Cell. Mol. Life Sci. 65 (2008) 1285 – 1294
1420-682X/08/071285-10
DOI 10.1007/s00018-008-8070-x
© Birkhäuser Verlag, Basel, 2008

Cellular and Molecular Life Sciences

Research Article

# Positive selection targeting the cathelin-like domain of the antimicrobial cathelicidin family

**S. Zhu**

Group of Animal Innate Immunity, State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Datun Road, Chaoyang District, Beijing 100101 (China), Fax: +86-010-64807099, e-mail: Zhusy@ioz.ac.cn

Received 8 February 2008; accepted 13 February 2008
Online First 7 March 2008

**Abstract.** The cathelin-like domain (CLD) of cathelicidins is grouped in the same superfamily with cystatins, natural cysteine protease inhibitors, due to their structural similarity. Intriguingly, human hCAP-18/LL37 and pig protegrin-3 (PG3) CLDs exhibit opposite effects against cathepsin L. Here, I evaluated the functional importance of the CLD through identifying whether positive selection has driven adaptive evolution of this domain. As a result, four positively selected sites were detected and three of them are located on a loop region previously recognized as a key determinant of the activating effect of the PG3 CLD. Analysis of amino acid variability of the CLD led to the discovery of a conserved region and three highly variable regions, in which two are subjected to positive selection. Positive selection targeting the variable regions provides a starting point for experimentally establishing a direct link between the observed amino acid changes and functional divergence of the CLD family.

**Keywords.** Phylogeny, codon-substitution model, innate immunity, antimicrobial peptides, cystatin.

## Introduction

Cathelicidins are one essential effector component of mammalian innate immunity, which provide hosts the first-line defense for rapidly clearing the infection of invading pathogenic microorganisms. Compared with other immune effectors, cathelicidins are unusual in that they have a substantial heterogenic C-terminal antimicrobial domain, which is linked to an evolutionarily conserved N-terminal cathelin-like domain (CLD) of approximately 99–114 residues [1–3]. In general, the cathelicidin gene is translated as a precursor in the cytoplasmic granules of neutrophil/polymorphonuclear leukocytes (PMN) with a signal peptide removed during translation to yield the proform with two-domains including the CLD and the antimicrobial domain. During the inflammatory response, the proform will be further processed through protease cleavage to release the antimicrobial domain to sites of microbial infection.

Extensive studies have focused on the antimicrobial domain of the cathelicidin family due to their central roles in both innate and adaptive immunity through direct antimicrobial activity and as immune modulators and mediators of inflammation. It is remarkable that the absence of the cathelicidin hCAP-18/LL37 in human neutrophils has been considered as a cause of morbus Kostmann disease [4], whereas mice with a knockout mutation in the cathelin-related antimicrobial peptide (CRAMP) has been found to be abnormally sensitive to *Streptococcal* A infections [5]. In addition, prevention of activation of the cathelicidin by inhibition of neutrophil elastase, a processing enzyme for the maturation of the C-terminal domain, impairs clearance of bacteria from wounds in pigs [6]. Expression of the pig PR-39 gene in mice can result in decreased bacterial load and mortality of experimental mice following challenge with bacteria [7]. Recent work performed by Carretero et al. (cited in [8]) further extends the effect of cathelicidins by finding the cell-signaling role of human cathelicidin whose expression can significantly alter multiple signaling pathways in a keratinocyte cell line and enhance wound re-epithelialization in ob/ob mice. While decreased cathelicidin expression in humans and mice has been associated with increased susceptibility to infection, their expression levels need to be accurately regulated. It has been shown that abnormally high level of expression of cathelicidin in rosacea represents an exacerbated innate immune response, which can reproduce elements of this disease [9].

Although the importance of the antimicrobial domain of cathelicidins in immune defense has been well documented, the CLD has attracted little attention despite their conservation within mammals throughout at least 100 million years of evolutionary history. From a structural perspective, CLDs and cystatins, a natural cysteine protease inhibitor widely distributed in multicellular organisms, can be grouped in the same superfamily due to their conservation of disulfide bridge pattern and global folding similarity, as identified by a helix cradled by a five-stranded β-sheet together with an appending domain [10]. Such structural conservation hints at a possible functional similarity, which has been confirmed by the work of Zaiou et al. [11] who demonstrated that human hCAP-18/LL37 CLD was able to inhibit protease activity of cathepsin L. Unexpectedly, our recent work showed that pig protegrin-3 (PG3) CLD can activate protease activity of cathepsin L and this activating effect is associated with one loop region of the CLD [12].

An emerging trend in the study of evolutionary mechanism of functional diversification of immune-related protein families is the observations of functional determinants of proteins that have frequently been subject to positive selection, which drives adaptive evolution and functional innovation of these proteins [13, 14]. Moreover, it is generally accepted that if the variability of a protein family can be shown to be driven by positive selection, its functional importance is established [15]. To evaluate the functional importance of CLDs, a class of often ignored immune molecules, I used a maximum likelihood method with codon-substitution models to detect positive selection signals in this family. My results provided convincing statistical evidences for adaptive evolution occurring at two variable regions of the CLD, with adaptive amino acid changes in these regions possibly leading to structural and functional alterations of this family.

## Materials and methods

**Sequence sources.** Sixty-three CLD sequences were obtained by BLASTP search of GenBank database (http://www.ncbi.nlm.nih.gov) using human CLD sequence as a query; all are derived from mammals besides two from chicken. Sequence names, species (organism) and GenBank accession numbers are listed in Table 1.

**Statistical analysis.** DNA sequences of CLDs were aligned by using Clustal X (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/) and further refined in reference to the multiple protein sequence alignment. Codons encoding amino acids of the CLD with the signal peptide and antimicrobial domain removed were used in my analysis, where seven sites were excluded due to the existence of indels. The neighbor-joining method implemented in MEGA 3.1 (http://www.megasoftware.net/) was selected to obtain tree topologies based on the aligned nucleotide sequences. The reliability of interior branches of the phylogeny was assessed with 1000 bootstrap resamplings by using nucleotide sequences with Kimura 2-parameter. The phylogeny was used to estimate nonsynonymous-to-synonymous rate ratio ($\omega = dN/dS$) by the maximum likelihood (ML) implemented in CODEML program of the PAML 3.15 software package (http://abacus.gene.ucl.ac.uk/software/paml.html).

