal location	Strand	Precursor (aa)	Mature peptide (aa)	Signal peptide	Net charge
CATH1	6:171651887-171654692	+	179	52	Y	+15
CATH2	6:171685696-171688545	+	183	56	Y	+16
CATH3	Un:9535394-9538034	+	157	29	Y	+6
CATH4	Un:10967782-10970695	+	144	22	Y	+6
CATH5	Un:11018996-11021899	+	160	35	Y	+4
CATH6	Un:10918443 -10921985	-	181	54	Y	+15
CATH7	Un:10893843-10897043	-	181	54	Y	+15
CATH8	Un:9632230-9639391	-	123*	-	Y	-
CATH9	Un:9562756-9566809	-	186	43	Y	+11
CATH10	Un:9548850-9550073	+	96*	49	Y	0
CATH11	Un:9552820-9556577	+	164	36	Y	+10
CATH12	6:171694019-171696768	+	177	50	Y	+6
DEFA1	1:559105867-559109247	+	91	33	Y	+7
DEFB1	1:557806052-557808079	-	84	61	Y	+3
DEFB2	1:557806052-557808079	-	84	61	Y	+3
DEFB3	1:557930103-557934790	-	90	68	Y	+6
DEFB4	1:558170800-558178626	-	62	37	Y	+4
DEFB5	1:558230783-558238488	-	62	41	Y	+2
DEFB6	1:558471629-558473851	-	104	87	Y	-4
DEFB7	1:558585282-558593538	-	77	58	Y	+3
DEFB8	1:558652397-558660073	-	62	41	Y	-1
DEFB9	1:558738001-558740316	-	62	41	Y	+6
DEFB10	1:558763225-558770469	-	62	41	Y	+2
DEFB11	1:558813928-558814053	-	93	74	Y	+1
DEFB12	1:558813928-558814053	-	88	-	N	+4

DEFB13	1:558844942-558845067	-	78	78	N	+9
DEFB14	1:558844942-558845067	-	78	-	N	+10
DEFB15	1:558870902-558878768	-	62	41	Y	+5
DEFB16	1:558907522-558909890	-	80	59	Y	+3
DEFB17	1:559006360-559008431	-	98	76	Y	+4
DEFB18	1:559075594-559077804	+	100	83	Y	-1
DEFB19	1:559124410-559126382	-	94	75	Y	+3
DEFB20	1:559212883-559215610	+	64	40	Y	+7
DEFB21	1:559275646-559279649	-	67	45	Y	+9
DEFB22	1:559370768 559370857	-	30*	-	-	+2
DEFB23	1:559370771-559374978	-	69	47	Y	+2
DEFB24	1:559396427-559396552	+	61	45	Y	+7
DEFB25	1:559415317-559421415	+	62	40	Y	+4
DEFB26	1:559430339-559432715	+	69	47	Y	+5
DEFB27	1:559452653-559469465	+	66	47	Y	0
DEFB28	1:559511311-559522123	+	69	50	Y	+7
DEFB29	Un:122945844-122946023	-	96	-	N	+4
DEFB30	2:297138592-297146394	-	77	55	Y	0
DEFB31	2:297209052-297209156	-	35*	-	-	+2
DEFB32	2:297279692-297279790	-	33*	-	-	0
DEFB33	2:297403241-297405062	-	70	56	Y	+3
DEFB34	1:422345997-422346098	+	34*	-	-	+6
DEFB35	1:422395278-422396457	-	70	49	Y	-4
DEFB36	1:422535545-422535646	+	34*	-	-	+7
DEFB37	1: 422655284-422656341	-	70	48	Y	+6

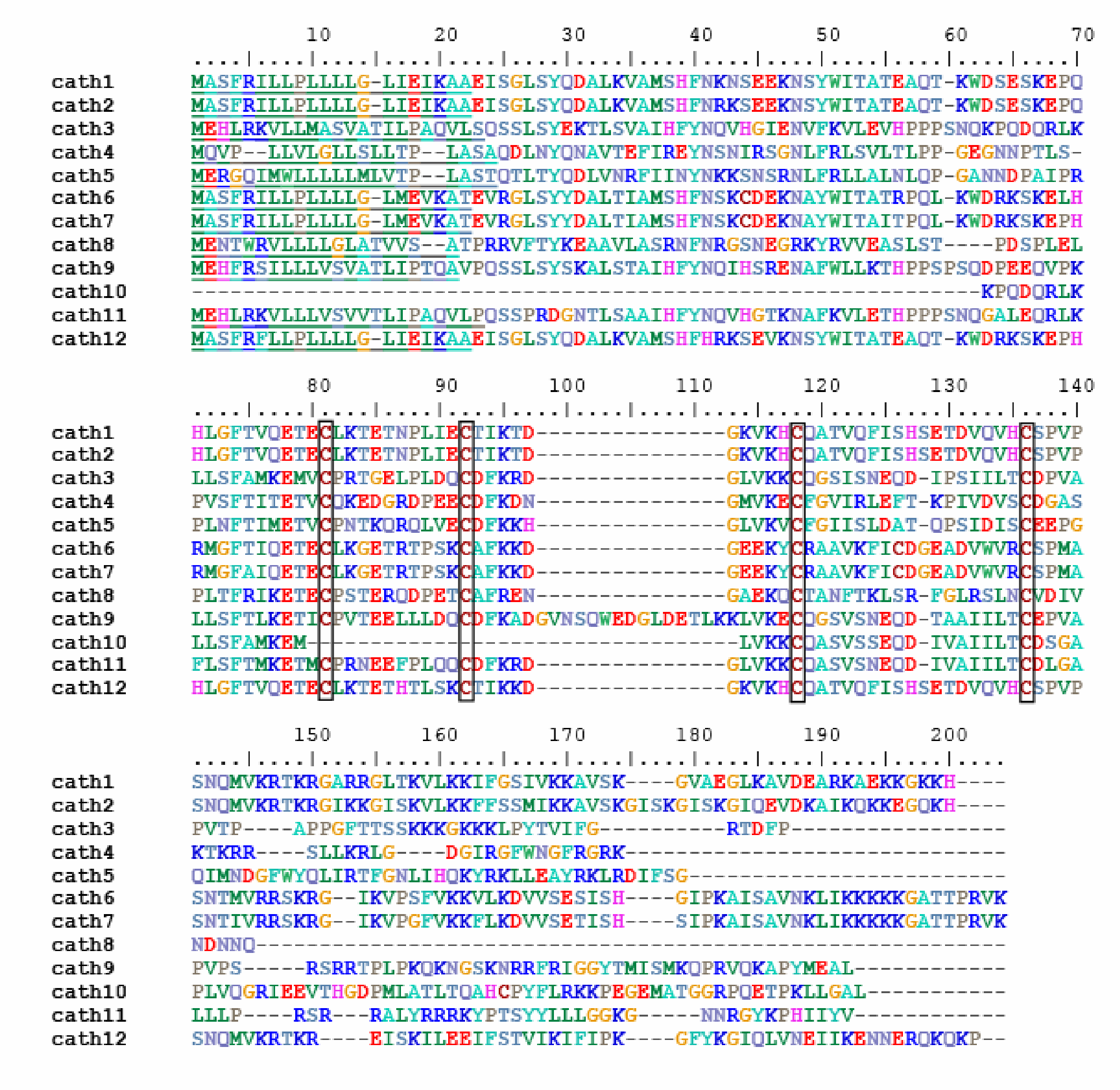
Cathelicidins

Cathelicidins exhibit broad spectrum antibiotic activity against several gram-positive and gram-negative bacteria, fungi, protozoa and enveloped viruses (Zaiou and Gallo 2002). They can bind and neutralize endotoxin and induce chemotaxis of neutrophils, T cells and monocytes (Giacometti et al. 2004). Some cathelicidins can act as immune modulators and mediators of inflammation, playing a part in cell proliferation and migration, wound healing, angiogenesis and the release of cytokines and histamine (Bals and Wilson 2003). Cathelicidins were previously only known in mammals, but have recently been discovered in chickens, salmonids and hagfish (Chang et al. 2006; Uzzell et al. 2003; Xiao et al. 2006).

Cathelicidins are characterized by a highly conserved prosequence at the N-terminus and a structurally variable mature peptide at the C-terminus. The conserved region (cathelin domain) contains four cysteines which form two disulphide bonds. Predicted opossum cathelicidin genes share common characteristics with cathelicidin genes found in all vertebrates, including four exons and three introns (Xiao et al. 2006), except *MdoCATH8* which is missing the terminal exon and *MdoCATH10* which is missing the first exon (and hence signal sequence). An alignment of cathelicidins is shown in Supplementary Figure 1, and a summary of the AMP gene features are shown in Supplementary Table 2. The first three exons contain the conserved preproregion including signal peptide and the cathelin domain. All predicted peptides contain a signal peptide as predicted by SignalP. The opossum signal peptides range from 19-23 amino acids in length and are comparatively shorter than the signal peptides found in eutherian mammals (29-30 aa) (Chang et al. 2006), and more similar in length to the signal peptides found in fish (22-26aa) (Chang et al. 2006). The signal peptide is followed by a conserved cathelin-like domain (the pro-region), which is highly conserved across all vertebrate species and contains four conserved cysteine residues, which are arranged to form two disulfide bonds. These residues are conserved in the opossum sequences. In the opossum, the cathelin domains range from 96-120 in length, similar to cathelin domains in eutherian mammals (94-114 aa) (Chang et al. 2006) and shorter than cathelin domains in fish (115-127 aa) (Chang et al. 2006). All ten full length cathelicidin genes encode mature peptides that are cationic, suggesting that the peptides will interact with negatively charged microbial membranes resulting in membrane disruption.

