Research

Duplication and selection in the evolution of primate β-defensin genes

Colin AM Semple, Mark Rolfe and Julia R Dorin

Address: MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK.

Correspondence: Colin AM Semple. E-mail: Colin.Semple@hgu.mrc.ac.uk

Published: 17 April 2003

Genome ***Biology*** 2003, **4**:R31

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/5/R31

Received: 27 February 2003
Revised: 18 March 2003
Accepted: 3 April 2003

Abstract

Background: Innate immunity is the first line of defense against microorganisms in vertebrates and acts by providing an initial barrier to microorganisms and triggering adaptive immune responses. Peptides such as β-defensins are an important component of this defense, providing a broad spectrum of antimicrobial activity against bacteria, fungi, mycobacteria and several enveloped viruses. β-defensins are small cationic peptides that vary in their expression patterns and spectrum of pathogen specificity. Disruptions in β-defensin function have been implicated in human diseases, including cystic fibrosis, and a fuller understanding of the variety, function and evolution of human β-defensins might form the basis for novel therapies. Here we use a combination of laboratory and computational techniques to characterize the main human β-defensin locus on chromosome 8p22-p23.

Results: In addition to known genes in the region we report the genomic structures and expression patterns of four novel human β-defensin genes and a related pseudogene. These genes show an unusual pattern of evolution, with rapid divergence between second exon sequences that encode the mature β-defensin peptides matched by relative stasis in first exons that encode signal peptides.

Conclusions: We conclude that the 8p22-p23 locus has evolved by successive rounds of duplication followed by substantial divergence involving positive selection, to produce a diverse cluster of paralogous genes established before the human-baboon divergence more than 23 million years ago. Positive selection, disproportionately favoring alterations in the charge of amino-acid residues, is implicated as driving second exon divergence in these genes.

Background

The vertebrate innate immune system provides protection against a wide range of pathogenic microorganisms, and defensins are an important component of this response as well as having a role in adaptive immunity. In mammals, the defensins can be divided into the α- and β-defensin subfamilies on the basis of differences in the spacing of six, conserved cysteine residues. The α-defensins are produced by neutrophils and intestinal Paneth cells, whereas the β-defensins are mainly produced by epithelial cells in contact with the environment. The functions of human β-defensins seem to be disrupted in cystic fibrosis and inflammatory skin lesions such as psoriasis [1,2]. A fuller knowledge of the human complement of β-defensins may therefore be useful in

understanding human disease as well as in the design of novel, synthetic antimicrobial peptides.

The known human β-defensin genes show a conserved two-exon structure: the first exon encodes a signal peptide whereas the second exon encodes a short propiece and the mature defensin peptide with a characteristic six-cysteine motif and many basic amino-acid residues [3]. The β-defensin genes are present at five syntenic loci in the human and mouse genomes, with the main locus on human chromosome 8p22-23 and mouse chromosome 8A3 [4]. All four, full-length, human β-defensins that are present in the public databases are from 8p22-23 (GenBank sequence accession numbers are human β-defensin 1 or *DEFB1*, Q09753; *DEFB4* (formerly *DEFB2*), O15263; *DEFB103* (formerly *DEFB3*), NP_061131; *DEFB104* (formerly *DEFB4*), CAC85520), but there are substantial differences in their coding sequences, expression patterns and antimicrobial activities. *DEFB1* is constitutively expressed in many tissues (respiratory tract, kidney, urogenital and oral cavity epithelia) whereas *DEFB4* is expressed in response to bacterial infection or proinflammatory agonists in respiratory tract epithelial cells, and epidermal and gingival keratinocytes. Both DEFB1 and DEFB4 proteins have salt-sensitive, bactericidal activity against a spectrum of Gram-positive and Gram-negative enteric, urinary tract, and respiratory bacteria *in vitro* [5]. DEFB103 is expressed in epithelial cells, adult heart, skeletal muscle, placenta and fetal thymus, it has broad-spectrum antimicrobial activity under conditions of low salt and (unusually among β-defensins) it retains activity against *Staphylococcus aureus* even in physiological saline. *DEFB104* achieves highest expression in the testis (with lower levels in gastric antrum, neutrophils, uterus, thyroid, lung and kidney) and was found to be inducible in the respiratory epithelium upon exposure to *Pseudomonas aeruginosa* or *Streptococcus pneumoniae* [3].