To test for heterogeneous selective pressure among lineages [16], models of variable ω ratios among lineages were fitted by ML to the alignment of the 63 CLD sequences. The one-ratio model assumes the same ratio for all branches in the phylogeny, whereas the free-ratios model assumes an independent ratio for each branch. The two models can be compared using the likelihood ratio test (LRT) to see whether the ω ratios are different among lineages. Twice the log likelihood difference between the two models ($2\Delta l$) was compared against $\chi^2$ with critical values with $df = 123$ to be 149.885 and 162.398 at 5% and 1% significance levels, respectively.

To test for heterogeneous selective pressure at amino acid sites [17], I calculated ω ratio distribution among sites under two pairs of models using CODEMLsites program: M1a (Nearly Neutral) against M2a (Positive Selection), and M7 (Beta) against M8 (Beta & ω). Positive selection is defined as presence of some codons at which $\omega > 1$. M1a and M7 belong to null models that do not allow for any codons with $\omega > 1$, while M2a and M8 represent more general models

**Table 1.** Accession numbers (ANs) of the 63 cathelicidin genes used in the analysis.

| Gene | Species (organism) | AN | Gene | Species (Organism) | AN |
|---|---|---|---|---|---|
| *Saurischia* | | | | | |
| GgCAMP1(Fowlicidin-1) | *Gallus gallus* (Chicken) | DQ092351 | BtCAMP4(CATHL4) | *Bos tauru*(Cow) | X67340 |
| GgCAMP2(Fowlicidin-2) | *Gallus gallus* (Chicken) | DQ092352 | BtCAMP5(CATHL5) | *Bos taurus* (Cow) | X97609 |
| *Glires* | | | BtCAMP6(CATHL6) | *Bos taurus* (Cow) | X97608 |
| CpCAMP(CAP11) | *Cavia porcellus* (Guinea pig) | D87405 | BtCAMP7(CATHL7) | *Bos taurus* (Cow) | Y12728 |
| MmCAMP1(Bactenecin F1) | *Mus musculus* (Mouse) | U95002 | BtCAMP8(BAC5) | *Bos taurus* (Cow) | L02650 |
| MmCAMP2(MmNgp) | *Mus musculus* (Mouse) | NM_008694 | BbCAMP1(CATHL-1) | *Bubalus bubalis* (Water buffalo) | DQ832665 |
| MmCAMP3(MCLP) | *Mus musculus* (Mouse) | X94353 | BbCAMP2(CATHL2) | *Bubalus bubalis* (Water buffalo) | EF050453 |
| OcCAMP(p15(R3,W88)) | *Oryctolagus cuniculus* (Rabbit) | L07588 | BbCAMP4(CATHL-4) | *Bubalus bubalis* (Water buffalo) | AJ812216 |
| OcCAMP1(CAP18) | *Oryctolagus cuniculus* (Rabbit) | M73998 | ChCAMP1(BAC7.5) | *Capra hircus* (Goat) | AJ243125 |
| OcCAMP2(p15R) | *Oryctolagus cuniculus* (Rabbit) | S68154 | ChCAMP2(MAP28) | *Capra hircus* (Goat) | AJ243126 |
| RnCAMP1(CRAMP) | *Rattus norvegicus* (Rat) | XM_236642 | OaCAMP1(BAC7.5’) | *Ovis aries* (Sheep) | U60598 |
| RnCAMP2(CRAMP) | *Rattus norvegicus* (Rat) | AF484553 | OaCAMP2(BAC7.5) | *Ovis aries* (Sheep) | NM_001009301 |
| RnCAMP3(NGP) | *Rattus norvegicus* (Rat) | XM_236646 | OaCAMP3(BAC6) | *Ovis aries* (Sheep) | AH005454 |
| *Primates* | | | OaCAMP4(BAC11) | *Ovis aries* (Sheep) | AH005455 |
| AfCAMP1 | *Ateles fusciceps* (Black-headed spider monkey) | DQ471355 | OaCAMP5(BAC5’) | *Ovis aries* (Sheep) | NM_001009787 |
| AfCAMP2 | *Ateles fusciceps* (Black-headed spider monkey) | DQ471354 | OaCAMP6(BAC5) | *Ovis aries* (Sheep) | U60599 |
| CjCAMP | *Callithrix jacchus* (Marmoset) | DQ471358 | OaCAMP7(MAP-29) | *Ovis aries* (Sheep) | U60600 |
| CaCAMP | *Cercopithecus aethiops* (Vervet monkey) | DQ471356 | OaCAMP8(SC5) | *Ovis aries* (Sheep) | X92757 |
| CcCAMP | *Cebus capucinus* (White-faced capuchin) | DQ471357 | OaCAMP9(MAP-34) | *Ovis aries*) (Sheep) | U60597 |
| GgorCAMP | *Gorilla gorilla* (Gorilla) | DQ471359 | OaCAMP10(DODE) | *Ovis aries* (Sheep) | NM-001009772 |
| HsCAMP(CAP-18) | *Homo sapiens* (Human) | X89658 | SsCAMP1(PR2) | *Sus scrofa* (Pig) | NM_213863 |
| HmCAMP | *Hylobates moloch* (Silvery gibbon) | DQ471364 | SsCAMP2(PMAP37) | *Sus scrofa* (Pig) | L39641 |
| MmulCAMP(CAP-18) | *Macaca mulatta* (Rhesus monkey) | AF181954 | SsCAMP3(PG3) | *Sus scrofa* (Pig) | X83267 |
| HcCAMP | *Nomascus gabriellae* (Yellow-cheeked crested gibbon) | DQ471360 | SsCAMP4(PMAP36) | *Sus scrofa* (Pig) | L29125 |
| NlCAMP | *Nomascus leucogenys* (White-cheeked crested gibbon) | DQ471363 | SsCAMP5(PG5) | *Sus scrofa* (Pig) | X84096 |
| NlCAMP | *Nomascus leucogenys* (White-cheeked crested gibbon) | DQ471361 | SsCAMP6(PR39’) | *Sus scrofa* (Pig) | X89201 |
| PyCAMP | *Pongo pygmaeus* (Sumatran orangutan) | DQ471370 | SsCAMP7(PR39) | *Sus scrofa* (Pig) | NM-214450 |
| PcCAMP | *Presbytis cristata* (Silvered leaf monkey) | DQ471367 | SsCAMP8(PMAP23) | *Sus scrofa* (Pig) | L26053 |
| PoCAMP | *Presbytis obscura* (Dusky leaf monkey) | DQ471368 | *Perissdactyls* | | |
| SoCAMP | *Saguinus oedipus* (Cotton-top tamarin) | DQ471372 | EcCAMP1(eCATH-1) | *Equus caballus* (Horse) | AJ224927 |
| *Cetartiodactyls* | | | EcCAMP2(eCATH-2) | *Equus caballus* (Horse) | AJ224928 |
| BtCAMP1(CATHL1) | *Bos taurus* (Cow) | Y09472 | EcCAMP3(eCATH-3) | *Equus caballus* (Horse) | AJ224929 |
| BtCAMP2(FALL-39) | *Bos taurus* (Cow) | XM_602399 | *Carnivora* | | |
| BtCAMP3(CATHL3) | *Bos taurus* (Cow) | Y09471 | CfCAMP | *Canis familiaris* (Dog) | NM_001003359 |