The size of opossum cathelicidin mature peptides ranges from 29-58 residues, as compared to a range from 12-80 in eutherian mammals.

Supplementary Figure 1. Amino acid sequence alignment of opossum cathelicidin genes. The signal peptide is underlined. The putative propeptide cleavage site is shown. The conserved cysteines are boxed.



```
                   10        20        30        40        50        60        70
           ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
cath1      MASFRILLPLLLLG-LIEIKAAEISGLSYQDALKVAMSHFNKNSEEKNSYWITATEAQT-KWDSESKEPQ
cath2      MASFRILLPLLLLG-LIEIKAAEISGLSYQDALKVAMSHFNRKSEEKNSYWITATEAQT-KWDSESKEPQ
cath3      MEHLRKVLLMASVATILPAQVLSQSSLSYEKTLSVAIHFYNQVHGIENVFKVLEVHPPPSNQKPQDQRLK
cath4      MQVP--LLVLGLLSLLTP--LASAQDLNYQNAVTEFIREYNSNIRSGNLFRLSVLTLPP-GEGNNPTLS-
cath5      MERGQIMWLLLLLMLVTP--LASTQTLTYQDLVNRFIINYNKKSNSRNLFRLLALNLQP-GANNDPAIPR
cath6      MASFRILLPLLLLG-LMEVKATEVRGLSYYDALTIAMSHFNSKCDEKNAYWITATRPQL-KWDRKSKELH
cath7      MASFRILLPLLLLG-LMEVKATEVRGLSYYDALTIAMSHFNSKCDEKNAYWITAITPQL-KWDRKSKEPH
cath8      MENTWRVLLLLGLATVVS--ATPRRVFTYKEAAVLASRNFNRGSNEGRKYRVVEASLST----PDSPLEL
cath9      MEHFRSILLLVSVATLIPTQAVPQSSLSYSKALSTAIHFYNQIHSRENAFWLLKTHPPSPSQDPEEQVPK
cath10     --------------------------------------------------------------KPQDQRLK
cath11     MEHLRKVLLLVSVVTLIPAQVLPQSSPRDGNTLSAAIHFYNQVHGTKNAFKVLETHPPPSNQGALEQRLK
cath12     MASFRFLLPLLLLG-LIEIKAAEISGLSYQDALKVAMSHFHRKSEVKNSYWITATEAQT-KWDRKSKEPH

                   80        90       100       110       120       130       140
           ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
cath1      HLGFTVQETECLKTETNPLIECTIKTD---------------GKVKHCQATVQFISHSETDVQVHCSPVP
cath2      HLGFTVQETECLKTETNPLIECTIKTD---------------GKVKHCQATVQFISHSETDVQVHCSPVP
cath3      LLSFAMKEMVCPRTGELPLDQCDFKRD---------------GLVKKCQGSISNEQD-IPSIILTCDPVA
cath4      PVSFTITETVCQKEDGRDPEECDFKDN---------------GMVKECFGVIRLEFT-KPIVDVSCDGAS
cath5      PLNFTIMETVCPNTKQRQLVECDFKKH---------------GLVKVCFGIISLDAT-QPSIDISCEEPG
cath6      RMGFTIQETECLKGETRTPSKCAFKKD---------------GEEKYCRAAVKFICDGEADVWVRCSPMA
cath7      RMGFAIQETECLKGETRTPSKCAFKKD---------------GEEKYCRAAVKFICDGEADVWVRCSPMA
cath8      PLTFRIKETECPSTERQDPETCAFREN---------------GAEKQCTANFTKLSR-FGLRSLNCVDIV
cath9      LLSFTLKETICPVTEELLLDQCDFKADGVNSQWEDGLDETLKKLVKECQGSVSNEQD-TAAIILTCEPVA
cath10     LLSFAMKEM----------------------------------LVKKCQASVSSEQD-IVAIILTCDSGA
cath11     FLSFTMKETMCPRNEEFPLQQCDFKRD---------------GLVKKCQASVSNEQD-IVAIILTCDLGA
cath12     HLGFTVQETECLKTETHTLSKCTIKKD---------------GKVKHCQATVQFISHSETDVQVHCSPVP

                  150       160       170       180       190       200
           ....|....|....|....|....|....|....|....|....|....|....|....|....
cath1      SNQMVKRTKRGARRGLTKVLKKIFGSIVKKAVSK----GVAEGLKAVDEARKAEKKGKKH----
cath2      SNQMVKRTKRGIKKGISKVLKKFFSSMIKKAVSKGISKGISKGIQEVDKAIKQKKEGQKH----
cath3      PVTP----APPGFTTSSKKKGKKKLPYTVIFG----------RTDFP-----------------
cath4      KTKRR----SLLKRLG----DGIRGFWNGFRGRK-------------------------------
cath5      QIMNDGFWYQLIRTFGNLIHQKYRKLLEAYRKLRDIFSG-------------------------
cath6      SNTMVRRSKRG--IKVPSFVKKVLKDVVSESISH----GIPKAISAVNKLIKKKKKGATTPRVK
cath7      SNTIVRRSKRG--IKVPGFVKKFLKDVVSETISH----SIPKAISAVNKLIKKKKKGATTPRVK
cath8      NDNNQ-----------------------------------------------------------
cath9      PVPS-----RSRRTPLPKQKNGSKNRRFRIGGYTMISMKQPRVQKAPYMEAL------------
cath10     PLVQGRIEEVTHGDPMLATLTQAHCPYFLRKKPEGEMATGGRPQETPKLLGAL-----------
cath11     LLLP----RSR---RALYRRRKYPTSYYLLLGGKG-----NNRGYKPHIIYV------------
cath12     SNQMVKRTKR----EISKILEEIFSTVIKIFIPK----GFYKGIQLVNEIIKENNERQKQKP--
```

Defensins

- **β defensins**

β defensins are cationic, amphipathic molecules, which range from 38-42 amino acids in length. They are expressed in epithelial tissues of vertebrates and invertebrates as well granulocytes, macrophages and thrombocytes. β defensin peptides have broad spectrum antimicrobial activity and resemble chemokines both structurally and functionally. β defensin genes encode a precursor protein which contains a signal sequence, a pro-sequence and a mature peptide that contains six highly conserved cysteine residues (reviewed in Ganz 2003; Selsted 2004).

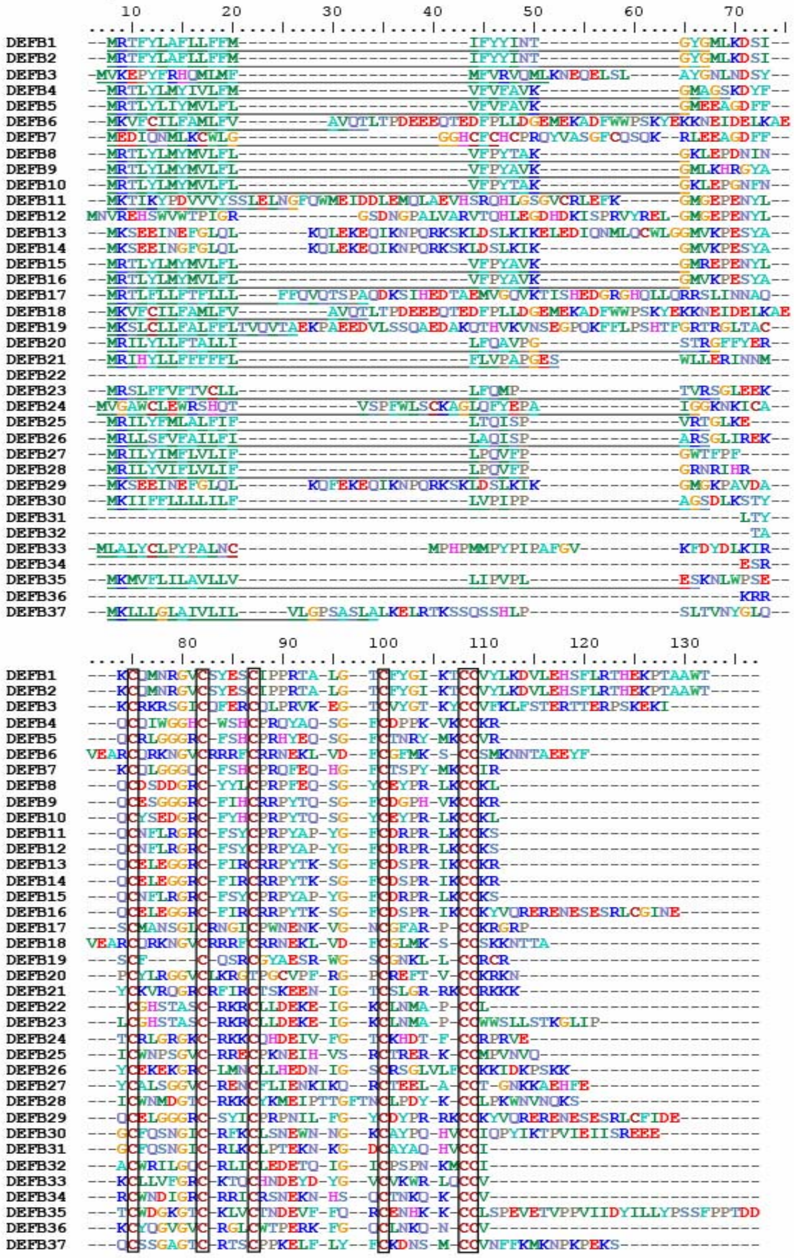
The number of β defensin genes varies between different mammalian genomes, with 39 in humans, 52 in the mouse, 42 in the rat, 43 in the dog and 13 in the chicken (reviewed in Patil et al. 2005).