The evolution of various genes involved in the vertebrate immune system has involved duplication followed by selection to provide responses to a wide range of pathogens, with well documented examples in immunoglobulin [6] and major histocompatibility complex genes [7]. Hughes and Yeager [8] studied the evolution of α-defensins and found evidence for duplication followed by diversification driven by positive selection. Similar phenomena were also implicated in the evolution of bovine β-defensins [9] and amphibian antimicrobial peptides [10]. In contrast, *DEFB1* was found not to vary significantly across primates [11].

Here we describe a combined strategy to identify further β-defensin genes in the draft human genome sequence using computational techniques and verification using reverse transcription-PCR. Four full-length, novel genes (*DEFB105*, *DEFB106*, *DEFB107*, *DEFB108*) and a related pseudogene *DEFB109p* are reported, as well as their expression patterns and evidence for their evolution by duplication and positive selection.

Results

All TBLASTX [12] matches to human bacterial artificial chromosome (BAC) clone sequences were in the 8p22-p23 region in a subsection of FPC contig ctg45 (1 April freeze WashU Accession Map Layout Files [13]) bounded by the BAC clones RP11-161B1 (AC079018) and SCb-177K12 (AF252831) and consisting of 53 BACs in total. These 53 BACs were deemed to represent the human β-defensin gene family locus and were masked for repetitive sequences using RepeatMasker [14]. This locus was the subject of a more sensitive search for the presence of novel human β-defensins, using a hidden Markov model constructed from an alignment of the GenBank β-defensin sequences mentioned above. As well as the known, full-length human β-defensins and the related epididymis-specific *SPAG11* (formerly *EP2*) gene [15], two novel β-defensin genes, *DEFB105* and *DEFB106*, were identified in this search and were then incorporated into the previous hidden Markov model. Further searches with this revised model identified a further three genes: *DEFB107*, *DEFB108* and *DEFB109p*. None of these five genes was found in the EMBL sequence database (24 June 2002 release) or in the Ensembl genomic annotation database (version 6.28.1 [16]). The novel gene *DEFB109p* appears to be a pseudogene as it contains a premature stop codon within its first exon, as observed in three independently sequenced, overlapping 8p22-p23 BAC sequences (accession numbers AC068974, AC087203 and AF252830). Given the absence of premature stop codons in the other four genes, despite considerable divergence among them (they encode only 18-28% identical amino-acid residues), it is unlikely that they too are pseudogenes.

Many putative final exon fragments from novel mouse β-defensins were recently reported by Schutte *et al.* [4]. However, their data were incomplete and were not supported by experimental verification. They used computational techniques to identify sequences matching a central portion of the mature defensin peptide including the six-cysteine motif characteristic of β-defensins. At best, this method could only identify incomplete final exons encoding this region of the peptide. No attempt was made to delineate precisely the boundaries of final exons or to identify first exons by these authors. In the present study we restrict our attention to complete genes, present in a BAC-clone-based map of the region and verified as encoding real transcripts by RT-PCR. Our full-length novel genes correspond to five final exon fragments (*DEFB5*, *DEFB6*, *DEFB7*, *DEFB8* and *DEFB9*) reported by Schutte *et al.* [4] and we have adopted the official HUGO Human Gene Nomenclature Committee [17] names for these fragments: *DEFB105*, *DEFB106*, *DEFB107*, *DEFB108* and *DEFB109p* (amended from *DEFB109*) respectively.

RT-PCR amplification confirmed the presence of the computationally predicted, functional genes, but sequencing of these products and subsequent alignment to genomic

sequence also uncovered several instances where the actual gene structures differed from those that were predicted. In two cases, *DEFB106* and *DEFB107*, splice sites other than those predicted were used and it was found that the predicted coding sequences had been respectively 36 base-pairs (bp) longer and 9 bp shorter than the actual coding sequences. In the case of *DEFB105* RT-PCR verified the two predicted exons but also uncovered an additional, 42-bp intervening exon. The RT-PCR analysis also produced interesting results with respect to the coding sequence of *DEFB108*. In four independent, sequenced RT-PCR products from one mRNA sample, the amplified *DEFB108* sequence was consistently found to differ from the computationally predicted sequence from the human public draft sequence at three nucleotide sites, but was otherwise identical. The three differences observed between predicted and amplified sequences respectively were an A to G at nucleotide 62 (which causes a conservative amino-acid change from lysine to arginine), a T to C at nucleotide 111 (which is a synonymous change) and a C to T at nucleotide 120 (also synonymous). This indicates novel human polymorphisms in *DEFB108* and is consistent with observations of high degrees of polymorphism for other human β-defensins [18] (see also Additional data files 1-6 for the novel gene sequences confirmed by RT-PCR and the computationally predicted sequence of *DEFB109p*).