that do. Two LRTs that calculated $2\Delta l$ were compared against $\chi^2$ with df = 2 with critical values to be 5.99 and 9.21 at 5 % and 1 % significance levels, respectively. When the LRT suggests positive selection, the Bayes empirical Bayes (BEB) method was used to calculate the posterior probabilities that each codon is from the site class of positive selection under models M2a and M8 [17].

Sites with a high probability of coming from the class with $\omega > 1$ are likely to be under positive selection and were mapped on the CLD structure using the MOLMOL program (http://hugin.ethz.ch/wuthrich/software/molmol/).

**Mapping of variability and conservation of CLD sites.** To calculate conservation scores for sites in the CLD, I used the ConSurf program (http://consurf.tau.ac.il/) to analyze the 63 amino acid sequences of CLDs under default parameters. ConSurf identifies functionally important regions on the surface of a protein based on the sequence alignment and phylogenetic tree through an empirical Bayesian approach [18].

## Results

**Ancient and lineage-specific gene duplications.** Firstly, I conducted phylogenetic analyses using 63 aligned nucleotide sequences of the CLD proteins to construct the neighbor-joining (NJ) tree (Fig. 1), which is used in the analysis of variation in the ω = dN/dS by the models of codon-substitutions. This tree divides mammalian cathelicidin family members into two distinct groups, designated as the α- and β-subfamilies. The α-subfamily includes the majority members of this family and its distribution is across the whole evolutionary lineage of vertebrates (cetartiodactyls, perissodactyls, glires, primates, carnivora, saurischia), whereas the β-subfamily represents a smaller group comprising only five sequences that all are derived from glires. A slightly shorter L1 loop is a feature of the β-subfamily (data not shown). Because both α- and β-subfamilies of cathelicidins are present in the glire lineage, it appears that an ancient duplication might have happened in the ancestor of all the vertebrate lineages and members from β-subfamily have been lost in the most lineages in the subsequent evolution of vertebrates. This explanation is consistent with the fact that gene loss, as a general evolutionary event, frequently occurred in human and other vertebrates [19].

In addition to ancient duplication, phylogenetic reconstruction also identified some lineage-specific gene duplications in the genomes of laurasiatheria, such as cetartiodactyls and perissodactyls. This event sharply contrasts with primates, which possess only single copy of cathelicidin gene. Wide duplication in Laurasiatheria often generates a conserved CLD with considerably variable antimicrobial repertoire of host defense molecules by a yet unknown mechanism [3]. According to the clustering pattern of the CLD in the tree, I found that the cathelicidin gene has repeatedly undergone independent duplication in the pig lineage after emerged from other catartiodactyls. The phylogeny clearly divided these duplicates into three distinct clades: clade I includes PG3, PG5, PMAP-37 and PR-2; and clade II includes PMAP-36, PR39 and PR39'. PMAP-23 alone constitutes clade III that shares a closer evolutionary distance to human CLD, supporting their orthologous relationship.

**Detection of positive selection signal.** To test for variable ω ratios among lineages, I undertook the LRT test to compare the one-ratio model, which assumes the same ω for all lineages, and the free-ratio model, which assumes an independent ω ratio for each branch. The log likelihood value under the one-ratio is $l0 = -5201.9681$, while the value for the free-ratio model is $l1 = -5120.5879$. Twice the log likelihood difference, $2\Delta l = 2(l1 - l0) = 162.7602$, which is slightly greater than the critical value (162.398) at 1 % significance level from a $\chi^2$ distribution with df = 123 (since the free-ratio model involves 124 ω parameters for the 124 branches, while the one-ratio model assume one), revealing a heterogeneous selective pressure among lineages. When checking ω values along each branch calculated by the free-ratio models, I found that many branches of the vertebrate CLD phylogeny, including some internal branches have $\omega > 1$, which shows evidence for its evolution under positive selection (Fig. 1). It thus appears that the CLD may have been subjected to positive selection within mammals for at least 100 million years and this selection is remarkably variable among lineages.

I next tested whether there are variable ω ratios at the amino acid sites. The log-likelihood values and parameter estimates of the CLD under various site models are shown in Table 2. In the model M0, a single ω is assumed for all sites in the alignment across all lineages. The ω value under M0 was estimated to be 0.3357 with the log-likelihood score $l = -5201.9681$. To test positive selection of the CLD, I selected four models (M1a, M2a, M7, and M8) to construct two LRTs (M1a/M2a and M7/M8). Parameter estimates from the models suggest that M0 shows a worse fit for the data than any other models because of its much lower log-likelihood value than all other models. This discards the possibility all sites in the CLD alignment have the same ω ratio. In contrast, two selection models (M2a and M8) do detect the existence of a