To identify opossum β defensin genes we searched the genome with a HMMer profile of the mature peptide. Gene prediction resulted in the identification of 32 putative β defensin genes and four putative pseudogenes for which we were unable to predict a first exon. The availability of a single opossum defensin cDNA sequence on GenBank (accession EC091475) provided support for our gene prediction methodology, as this sequence was identical, bar a single amino acid difference, to β defensin *MdoDEFB18.*

In vertebrates, β defensin genes contain either two, three or four exons (Zou et al. 2007). Bird and fish β defensins consist of four exons and three exons respectively, although both have only three coding exons. In fish, the conserved cysteine residues are spread over two exons, with four cysteines found in exon 2 and two in exon 3, while in birds all six cysteines are found in exon 3. In eutherian mammals, β defensins are encoded by two exons, the second exon encoding the entire mature peptide. It has been suggested that the fusion of defensin exons in mammals was an adaptive event, allowing faster mobilization of innate host immunity (Xiao et al. 2004). This fusion event most likely occurred prior to the divergence of marsupials and eutherians, as no three exon gene predictions were observed prior to manual curation. However, our approach has low sensitivity for detecting defensins containing three exons and it is possible that these genes have been overlooked.

An alignment of the opossum β defensins is shown in Supplementary Figure 2, and their features are summarized in Supplementary Table 2. The percentage amino acid identity of opossum mature peptides ranged from 7-92% and the precursors from 8-94%. The majority of predicted precursor proteins contained a signal sequence, with only defensins *MdoDEFB12, 13, 14* and *29* having no predicted signal sequence. Absence of signal peptides could be the result of inaccurate gene prediction or SIGNALP false negatives. Alternatively these genes may be pseudogenes. The mature peptides range from 37-83 amino acids in length. The majority of the predicted peptides were cationic, with only four having a negative charge. The six conserved cysteine residues are highlighted in the figure.

Supplementary Figure 2. Amino acid sequence alignment of opossum β defensins. The signal peptide is underlined. The six conserved cysteines are boxed.

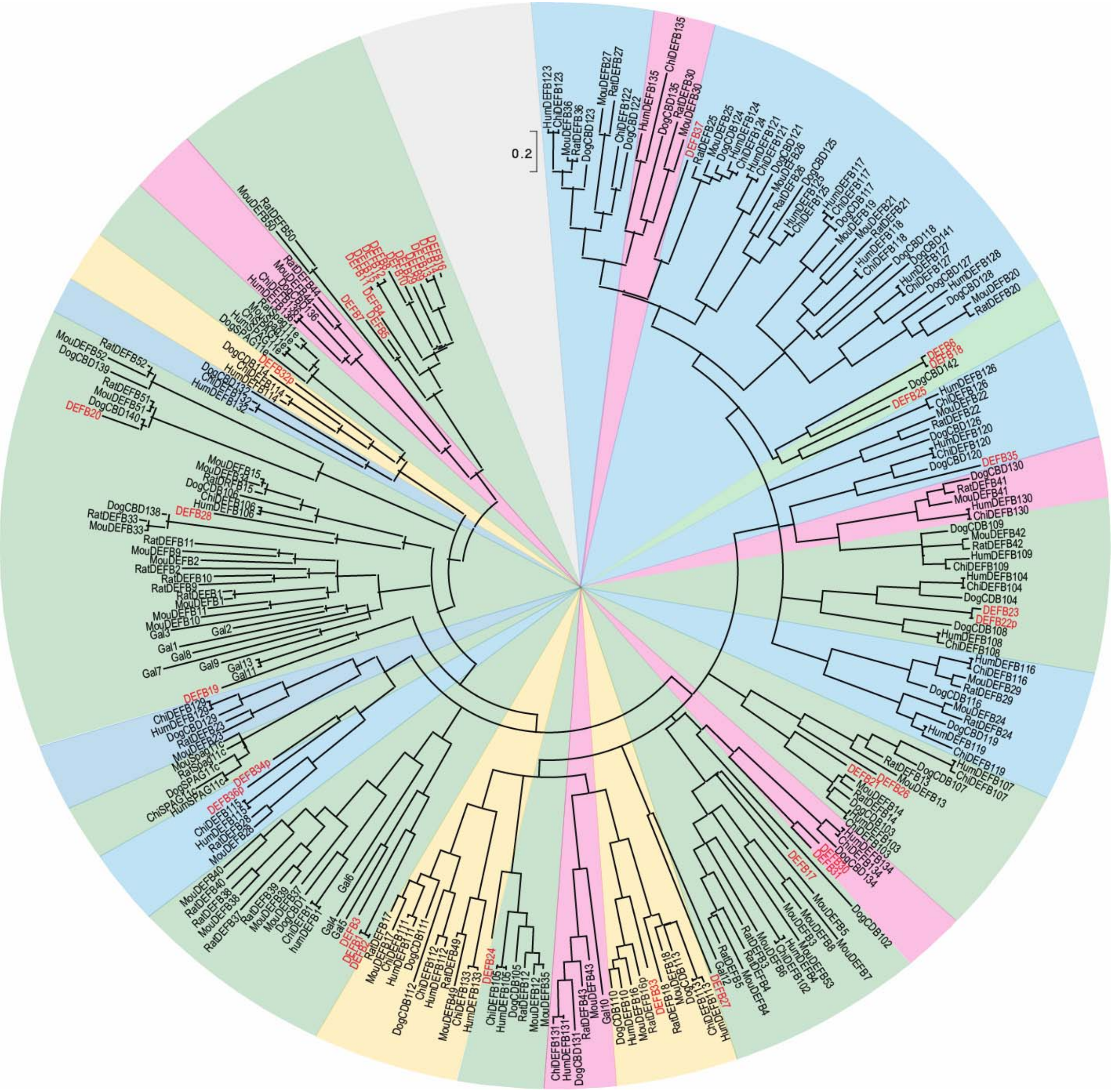


```
             10        20        30        40        50        60        70
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
DEFB1     --MRTFYLAFLLFFM-----------------------IFYYINT--------------GYGMLKDSI--
DEFB2     --MRTFYLAFLLFFM-----------------------IFYYINT--------------GYGMLKDSI--
DEFB3     -MVKEPYFRHQMLMF-----------------------MFVRVQMLKNEQELSL-----AYGNLNDSY--
DEFB4     --MRTLYLMYIVLFM-----------------------VFVFAVK--------------GMAGSKDYF--
DEFB5     --MRTLYLIYMVLFL-----------------------VFVFAVK--------------GMEEAGDFF--
DEFB6     --MKVFCILFAMLFV---------AVQTLTPDEEEQTEDFPLLDGEMEKADFWWPSKYEKKNEIDELKAE
DEFB7     --MEDIQNMLKCWLG--------------------GGHCFCHCPRQYVASGFCQSQK--RLEEAGDFF--
DEFB8     --MRTLYLMYMVLFL-----------------------VFPYTAK--------------GKLEPDNIN--
DEFB9     --MRTLYLMYMVLFL-----------------------VFPYAVK--------------GMLKHRGYA--
DEFB10    --MRTLYLMYMVLFL-----------------------VFPYTAK--------------GKLEPGNFN--
DEFB11    --MKTIKYPDVVVYSSLELNGFQWMEIDDLEMQLAEVHSRQHLGSGVCRLEFK------GMGEPENYL--
DEFB12    MNVREHSWVWTPIGR------------GSDNGPALVARVTQHLEGDHDKISPRVYREL-GMGEPENYL--
DEFB13    --MKSEEINEFGLQL-------KQLEKEQIKNPQRKSKLDSLKIKELEDIQNMLQCWLGGMVKPESYA--
DEFB14    --MKSEEINGFGLQL-------KQLEKEQIKNPQRKSKLDSLKIK--------------GMVKPESYA--
DEFB15    --MRTLYLMYMVLFL-----------------------VFPYAVK--------------GMREPENYL--
DEFB16    --MRTLYLMYMVLFL-----------------------VFPYAVK--------------GMVKPESYA--
DEFB17    --MRTLFLLFTFLLL----FFQVQTSPAQDKSIHEDTAEMVGQVKTISHEDGRGHQLLQRRSLINNAQ--
DEFB18    --MKVFCILFAMLFV---------AVQTLTPDEEEQTEDFPLLDGEMEKADFWWPSKYEKKNEIDELKAE
DEFB19    --MKSLCLLFALFFLTVQVTAEKPAEEDVLSSQAEDAKQTHVKVNSEGPQKFFLPSHTFGRTRGLTAC--
DEFB20    --MRILYLLFTALLI-----------------------LFQAVPG--------------STRGFFYER--
DEFB21    --MRIHYLLFFFFFL-----------------------FLVPAPGES------------WLLERINNM--
DEFB22    ----------------------------------------------------------------------
DEFB23    --MRSLFFVFTVCLL-----------------------LFQMP----------------TVRSGLEEK--
DEFB24    -MVGAWCLEWRSHQT------------VSPFWLSCKAGLQFYEPA--------------IGGKNKICA--
DEFB25    --MRILYFMLALFIF-----------------------LTQISP---------------VRTGLKE----
DEFB26    --MRLLSFVFAILFI-----------------------LAQISP---------------ARSGLIREK--
DEFB27    --MRILYIMFLVLIF-----------------------LPQVFP---------------GWTFPF-----
DEFB28    --MRILYVIFLVLIF-----------------------LPQVFP---------------GRNRIHR----
DEFB29    --MKSEEINEFGLQL-------KQFEKEQIKNPQRKSKLDSLKIK--------------GMGKPAVDA--
DEFB30    --MKIIFFLLLLILF-----------------------LVPIPP---------------AGSDLKSTY--
DEFB31    -----------------------------------------------------------------LTY--
DEFB32    ------------------------------------------------------------------TA--
DEFB33    -MLALYCLPYPALNC-------------------MPHPMMPYPIPAFGV----------KFDYDLKIR--
DEFB34    -----------------------------------------------------------------ESR--
DEFB35    --MKMVFLILAVLLV-----------------------LIPVPL---------------ESKNLWPSE--
DEFB36    -----------------------------------------------------------------KRR--
DEFB37    --MKLLLGLAIVLIL-----VLGPSASLALKELRTKSSQSSHLP---------------SLTVNYGLQ--

                  80        90       100       110       120       130
          ....|....|....|....|....|....|....|....|....|....|....|....|....|..
DEFB1     ---KCQMNRGVCSYESCIPPRTA-LG--TCFYGI-KTCCVYLKDVLEHSFLRTHEKPTAAWT-----
DEFB2     ---KCQMNRGVCSYESCIPPRTA-LG--TCFYGI-KTCCVYLKDVLEHSFLRTHEKPTAAWT-----
DEFB3     ---KCRKRSGICQFERCQLPRVK-EG--TCVYGT-KYCCVFKLFSTERTTERPSKEKI---------
DEFB4     ---QCQIWGGHC-WSHCPRQYAQ-SG--FCDPPK-VKCCKR--------------------------
DEFB5     ---QCRLGGGRC-FSHCPRHYEQ-SG--FCTNRY-MKCCVR--------------------------
DEFB6     VEARCQRKNGVCRRRFCRRNEKL-VD--FCGFMK-S-CCSMKNNTAEEYF-----------------
DEFB7     ---KCQLGGGQC-FSHCPRQFEQ-HG--FCTSPY-MKCCIR--------------------------
DEFB8     ---QCDSDDGRC-YYLCPRPFEQ-SG--YCEYPR-LKCCKL--------------------------
DEFB9     ---QCESGGGRC-FIHCRRPYTQ-SG--FCDGPH-VKCCKR--------------------------
DEFB10    ---QCYSEDGRC-FYHCPRPYTQ-SG--YCEYPR-LKCCKL--------------------------
DEFB11    ---QCNFLRGRC-FSYCPRPYAP-YG--FCDRPR-LKCCKS--------------------------
DEFB12    ---QCNFLRGRC-FSYCPRPYAP-YG--FCDRPR-LKCCKS--------------------------
DEFB13    ---QCELEGGRC-FIRCRRPYTK-SG--FCDSPR-IKCCKR--------------------------
DEFB14    ---QCELEGGRC-FIRCRRPYTK-SG--FCDSPR-IKCCKR--------------------------
DEFB15    ---QCNFLRGRC-FSYCPRPYAP-YG--FCDRPR-LKCCKS--------------------------
DEFB16    ---QCELEGGRC-FIRCRRPYTK-SG--FCDSPR-IKCCKYVQRERENESESRLCGINE--------
DEFB17    ---SCMANSGICRNGICPWNENK-VG--NCGFAR-P-CCKRGRP-----------------------
DEFB18    VEARCQRKNGVCRRRFCRRNEKL-VD--FCGLMK-S-CCSKKNTTA---------------------
DEFB19    ---SCF-----C-QSRCGYAESR-WG--SCGNKL-L-CCRCR-------------------------
DEFB20    ---PCYLRGGVCLKRGTPGCVPF-RG--PCREFT-V-CCKRKN------------------------
DEFB21    ---YCKVRQGRCRFIRCTSKEEN-IG--TCSLGR-RKCCRKKK------------------------
DEFB22    ----CGHSTASC-RKRCLLDEKE-IG--KCLNMA-P-CCL---------------------------
DEFB23    ---LCGHSTASC-RKRCLLDEKE-IG--KCLNMA-P-CCWWSLLSTKGLIP----------------
DEFB24    ---TCRLGRGKC-RKKCQHDEIV-FG--TCKHDT-F-CCRPRVE-----------------------
DEFB25    ---ICWNPSGVC-RRECPKNEIH-VS--RCTRER-K-CCMPVNVQ----------------------
DEFB26    ---YCEKEKGRC-LMNCLLHEDN-IG--SCRSGLVLFCCKKIDKPSKK-------------------
DEFB27    ---YCALSGGVC-RENCFLIENKIKQ--RCTEEL-A-CCT-GNKKAEHFE-----------------
DEFB28    ---ICWNMDGTC-RKKCYKMEIPTTGFTNCLPDY-K-CCLPKWNVNQKS------------------
DEFB29    ---QCELGGGRC-SYICPRPNIL-FG--YCDYPR-RKCCKYVQRERENESESRLCFIDE--------
DEFB30    ---GCFQSNGIC-RFKCLSNEWN-NG--KCAYPQ-HVCCIQPYIKTPVIEIISREEE----------
DEFB31    ---GCFQSNGIC-RLKCLPTEKN-KG--DCAYAQ-HVCCI---------------------------
DEFB32    ---ACWRILGQC-RLICLEDETQ-IG--ICPSPN-KMCCI---------------------------
DEFB33    ---KCLLVFGRC-KTQCHNDEYD-YG--VCVKWR-LQCCV---------------------------
DEFB34    ---RCWNDIGRC-RRICRSNEKN-HS--QCTNKQ-K-CCV---------------------------
DEFB35    ---TCWDGKGTC-KLVCTNDEVF-FQ--RCENHK-K-CCLSPEVETVPPVIIDYILLYPSSFPPTDD
DEFB36    ---KCYQGVGVC-RGLCWTPERK-FG--QCLNKQ-N-CCV---------------------------
DEFB37    ---QCSSGAGTC-RTSCPPKELF-LY--FCKDNS-M-CCVNFFKMKNPKPEKS--------------
```