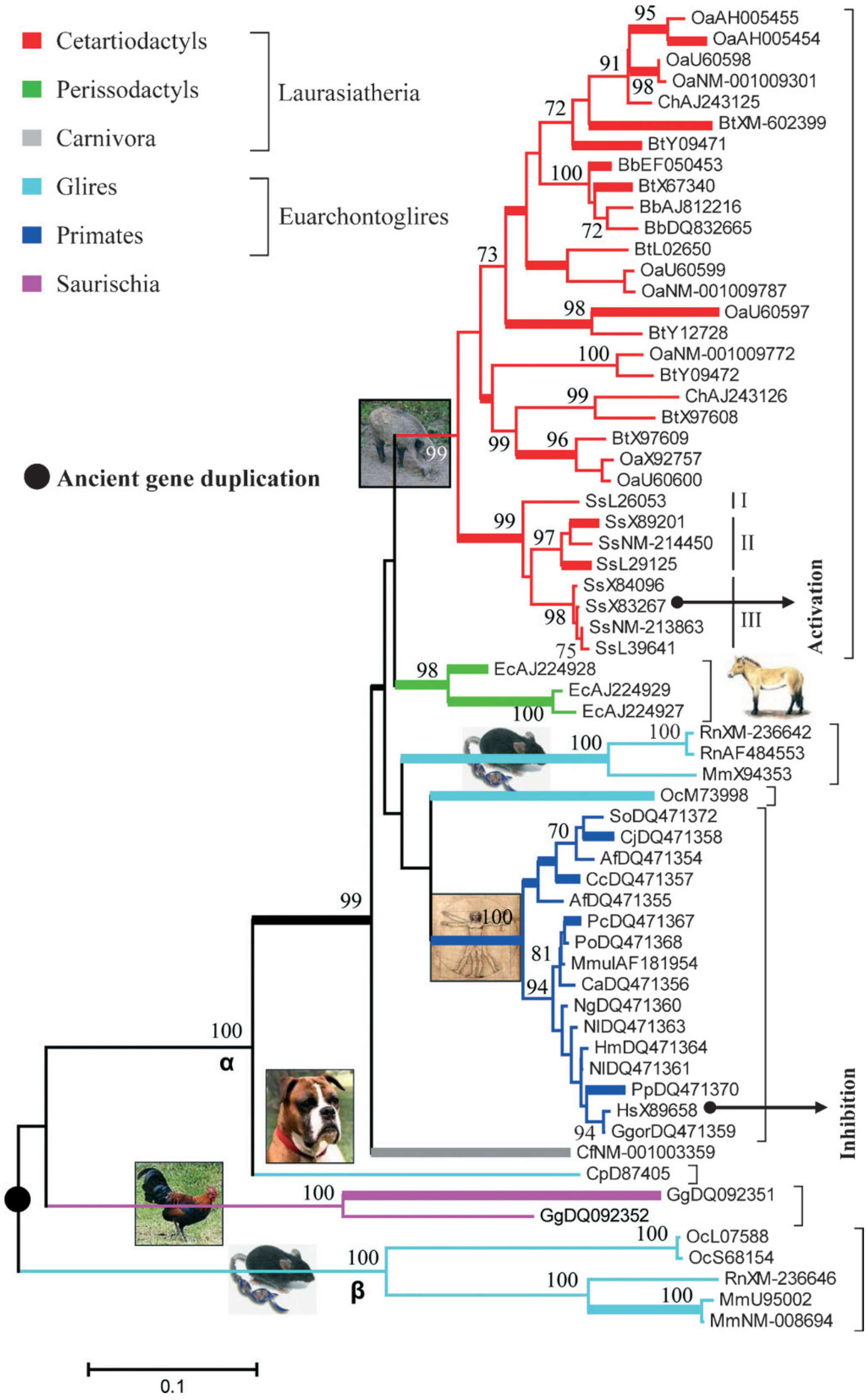

**Figure 1.** Phylogeny of CLDs. To minimize confusion, all sequences are referred to by their CoreNucleotide accession numbers (http://www.ncbi.nlm.nih.gov) beginning with the first letters of the genus and species names. For the species abbreviations, see Table 1. Only bootstrap values ≥70 % are shown here. Bars show total nucleotide distance. Branches with rates of numbers of nonsynonymous and synonymous substitutions >1 are indicated by thick lines, with the exception of $dS = 0$.

substantial proportion of positively selected sites. In particular, these two models suggest a similar proportion (0.0395–0.0439) of sites evolved by positive selection with a similar ω (1.8932–2.4560). LRTs indicate that these models significantly increase the likelihood scores compared with models with no positive selection. For example, M1a does not allow for sites with ω >1, whereas the selection model M2a adds an additional site class, with the ω ratio estimated to be 2.4560. $2\Delta l$ is 14.32, which is much greater than the critical value from a χ2 distribution with df = 2, suggesting that M2a fits the data better than M1a and thus indicates the existence of positively selected sites with ω >1. Additional LRT performed by comparing the $2\Delta l$ values between M7 and M8 led to consistent results (Table 2).

**Table 2.** Maximum likelihood estimates of parameters and sites inferred to be under positive selection for the cathelin-like domain of the cathelicidin family

| Model | $p$ | $l$ | $\kappa$ | Estimates of parameters | Positively selected sites |
|---|---|---|---|---|---|
| M0 (one-ratio) | 1 | –5201.9681 | 3.0265 | $\omega$= 0.3357 | None |
| M1a (Nearly Neutral) | 2 | –5081.3203 | 3.5257 | p0=0.5601, $\omega$0=0.1517<br>p1=0.4400, $\omega$1=1.0000 | Not allowed |
| M2a (Positive Selection) | 4 | –5074.1603 | 3.5932 | p0=0.5473, $\omega$0= 0.1570<br>p1=0.4132, $\omega$1=1.0000<br>p2=0.0395, $\omega$2= 2.4560 | 89T$^{*}$, 115Q$^{**}$, 117K$^{**}$, 119P |
| M7 (beta) | 2 | –5049.2317 | 3.4017 | p=0.5568, q= 0.8915 | Not allowed |
| M8 (beta & $\omega$) | 4 | –5042.0825 | 3.4551 | p1= 0.0439, $\omega$=1.8932<br>p0= 0.9561<br>p=0.6386, q=1.1783 | 89T$^{*}$, 115Q$^{**}$, 117K$^{**}$, 119P |

NOTE. $p$ is the number of parameters in the $\omega$ distribution; $l$ is the log likelihood; $\kappa$ is transition/transversion rate ratio. 2$\Delta l$: M1a/M2a = 14.32; M7/M8 = 14.30. Positively selected sites identified by the Bayes empirical Bayes (BEB) method under models M2a and M8 with posterior probabilities (p) $\geq$0.95 and those with p$\geq$0.99 are indicated by $^{*}$ and $^{**}$.

**Identification of positively selected sites.** Because the LRT suggested positive selection, the BEB method was used to calculate the posterior probabilities that each codon is from the site class of positive selection under models M2a and M8. Significantly, two selection models identified four identical sites evolved under positive selective pressure. They are 89T, 115Q, 117K, 119P (residues numbered according to the PG3 CLD), of which the latter three sites were all found to be located in the L2 loop of the PG3 CLD (V 111-D121) (Fig. 2). Among the three positively selected sites, two differ considerably between CLDs of pig PG-3 and human hCAP18 (117K/R and 119P/S). Given that previous structural analysis has identified main conformational changes in the L2 loop mediated by the *cis-trans* isomerization of D118-P119 amide bonds [10], it is remarkable that only several pig CLDs were found to possess a proline at site 119. This excludes that all except these CLDs have a conformational flexibility in the loop 2. The positively selected site 89 forms the end of a β-turn comprising 86PRPT89, which contains the R87-P88 amide bond that is also able to adopt both the *cis-trans* conformations. This site is conserved between CLDs of pig and human, in which all are occupied by a Thr residue. However, high variability in other lineages can be identified by the diversity of amino acid types, which include Thr, Ser, Ala, Ile, Met, Glu, Asn, Leu, and Gln.

**Variable and conserved sites of the CLD.** Given that variability driven by positive selection and conservation constrained by purifying selection reflect different aspects of biological functions of proteins [15], recognition of such regions will undoubtedly be useful for understanding of structure-function relationship of a protein family. Based on the ConSurf analysis of 63 CLD sequences, I can draw several remarkable observations in terms of their structural characteristics (Fig. 3): (1) Conformationally flexible L2 loop and the β-turn, both detected under positive selection, are highly variable in some sites with the scores =1, in which the four positively selected sites are included. In particular, six continuous variable sites ($^{113}$DQIKDP$^{118}$) in the L2 loop are fully exposed to the molecular surface that provides structural basis for protein-protein interaction. (2) Apart from the two variable positively selected regions, ConSurf identified a new variable region located in the L1 loop including three sites (65K, 66A and 68E) with the scores =1, and three other sites (63P, 70P and 71G) with the scores =3. Although the codon-substitution models M2a and M8 did not detect positive selection signal in this region, its variability could reflect function given two facts, *i.e.*, the equivalent loop region of cystatins has been identified as one major functional loop that directly interacts with the cysteine proteases, and in the complex model previously established by us, the L1 loop of PG3 CLD is close to the active site of cathepsin L [12], supporting its functional importance. Absence of selection in the variable L1 loop could be explained by the power limitation of the maximum likelihood approach in detecting weak positive selection or episodic evolution of some sites. (3) An interesting finding from this analysis is that ConSurf identified a conserved patch primarily contributed by two β-strands (Fig. 3) (here named β-patch), which include some residues (55R, 58L, 80V, 81K, 82E, 83T and 98F) with the scores =9. This patch is located on the same face with the L1 loop but is more buried in the molecular interior. Although the functional significance of this patch remains unknown at present, spatial clustering in the structure from noncontiguous primary sequences suggests a possible conserved role key for all the members in this family. Given that a conserved surface patch itself suggests function and the development of variable and

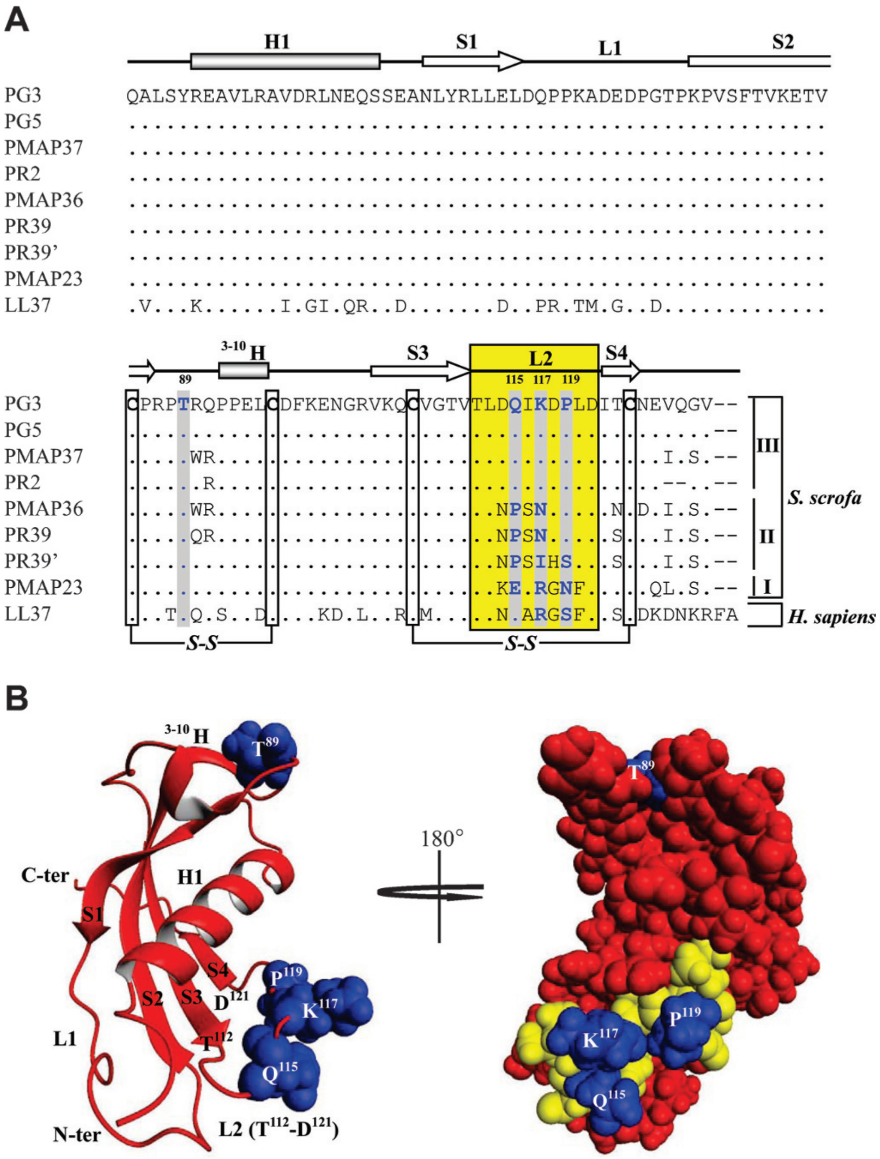
A
H1
S1
L1
S2
PG3 QALSYREAVLRAVDRLNEQSSEANLYRLLELDQPPKADEDPGTPKPVSFTVKETV
PG5
PMAP37
PR2
PMAP36
PR39
PR39'
PMAP23
LL37 .V...K......I.GI.QR..D.......D..PR.TM.G..D.............
3-10 H
S3
L2
S4
89
115 117 119
PG3 CPRPTRQPPELCDFKENGRVKQCVGTVTLDQIKDPLDITCNEVQGV--
PG5 --
PMAP37 .....WR ... .I.S.--
PR2 .....R ... --..--
PMAP36 .....WR ... NPSN ... N.D.I.S.--
PR39 .....QR ... NPSN ... S...I.S.--
PR39' NPSIHS ... S...I.S.--
PMAP23 KE.RGNF ... .QL.S.--
LL37 ...T.Q.S..D...KD.L..R.M.....N.ARGSF..S.DKDNKRFA
S-S
S-S
III
II
I
S. scrofa
H. sapiens
B
3-10 H
T89
C-ter
H1
S1
S4
P119
S2
S3
D121
K117
L1
T112
Q115
N-ter
L2 (T112-D121)
180°
T89
K117
P119
Q115