The α and θ defensins are believed to have evolved from an ancestral β defensin after the separation of the bird and mammal lineages (Xiao et al. 2004). Opossum sequences were aligned with the data set used by Patil et al. (Patil et al. 2005), with resulting phylogenetic trees supporting the topology obtained by Patil and colleagues and are shown in Supplementary figure 3.

Supplementary Figure 3. Phylogenetic tree representing evolutionary history of β defensins. Shading corresponds to genomic location as shown in Figure 3; Cluster A is green, Cluster B is pink, Cluster C is orange and Cluster D is blue. Opossum sequences are shown in red.



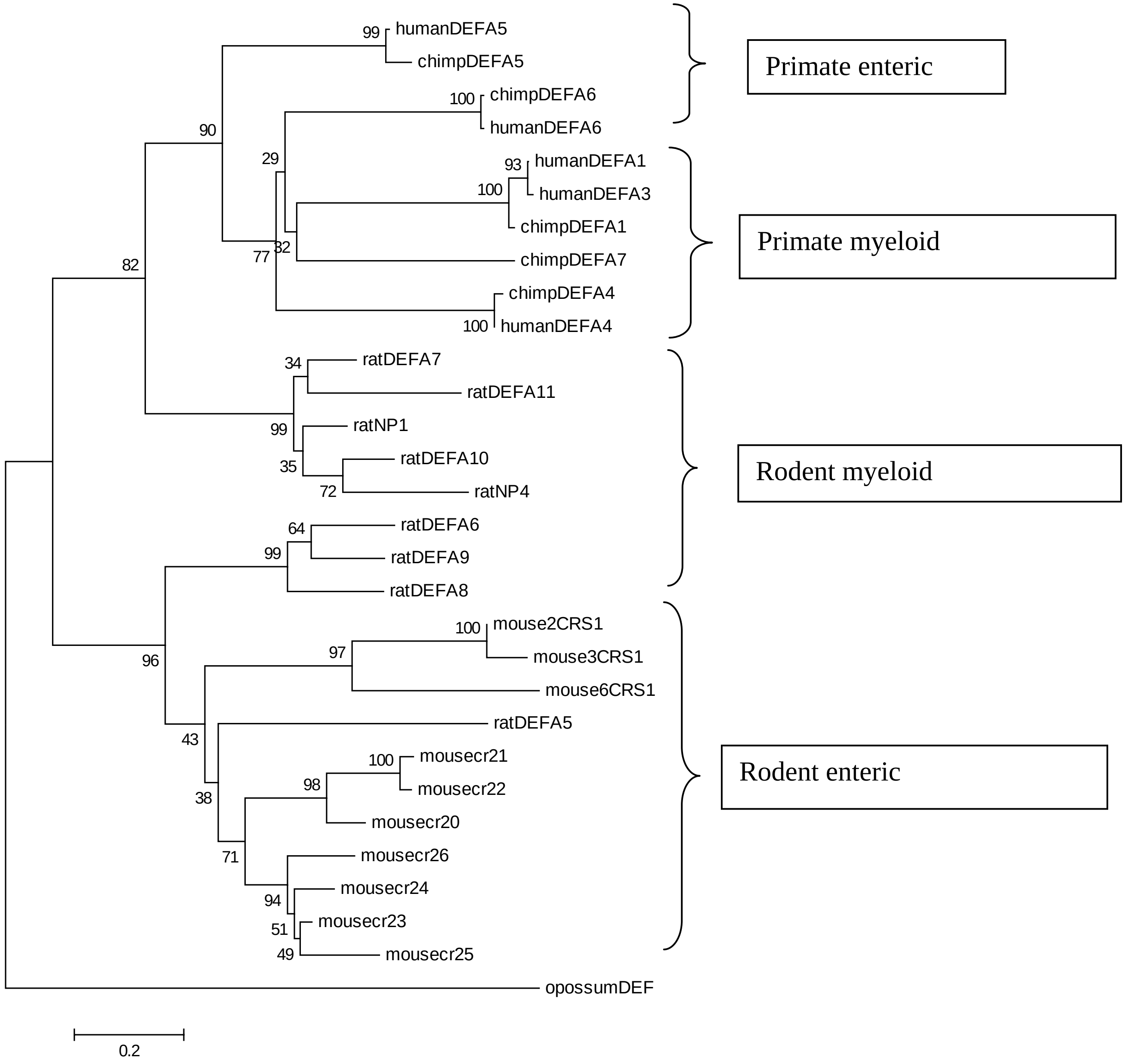
Nomenclature for defensin genes is confusing, as different names have been given to rodent and primate gene families. Here we numbered defensin genes based on their order in clusters. Figure 3 shows that opossum Cluster A is located on Chromosome 1 and contains 29 genes. Eutherian Clusters B and C evolved from a single cluster in early mammals, which remains in the opossum on chromosome 2 and contains two putative genes, and two putative pseudogenes. Cluster D, found on opossum chromosome 1, also contains two putative genes and two putative pseudogenes. A single defensin was found on the unordered chromosome but is likely to be part of Cluster A, based on its phylogenetic position.