**Figure 2.** The L2 loop is a hot spot for positive selection. (*A*) Multiple sequence alignment of the cathelin-like domain of the cathelicidin family from pig and human. The L2 loop shadowed in yellow contains three positively selected sites highlighted in blue. Sequence origin: PMAP-23 (L26053); PR-39' (X89201); PR-39 (NM_214450); PMAP-36 (L29125); PG5 (X84096); PG3 (X83267); PR-2 (NM_213863); PMAP37 (L39641). (*B*) Mapping positively selected sites onto the CLD structure (pdb entry 1N5H).

conserved functional regions in a protein structure is a common strategy due to functional requirement, the recognition of the β-patch in the CLD is of significance regarding its interaction with other components. Possibly, the β-patch represents a region for the entry of the CLD to antigen presentation cells (APCs) to regulate the activity of cathepsin L if this process occurs [12].

## Discussion

Studies have shown that host genes involved in immunity are frequently subjected to positive selection during evolution. Some typical examples include major histocompatibility complex (MHC) class I [20, 21], α-defensin [22], lymphocyte protein CD45 [23] and antiviral enzyme APOBEC3G [24]. Recently, Zelezetsky et al. [25] evaluated selective pressures acting on the different domains of the cathelicidin family from 21 primate species using the maximum-likelihood approach, and found the region encoding the antimicrobial domain presented 71 % of codons having evolved under positive selection. However, these authors showed that the whole signal sequence and the CLD evolved under purifying or neutral selection. The lack of such selection signal in the CLD appears to be due to highly similar sequences used by these authors, which decreased the power of this approach. As pointed out by Yang (http://abacus.gene.ucl.ac.uk/software/paml.html), the maximum likelihood analysis uses the number of synonymous and nonsynonymous changes to measure whether there is an excess of nonsynonymous changes relative to