- **α defensins**

The human genome contains a cluster of 10 α defensin genes and pseudogenes that span a region of 132kb on chromosome 8p23 (Patil et al. 2004). In eutherian mammals, the length of α defensin precursors range from 80 to105 amino acids and contain a signal sequence, prosegment and cationic mature peptide (Patil et al. 2004). The opossum α defensin is 91 amino acids in length and contains a 19 amino acid signal peptide. The genomic sequence has a single amino acid difference from an opossum α cDNA sequence available on Genbank (accession EC091476).

There are two α defensin types in mammals: myeloid, which are expressed in the bone marrow, and enteric, which are primarily expressed in Paneth cells (Patil et al. 2004). Phylogenetic analysis places the sole opossum α defensin at the base of both lineages, albeit without bootstrap support (Supplementary Figure 4). The tree topology indicates that the sole opossum α defensin may be ancestral to both the myeloid and enteric defensin lineages. Characterization of the function of the marsupial gene will help test this hypothesis.

Supplementary Figure 4. Phylogenetic tree showing relationship of opossum α defensins to eutherian α defensins genes.



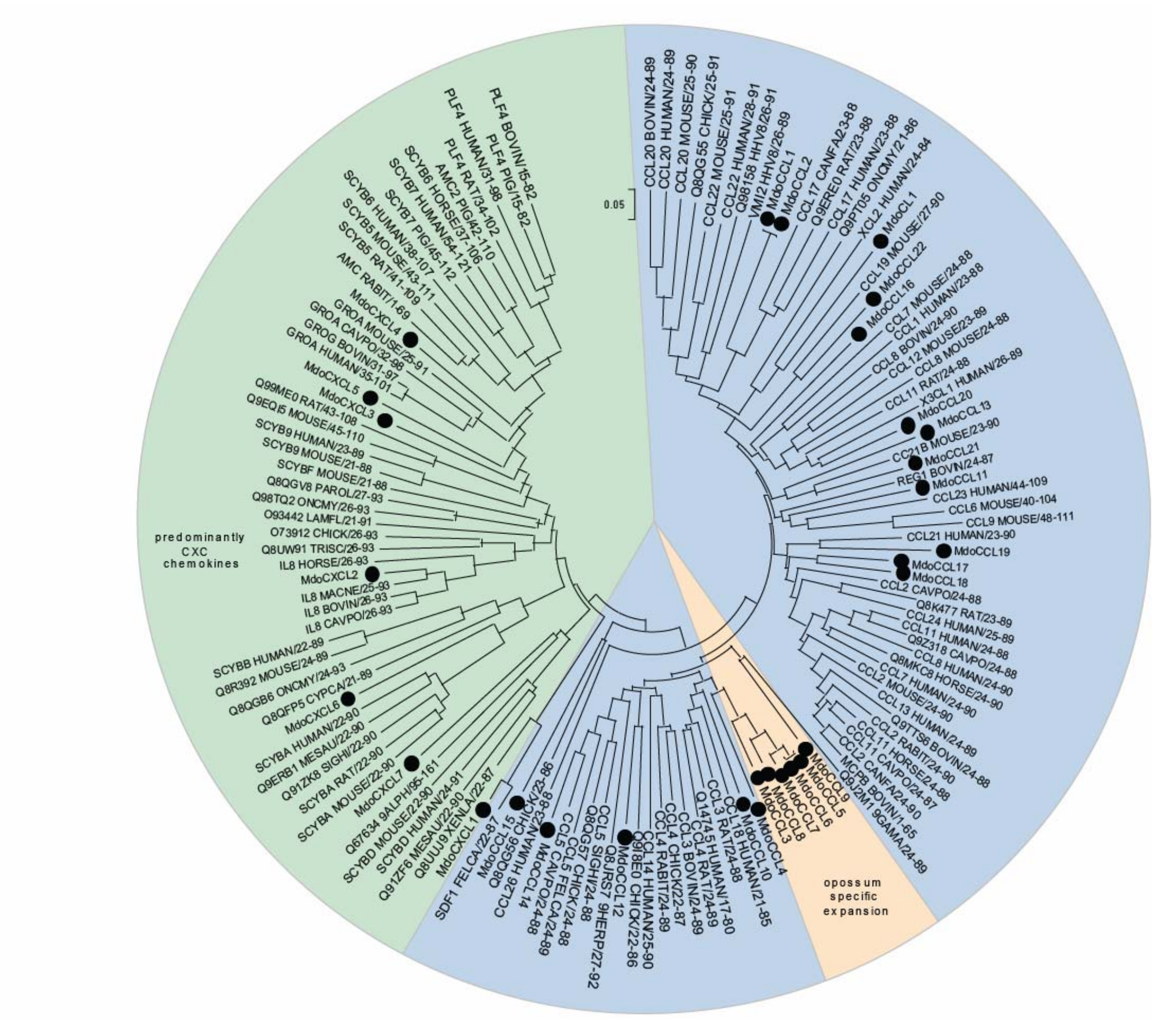
- **θ defensins**

θ defensins are cyclic peptides, derived from a mutated α defensin gene (Tang et al. 1999). They are expressed in several species of old world monkey and orangutan. A mutation in the human θ defensin gene has been implicated as a contributory cause for HIV and AIDs in humans and absence in old world monkeys (Selsted 2004). θ defensin genes were not identified using BLAST searches of the opossum genome.

The chemokines

Chemokines (CHEMotactic cytoKINES) are the largest family of cytokines in humans. They are small proteins that have a role in cell activation, migration, and differentiation (Laing and Secombes 2004). They are involved in many biological processes such as lymphocyte polarisation, angiogensesis, hematopoiesis, apoptosis and tumor metastasis (Esche et al. 2005). Many chemokines may also have defensin-like antimicrobial actions. In fact, chemokines are very similar to β defensins in that they are found in clusters and contain conserved cysteine motifs. Opossum chemokines were identified using a PFAM profile spanning the amino acids of exons 2 and 3. Assignment of genes to families was done phylogenetically (Supplementary Figure 5).

Supplementary Figure 5. Phylogenetic tree of mammalian chemokines. The opossum CCL expansion is shown in yellow. CXC chemokines are shown in green.



The human genome encodes 47 chemokines, while 24 have been identified in the genome of the chicken (Kaiser et al. 2005). We have identified 31 chemokine genes in the opossum genome. All chemokines contained the conserved cysteine motifs. The chemokine family is traditionally divided into four subgroups based on the arrangement of one or two N-terminal cysteine residues: CXC, CC, C and CX_3C.

In humans, 16 CXC-type chemokines have been identified. Seven CXC genes were identified in the opossum genome. Some CXC chemokines contain a three residue motif (ELR) immediately adjacent to the cysteines. This motif plays a role in the neutrophil attraction (Laing and Secombes 2004). Opossum chemokines containing an ELR motif (*MdoCXCL 2, 3* and *4*) are grouped with other ELR positive CXC chemokines from mammals and chickens on the phylogenetic tree. *MdoCXCL1* appears orthologous to eutherian CXCL12, with greater than 93% bootstrap support. MdoCXCL6 and *MdoCXCL7* are possible orthologues of *CXCL10* and *CXCL13* in eutherians.

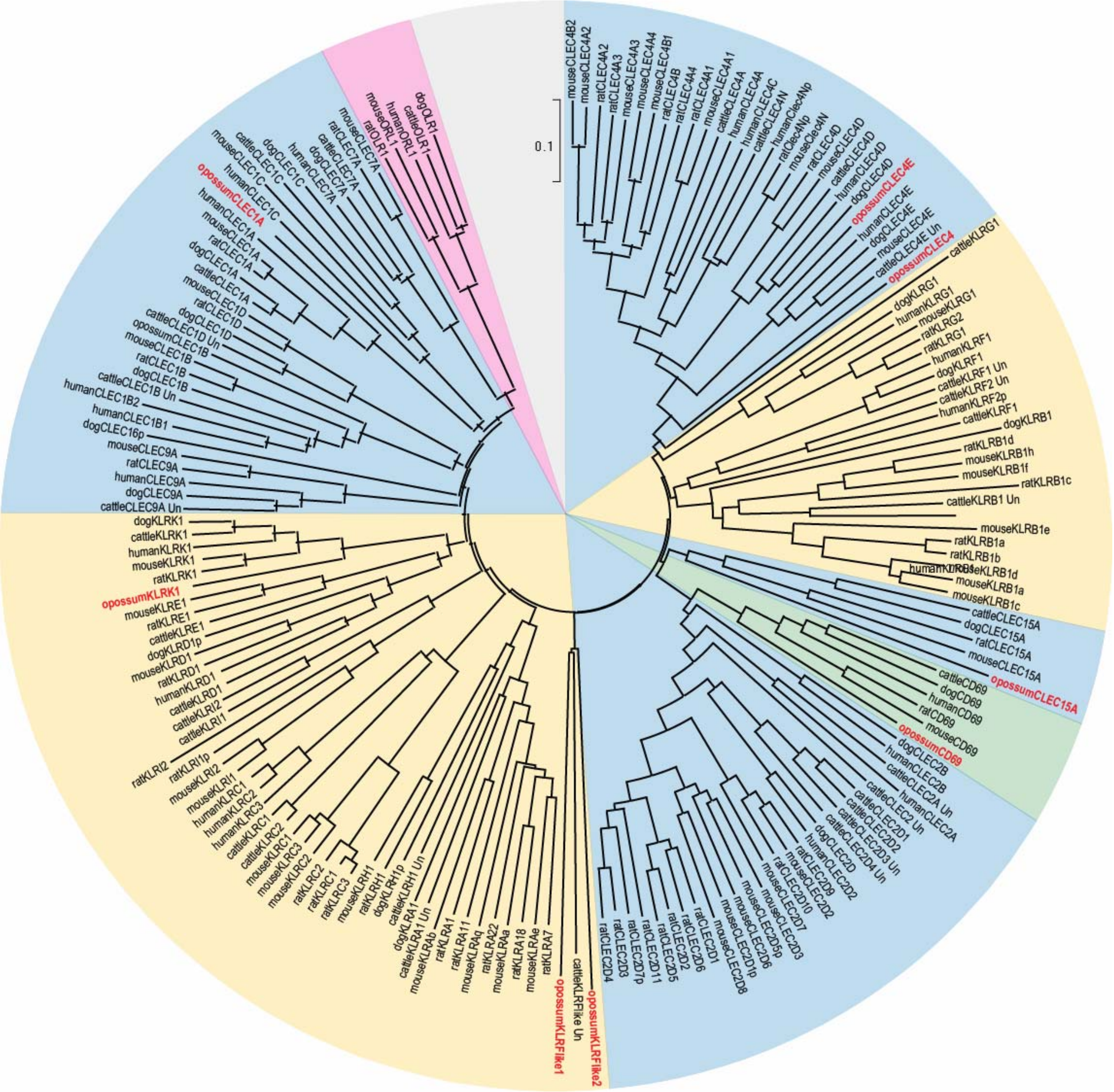
In opossum, as with eutherians and birds, the largest subgroup of chemokines is the CC family. There are 22 members of this subfamily in opossum. Humans and chickens have 28 and 14 members, respectively. A large cluster of CC-type chemokines is present on opossum chromosome 2. This cluster contains 16 putative genes distributed in just over 2 Mb. Phylogenetic analysis suggests that *MdoCCL 3, 5, 6, 7* and *8*, are likely to have resulted from either a marsupial- or an opossum-specific lineage expansion.

Two highly similar C chemokines have been identified in the human genome (Laing and Secombes 2004). A homologous protein is also present in chickens (Rossi et al. 1999). We identified a member of this group in the opossum genome, *MdoXCL1*. Only one member of the CX_3C family is found in eutherians and birds (Bazan et al. 1997, Kaiser, 2005 #168). A CX_3C homologue was identified in the opossum genome on chromosome 1, adjacent to *MdoCCL1* and *MdoCCL2*

The identification of 31 chemokine family members in the opossum suggests that chemokine numbers in marsupials are midway between those of chickens and humans. Chemokine diversification may have occurred in two waves; after the divergence of birds and mammals and after the divergence of marsupials and eutherians. However, it is more likely that chemokine diversification has been driven by pathogen pressures which have resulted in lineage specific chemokine expansions. Detailed characterization of gene expression profiles of opossum specific CC chemokines may lead to a greater understanding of immune function in marsupials.

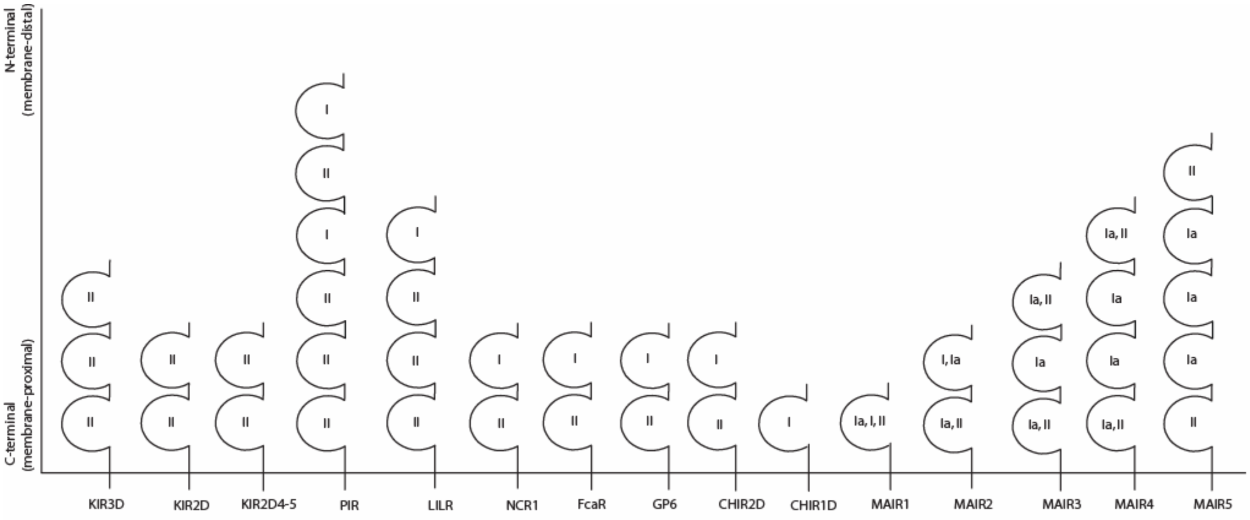
The Natural Killer Complex (NKC)

Supplementary Figure 6. Phylogenetic tree of the natural killer complex gene families. *CLEC* genes are highlighted in blue, *KLR* genes in yellow, *CD69* in green and *OLR1* genes in pink.

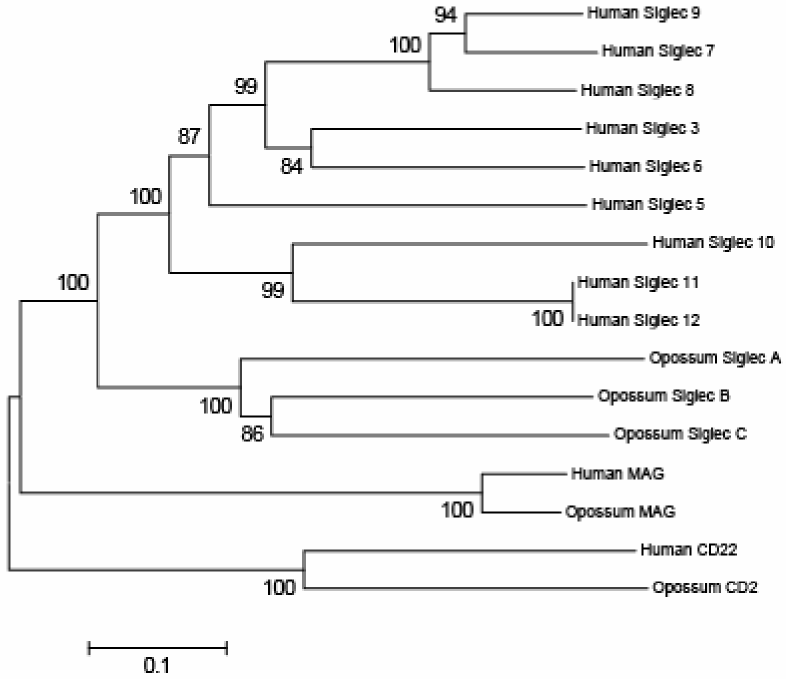


The Leukocyte Receptor Complex (LRC)

Supplementary Figure 7. A schematic diagram of the immunoglobulin domains of eutherian, marsupial and chicken LRC genes. Phylogenetic type of each domain is based on phylogenetic tree shown in Supplementary Figure 9.



Supplementary Figure 8. Phylogenetic tree showing relationship of opossum *SIGLEC* genes with those of humans. *MAG* and *CD22* orthologs were identified. Three opossum *SIGLECs* do not have eutherian orthologs and are basal to the eutherian CD33-related genes.