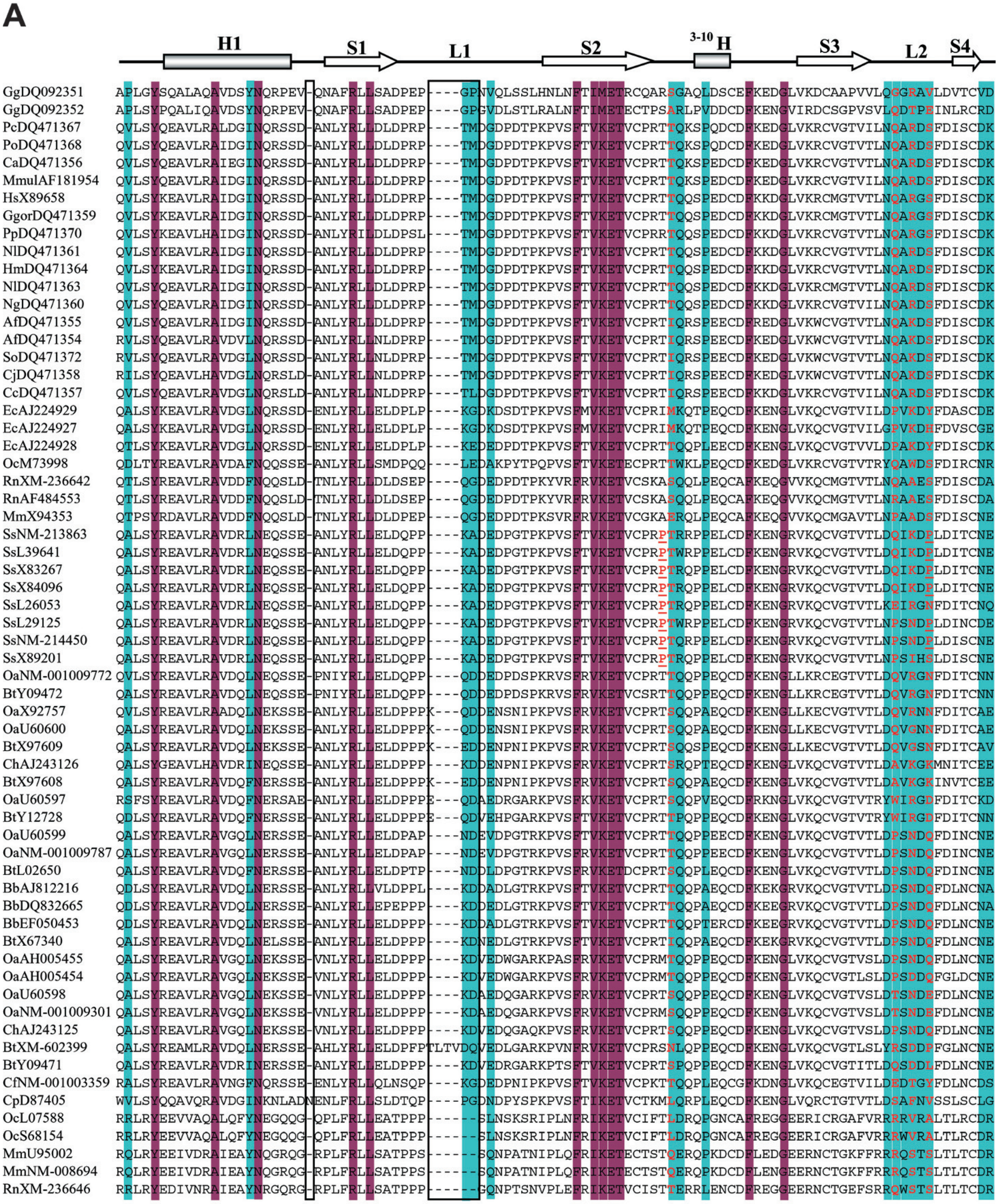

**Figure 3.** Variability and conservation of CLDs detected by the ConSurf program. (*A*) Multiple sequence alignment of CLDs. Variable sites with the scores =1 and conserved sites with the scores =9 are, respectively, shadowed in cyan and purple. Positively selected sites are highlighted in red. Two proline residues (88P and 119P) mediating *cis-trans* isomerization are underlined once. Sites not used in the evolutionary analysis due to the existence of gaps are boxed. Secondary structure elements extracted from the 3D coordinates of the CLD of protegrin-3 (pdb entry 1N5P) are indicated at the top of the alignment. (*B*) The conservation pattern calculated using ConSurf for the CLD (pdb entry 1N5H). The CLD is presented as a space-filled model, and colored according to the conservation scores.

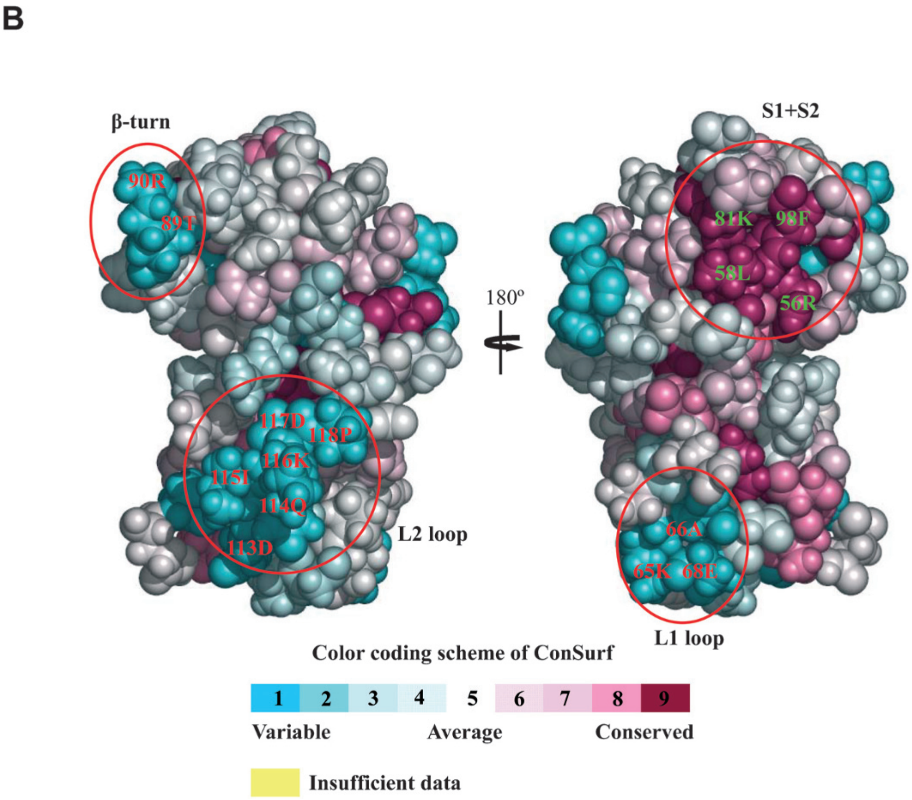

**Figure 3.** (continued)

synonymous changes. In this case, many sequences to accumulate changes at each site are needed in the site models. When adding more sequences derived from diverse lineages, I successfully detected adaptive evolution that affects several sites of the CLD and many lineages of the phylogeny.

For a long time, the function of the N-terminal CLD has been theorized to only act as a balancer to neutralize the C-terminal cationic antimicrobial domain during intracellular transport and storage to avoid potential intracellular cytotoxicity [26]. However, *in vitro* experiments have identified putative immune-related functions of human and pig CLDs [11, 12], while nothing is known regarding their *in vivo* effects. The detection of positive selection in the CLD has implications for establishing its fundamental biological importance given that natural selection has provided evolutionary innovations for molecular adaptation [15, 27]. In this regard, human and pig CLDs offer us valuable information: single copy of human CLD acts as an inhibitor of cathepsin L, whereas the multigene family of pig CLDs, originating from duplication and subsequent divergence, has developed an activating effect on cathepsin L in some members. Evidence for adaptive evolution combined with *in vitro* functional data clearly imply that functional diversification of the CLD represents an evolutionary advantage for undergoing ecological adaptation to local environment and diverse pathogens in a species-specific manner.

From an evolutionary perspective, extrinsic proteases often prompt adaptive change of their inhibitors through a host-pathogen arms race [28]. However, the driving force responsible for the CLD divergence in the pig lineage appears not to be derived from the evolutionary pressure of cathepsin L because these members target the same enzyme. In this case, indirect immune pressure could be a driving force that exerts its role through cathepsin L in regulating antigen-presentation efficiency [12]. Given that positive selection frequently drives functional innovation of protein families [29], it is interesting to study whether the CLD has also developed other functions that are unrelated to modulators of cathepsin L activity.

If we consider that the highly exposed L2 loop is one major functional determinant of PG3 CLD in activating cathepsin L [12], positive selection in this region is of considerable interest, and provides evolutionary evidence for its functional importance in developing an activating effect of the CLD through amino acid substitutions. Some similar examples can be found: (1) The selective force acting upon the MHC class I is concentrated on its antigen recognition site (ARS), a cleft recognizing and binding diverse foreign antigens [21, 29]. (2) Extrinsic proteases, as selective forces through the host-pathogen arms race, frequently prompt accelerated amino acid substitutions of their inhibitors, in which positive selection targets the reactive center regions of the inhibitors [28]. (3) A similar evolutionary scenario has also been observed in the inhibitor domains of porcine elafin family members [30]. (4) A critical species-specific retroviral restriction domain of TRIM5α (a major innate immunity protein of primates) has also been characterized as a hot spot for positive selection [31]. Of course, to reach a decisive conclusion in terms of roles of these positively selected sites in the functional switch of CLDs between human and pig, exchange of their positively selected site residues and confirmation of the functional consequence of this modification will be needed.

In conclusion, the work presented here highlights for the first time the conservation and variability characteristic of the CLD, and provides statistical data in favor of its variable regions subjected to positive selection. Evolutionary identification of such positively selected regions not only provides convincing evidence for adaptive evolution of the CLD, but also pinpoints crucial molecular targets for mutational analysis of its functional surfaces.

*Acknowledgements.* This work was supported by grants from the National Natural Science Foundation of China (90608009 and 30621003) and the 'Bairen Plan' from the Chinese Academy of Sciences to S. Z.

1 Zaiou, M. and Gallo, R. L. (2002) Cathelicidins, essential gene-encoded mammalian antibiotics. J. Mol. Med. 80, 549–561.

2 Zanetti, M. (2005) The role of cathelicidins in the innate host defenses of mammals. Curr. Issues Mol. Biol. 7. 179–196.

3 Tomasinsig, L. and Zanetti. M. (2005) The cathelicidins – Structure, function and evolution. Curr. Protein Peptide Sci. 6, 23–34.

4 Boman, H. G. (2003) Antibacterial peptides: Basic facts and emerging concepts. J. Intern. Med. 254, 97–215.

5 Nizet, V., Ohtake, T., Lauth, X., Trowbridge, J., Rudisill, J., Dorschner, R. A., Pestonjamasp, V., Piraino, J., Huttner, K. and Gallo, R. L. (2001) Innate antimicrobial peptide protects the skin from invasive bacterial infection. Nature 414, 454–457.

6 Cole, A. M., Shi, J., Ceccarelli, A., Kim, Y. H., Park, A. and Ganz, T. (2001) Inhibition of neutrophil elastase prevents cathelicidin activation and impairs clearance of bacteria from wounds. Blood 97, 297–304.

7 Lee, P. H., Ohtake, T. Zaiou, M., Murakami, M., Rudisill, J. A., Lin, K. H. and Gallo, R. L. (2005) Expression of an additional cathelicidin antimicrobial peptide protects against bacterial skin infection. Proc. Natl. Acad. Sci. USA 102, 3750–3755.

8 Gallo, R. L. (2008) Sounding the alarm: multiple functions of host defense peptides. J Invest Dermatol. 128, 5–6.

9 Yamasaki, K., A. Nardo, D., Bardan, A., Murakami, M., Ohtake, T., Coda, A., Dorschner, R. A., Bonnart, C., Descargues, P., Hovnanian, A., Morhenn, V. B. and Gallo, R. L. (2007) Increased serine protease activity and cathelicidin promotes skin inflammation in rosacea. Nat. Med. 13, 975–980.

10 Yang, Y., Sanchez, J. F., Strub, M. P., Brutscher, B. and Aumelas, A. (2003) NMR structure of the cathelin-like domain of the protegrin-3 precursor. Biochemistry 42, 4669–4680.

11 Zaiou, M., Nizet, V. and Gallo, R. L. (2003) Antimicrobial and protease inhibitory functions of the human cathelicidin (hCAP18/LL37) prosequence. J. Invest. Dermatol. 120, 810–816.

12 Zhu, S, Wei, L, Yamasaki, K. and Gallo, R. L. (2008) Activation of cathepsin L by the cathelin-like domain of protegrin-3. Mol. Immunol. doi:10.1016/j.molimm.2008.01.007.

13 Yang, Z. (2002) Inference of selection from multiple species alignments. Curr. Opin. Genet. Dev. 12, 688–694.

14 Yang, Z. (2005) The power of phylogenetic comparison in revealing protein function. Proc. Natl. Acad. Sci. USA 102, 3179–3180.

15 Yang Z. (2002) Inference of selection from multiple species alignments. Curr. Opin. Genet. Dev. 12, 688–694.

16 Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15, 568–573.

17 Yang, Z., Wong, W. S. W. and Nielsen, R. (2005) Bayes empirical Bayes inferences of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118.

18 Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal, N. (2005) ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. 33, W299–W302.

19 Wang X, Grus W. E. and Zhang, J. (2006) Gene losses during human origins. PLoS Biol. 4: e52. DOI:10.1371/journal.-pbio.0040052.

20 Yang, Z. and Swanson, W. J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. 19, 49–57.

21 Hughes, A. L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 1335, 167–170.

22 Lynn, D. J., Lloyd, A. T., Fares, M. A. and O'Farrelly, C. (2004) Evidence of positively selected sites in mammalian α-defensins. Mol. Biol. Evol. 21, 819–827.

23 Filip, L. C. and Mundy, N. I. (2004) Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates. Mol. Biol. Evol. 21, 1504–1511.

24 Sawyer, S. L., Emerman, M. and Malik, H. S. (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol. 2, 1278–1285.

25 Zelezetsky, I., Pontillo, A., Puzzi, L., Antcheva, N., Segat, L., Pacor, S., Crovella, S. and Tossi, A. (2006) Evolution of the primate cathelicidin. Correlation between structural variations and antimicrobial activity. J. Biol. Chem. 281, 19861–19871.

26 Martin, E., Ganz, T. and Lehrer, R. I. (1995) Defensins and other endogenous peptide antibiotics of vertebrates. J. Leukoc. Biol. 58, 128–136.

27 Anisimova, M. and Liberles, D. A. (2007) The quest for natural selection in the age of comparative genomics. Heredity 99, 567–579.

28 Robert, E. H. and Nicholas, D. H. (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 326, 96–99.

29 Yang Z (2006) Computational Molecular Evolution. Oxford University Press, Oxford.

30 Tamechika, I., Itakura, M., Saruta, Y., Furukawa, M., Kato, A., Tachibana, S. and Hirose, S. (1996) Accelerated evolution in inhibitor domains of porcine elafin family members. J. Biol. Chem. 271, 7012–7018.

31 Sawyer, S. L., Wu, L. I., Emerman, M. and Malik, H. S. (2005) Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. Proc. Natl. Acad. Sci. USA 102, 2832–2837.