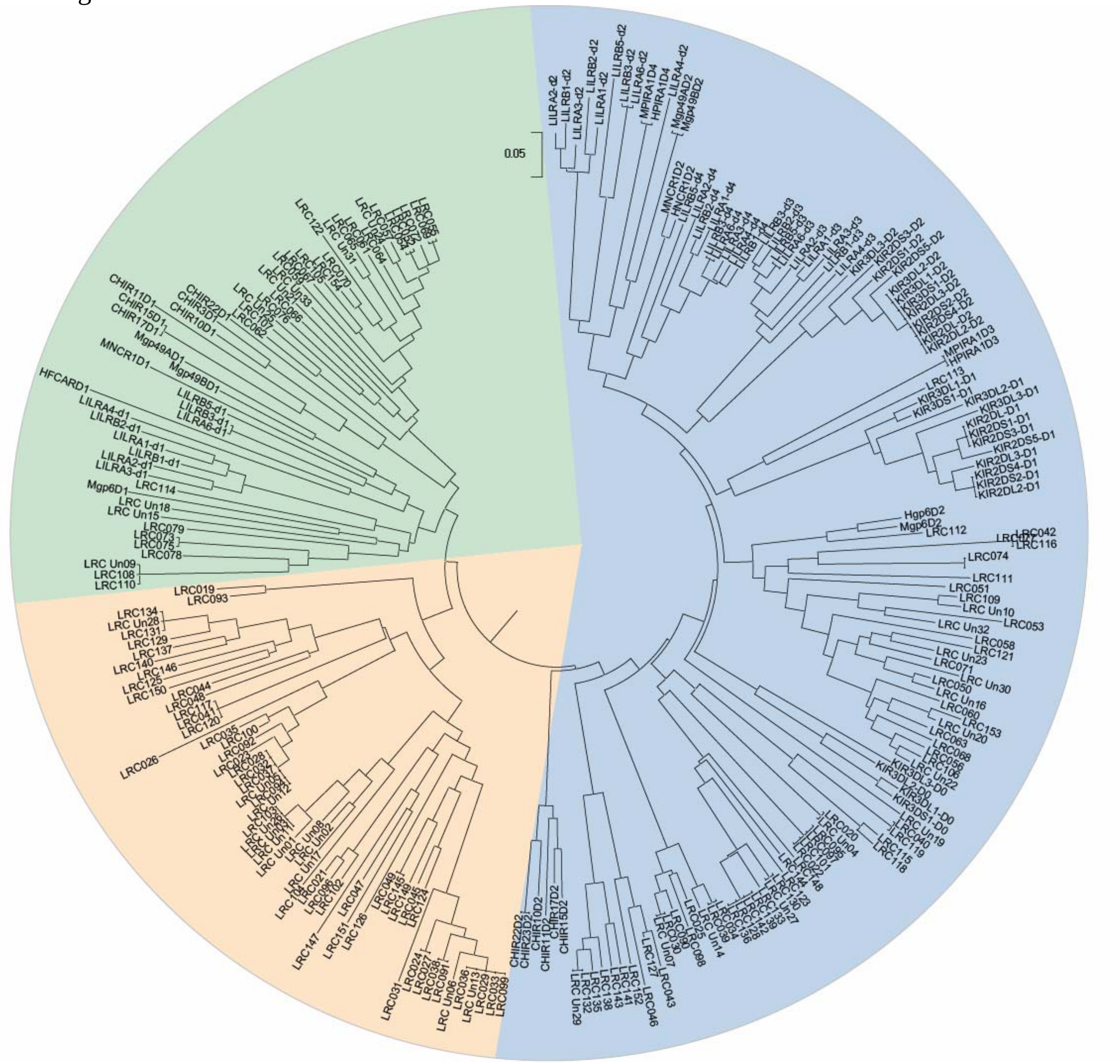


SIGLEC molecules, containing V set and C2-set Ig domains, recognize and bind sialylated glycans. There are eleven known SIGLECs in humans and eight in mice (Angata et al. 2004). Chaining of Ig domains and gene prediction in the opossum SIGLEC cluster resulted in the identification of *MAG (SIGLEC4)* and CD22 (*SIGLEC2*). The discovery of MAG is not surprising as it is a highly conserved molecule which has already been identified in birds, amphibians and fish. *CD22* has previously not been identified outside eutherian mammals. As in humans, *MAG* and *CD22* are located adjacent to each other in the opossum genome.

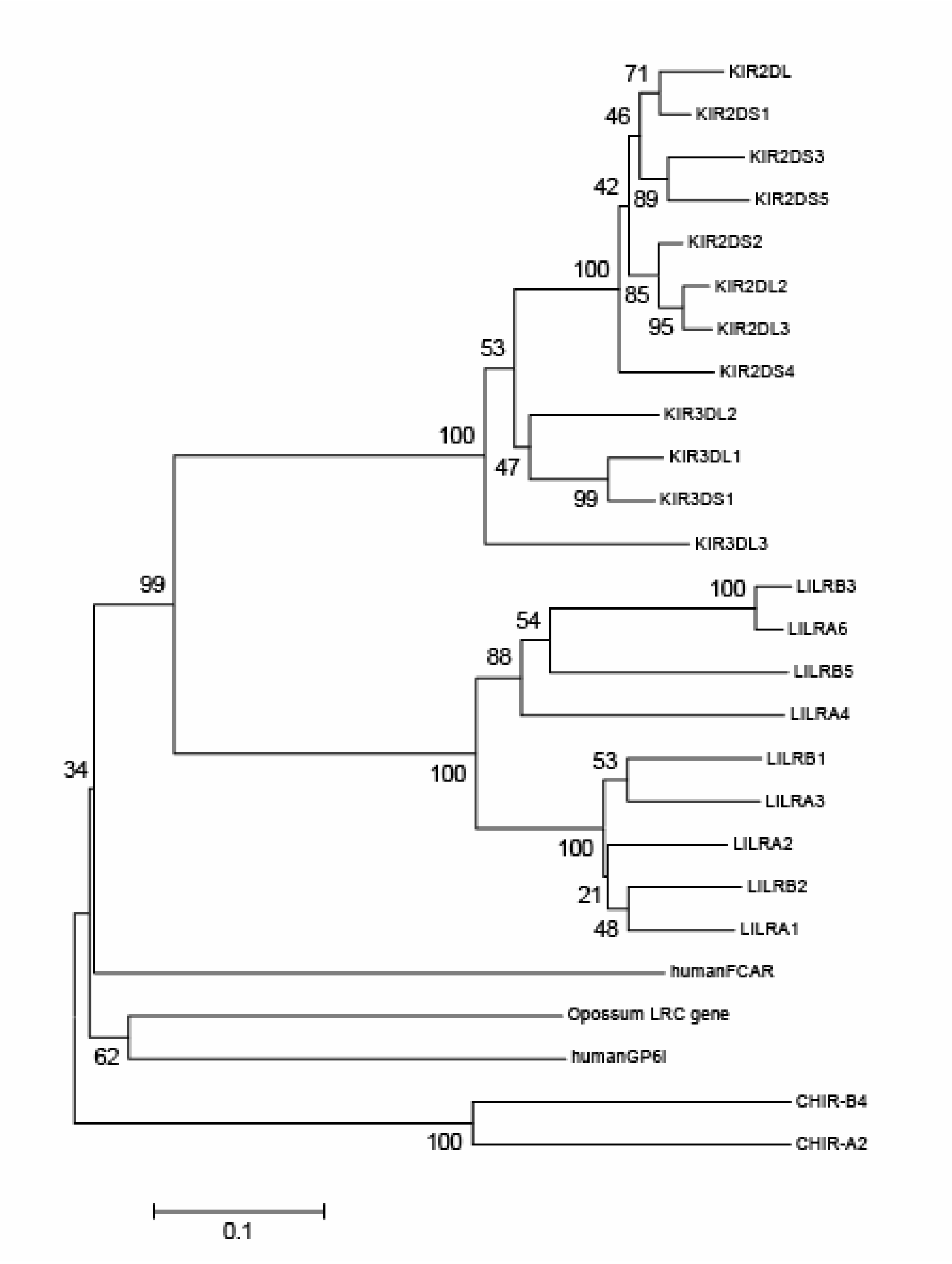
Aside from *MAG* and *CD22* the opossum genome also contains three other *SIGLEC* genes. These genes belong to the CD33-related group of highly related and rapidly evolving SIGLECs. Even within eutherian mammals, these genes differ significantly in composition and domain architecture due to their rapid evolution via multiple mechanisms including gene duplication, gene loss and exon shuffling (Angata et al. 2004). Phylogenetic comparison of *SIGLEC* genes (Supplementary Figure 8) shows clear orthology between opossum and human *MAG* and *CD22*. The other three opossum genes are basal to the human CD-33-related genes, indicating that these genes are not orthologous, and that human and opossum immune defence molecules have evolved to recognize different sialic acids on different pathogens.

Recent studies of catfish immunoglobulin-like receptors have identified receptors with a unique combination of Ig domains (Stafford et al. 2006). The distal domain is related to the Fc receptors, while proximal domains share homology with domains seen in the leukocyte receptor complex. This is interesting, because C2-set Ig domains, seen in the LRC, and Fc receptor Ig domains are believed to have shared a common ancestor.

Supplementary Figure 9. Phylogenetic tree showing the relationship between mammalian and chicken leukocyte receptor complex Ig domains. Chicken CI and eutherian MI genes are shown in green. Chicken CII and eutherian MII genes are shown in blue. Opossum Ig domains are located within both of these clades, and form an additional Ia clade, shown in orange.



Supplementary Figure 10. Phylogenetic tree showing that predicted opossum MAIR encompassing LRC domains 112, 113 and 114 is orthologous to *GP6*. We used mammalian Fc receptor Ig domains as an outgroup for two reasons. Firstly, Fc receptors and Ig-like receptors are believed to have evolved from a common ancestral element, and secondly, we wanted to test that opossum Ig-domains were not actually Fc type, as in catfish. No opossum LRC encoded Ig-domains clustered with these sequences and they are not shown on this figure.



Chaining of HMMer profiles

Both cathelicidins and immunoglobulin domain containing genes of the LRC were identified by chaining HMMer matches. To assess the reliability of this approach, we searched the human genome (build NCBI 36.1) with the cathelicidin (PF00666) and immunoglobulin (PF00047) Pfam profiles using HMMer. Since both HMMer profiles are likely to contain some of the human genes, this will only provide us with an estimate of the sensitivity and specificity for the chaining strategy, given our chosen E-value threshold and maximum gap size for each gene families.

As the conserved cathelin domain that is represented by the cathelicidin Pfam domain is encoded by exons 2 and 3, the fully local alignment model (fs) was used. An E-value threshold of 0.1 and maximum intron length of 2kb were chosen. The single cathelicidin gene encoded in the human genome was correctly identified with no false positives.

Chains of immunoglobulin domains with E-value<1 and maximum gap size of 10kb were compared with known LILR and KIR genes encoded in the human LRC. Genes were identified using ENSEMBL. Thirty-two genes and five pseudo-genes or poorly annotated features were identified using ENSEMBL. No false positives could be identified. The search showed 100% sensitivity and specificity at the gene level. Five out of the 73 immunoglobulin domains were not detected (93% sensitivity at the domain level), affecting the predicted structure prediction of 3 genes.

Methods

Phylogenetic analysis

Phylogenetic analysis for α and β defensins trees was conducted using MEGA 3.1 (Kumar et al. 2001). Amino acid alignments were generated using MUSCLE (Edgar 2004). Accession numbers for human, mouse, rat, dog and chicken defensins are provided in (Patil et al. 2004; Patil et al. 2005; Xiao et al. 2004). As per Patil (2005), the neighbour joining method was used to construct phylogenetic trees based on the proportion of amino acid differences (p distance) and the topology was tested by 1000 bootstrap replicates.

Phylogenetic analysis for chemokines was conducted on exon 2 and 3 chains using MEGA 3.1 (Kumar et al. 2001). Amino acid alignments were generated using MUSCLE (Edgar 2004). Sequences used for tree construction were obtained from Pfam.

Phyogenetic trees for NKC genes were constructed using MEGA3.1, as described above. Sequences used in alignments were obtained from (Hao et al. 2006; Nikolaidis et al. 2005).

Signal peptides and charge

Signal peptides in the defensins and cathelicins were predicted using SIGNALP (Bendtsen et al. 2004; Nielsen et al. 1997), while net charge was calculated using PROTPARAM (Gasteiger et al. 2005).

Physical mapping

To physically map opossum cathelicidin genes, which were assigned to the unordered chromosome, opossum BACs VMRC18_477I23, VMRC18_139C5, VMRC18_504P8 and VM18RC_156I9 were labelled by nick translation with digoxygenin-11-dUTP or biotin-16-dUTP (Roche Diagnostics, Basel, Switzerland), hybridized to male opossum metaphase chromosomes and fluorescent signals were detected following a previously described protocol (Alsop et al. 2005). Fluorescent signals were visualized on a Zeiss Axiolplan2 epifluorescence microscope (Carl Zeiss, Thornwood, NY, USA) and captured on a SPOT RT Monochrome CCD (charge-coupled device) camera (Diagnostic Instruments Inc., Sterling Heights, MI, USA). DAPI stained chromosome and fluorescent signal images were merged using IP Lab imaging software (Scanalytics Inc., Fairfax, VA, USA).

References

Alsop, A.E., P. Miethke, R. Rofe, E. Koina, N. Sankovic, J.E. Deakin, H. Haines, R.W. Rapkins, and J.A.M. Graves. 2005. Characterizing the chromosomes of the Australian model marsupial *Macropus eugenii* (tammar wallaby). *Chromosome Res.* 13: 627-636.

Angata, T., E.H. Margulies, E.D. Green, and A. Varki. 2004. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc Natl Acad Sci U S A* 101: 13251-13256.

Bals, R. and J.M. Wilson. 2003. Cathelicidins--a family of multifunctional antimicrobial peptides. *Cell Mol Life Sci* 60: 711-720.

Bazan, J.F., K.B. Bacon, G. Hardiman, W. Wang, K. Soo, D. Rossi, D.R. Greaves, A. Zlotnik, and T.J. Schall. 1997. A new class of membrane-bound chemokine with a CX3C motif. *Nature* 385: 640-644.

Bendtsen, J.D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783-795.

Chang, C.I., Y.A. Zhang, J. Zou, P. Nie, and C.J. Secombes. 2006. Two cathelicidin genes are present in both rainbow trout (Oncorhynchus mykiss) and atlantic salmon (Salmo salar). *Antimicrob Agents Chemother* 50: 185-195.

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

Esche, C., C. Stellato, and L.A. Beck. 2005. Chemokines: key players in innate and adaptive immunity. *J Invest Dermatol* 125: 615-628.

Ganz, T. 2003. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol* 3: 710-720.

Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook* (ed. J.M. Walker), pp. 571-607 Humana Press, Totowa, NJ.

Giacometti, A., O. Cirioni, R. Ghiselli, F. Mocchegiani, G. D'Amato, R. Circo, F. Orlando, B. Skerlavaj, C. Silvestri, V. Saba, M. Zanetti, and G. Scalise. 2004. Cathelicidin peptide sheep myeloid antimicrobial peptide-29 prevents endotoxininduced mortality in rat models of septic shock. *Am J Respir Crit Care Med* 169: 187-194.

Hao, L., J. Klein, and M. Nei. 2006. Heterogeneous but conserved natural killer receptor gene complexes in four major orders of mammals. *Proc Natl Acad Sci U S A* 103: 3192-3197.

Kaiser, P., T.Y. Poh, L. Rothwell, S. Avery, S. Balu, U.S. Pathania, S. Hughes, M. Goodchild, S. Morrell, M. Watson, N. Bumstead, J. Kaufman, and J.R. Young. 2005. A genomic analysis of chicken cytokines and chemokines. *J Interferon Cytokine Res* 25: 467-484.

Kumar, S., K. Tamura, I.B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17: 1244-1245.

Laing, K.J. and C.J. Secombes. 2004. Chemokines. *Dev Comp Immunol* 28: 443-460.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10: 1-6.

Nikolaidis, N., J. Klein, and M. Nei. 2005. Origin and evolution of the Ig-like domains present in mammalian leukocyte receptors: insights from chicken, frog, and fish homologues. *Immunogenetics* 57: 151-157.

Patil, A., A.L. Hughes, and G. Zhang. 2004. Rapid evolution and diversification of mammalian alpha-defensins as revealed by comparative analysis of rodent and primate genes. *Physiol Genomics* 20: 1-11.

Patil, A.A., Y. Cai, Y. Sang, F. Blecha, and G. Zhang. 2005. Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. *Physiol Genomics* 23: 5-17.

Rossi, D., J. Sanchez-Garcia, W.T. McCormack, J.F. Bazan, and A. Zlotnik. 1999. Identification of a chicken "C" chemokine related to lymphotactin. *J Leukoc Biol* 65: 87-93.

Selsted, M.E. 2004. Theta-defensins: cyclic antimicrobial peptides produced by binary ligation of truncated alpha-defensins. *Curr Protein Pept Sci* 5: 365-371.

Stafford, J.L., E. Bengten, L. Du Pasquier, R.D. McIntosh, S.M. Quiniou, L.W. Clem, N.W. Miller, and M. Wilson. 2006. A novel family of diversified immunoregulatory receptors in teleosts is homologous to both mammalian Fc receptors and molecules encoded within the leukocyte receptor complex. *Immunogenetics* 58: 758-773.

Tang, Y.Q., J. Yuan, G. Osapay, K. Osapay, D. Tran, C.J. Miller, A.J. Ouellette, and M.E. Selsted. 1999. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. *Science* 286: 498-502.

Uzzell, T., E.D. Stolzenberg, A.E. Shinnar, and M. Zasloff. 2003. Hagfish intestinal antimicrobial peptides are ancient cathelicidins. *Peptides* 24: 1655-1667.

Xiao, Y., Y. Cai, Y.R. Bommineni, S.C. Fernando, O. Prakash, S.E. Gilliland, and G. Zhang. 2006. Identification and functional characterization of three chicken cathelicidins with potent antimicrobial activity. *J Biol Chem* 281: 2858-2867.

Xiao, Y., A.L. Hughes, J. Ando, Y. Matsuda, J.F. Cheng, D. Skinner-Noble, and G. Zhang. 2004. A genome-wide screen identifies a single beta-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins. *BMC Genomics* 5: 56.

Zaiou, M. and R.L. Gallo. 2002. Cathelicidins, essential gene-encoded mammalian antibiotics. *J Mol Med* 80: 549-561.

Zou, J., C. Mercier, A. Koussounadis, and C. Secombes. 2007. Discovery of multiple beta-defensin like homologues in teleost fish. *Mol Immunol* 44: 638-647.