Figure 1a depicts the relative positions and orientations of the four novel genes, the novel pseudogene *DEFB109p* and four known genes in the vicinity: *DEFB2-4* and the related epididymis-specific *SPAG11*. This cluster is about 350 kilobases (kb) centromeric of the *DEFB1* gene. A phylogenetic tree relating functional human β-defensins (Figure 1b) reflects the spatial distribution of these genes, with the cluster of genes in Figure 1a appearing to form a clade separate from *DEFB1*. The simplest explanation for the origin of this cluster is a series of local duplication events followed by substantial divergence. In addition it seems that the four novel functional genes are more closely related to *DEFB104* than to *DEFB4* or *DEFB103*. Three of these genes, *DEFB106*, *DEFB105* and *DEFB108*, encode mature peptides exhibiting the same spacing of conserved cysteine residues as seen in *DEFB104*, although in *DEFB105* there is an extra cysteine residue (amino acid 43) towards the amino-terminal end of the mature peptide. *DEFB105* also encodes an unusually long propiece peptide region in the second of its three exons. Strikingly, the DEFB107 protein lacks the first canonical β-defensin cysteine altogether and instead has a serine residue at the same position (Figure 2). The changes in the number of cysteine residues seen in DEFB105 and DEFB107 are likely to have important functional consequences. The predicted mature peptides for the four novel, functional β-defensins presented here have a similar proportion of cationic residues to human DEFB104, with higher proportions of anionic residues (10-13%) than are seen in human DEFB1, DEFB4 and DEFB103 (less than 4%). Indeed the pI of DEFB107 and DEFB108 are 6.74 and 6.89 respectively, whereas all other β-defensins described to date are cationic. This relative increase in anionic residues is expected to affect function, as the action of defensins initially involves interactions between the cationic mature defensin peptides and anionic membrane lipids [19]. Expression analysis for the novel, functional human genes was carried out by RT-PCR on a panel of human RNA samples and the novel gene PCR products were confirmed by hybridization to an internal probe (Figure 3). Expression of all four novel genes was readily detected in testis. A longer exposure period revealed low levels of expression of *DEFB108* in the liver (Figure 3, left-hand panel of *DEFB108*). Expression was not detected in any of the other tissues analyzed.

Six genes highly similar (85-98% identical at the amino-acid level) to three of the novel human β-defensins (*DEFB105*, *DEFB106* and *DEFB107*) as well as to *DEFB4*, *DEFB103* and *DEFB104*, were found within two olive baboon (*Papio cynocephalus anubis*) draft genomic sequences (GenBank accession numbers AC116558 and AC116559) using BLAST. Full-length sequences were obtained for each baboon gene except the putative *DEFB4* ortholog (which lacks a first exon because of gaps in the draft genomic sequences); in all other cases the exonic structure of the putative baboon ortholog was identical to that of the human gene (Figure 2). It is very likely that these sequences originate from the baboon locus orthologous to the human region under study, but without more complete sequence or mapping data for the baboon genome it is impossible to be certain. These novel baboon gene sequences (see Additional data files for the accession numbers), together with the published sequence for olive baboon β-defensin 1 (AAK61474) and the full-length human β-defensins formed the basis for our evolutionary analyses.

Figure 4 shows d_N (number of nonsynonymous substitutions per nonsynonymous site) plotted against d_S (number of synonymous substitutions per synonymous site) for comparisons between the first exon (which encodes the signal peptide) and second exon (which encodes the mature defensin) from all human and baboon genes (the full d_N, d_S and d_N - d_S estimates for all first and second exon comparisons are available in Additional data file 7). Two major trends are observable. In the vast majority of first exon comparisons d_S exceeds d_N (Figure 4a) and this excess is statistically significant in almost every case according to two-tailed Z-test results for all human and baboon genes, but there were no significant excesses of d_S according to the more rigorous Fisher's exact test (data not shown). Thus the rates of substitution in first exons indicate that they are evolving approximately neutrally, perhaps under weak purifying selection. The pattern seen in the second exon comparisons is quite different (Figure 4b). In the second exons d_N often exceeds d_S, and even in these short sequences, for certain comparisons this excess reaches statistical significance. Significant excesses of d_N over d_S are seen between *DEFB1* and

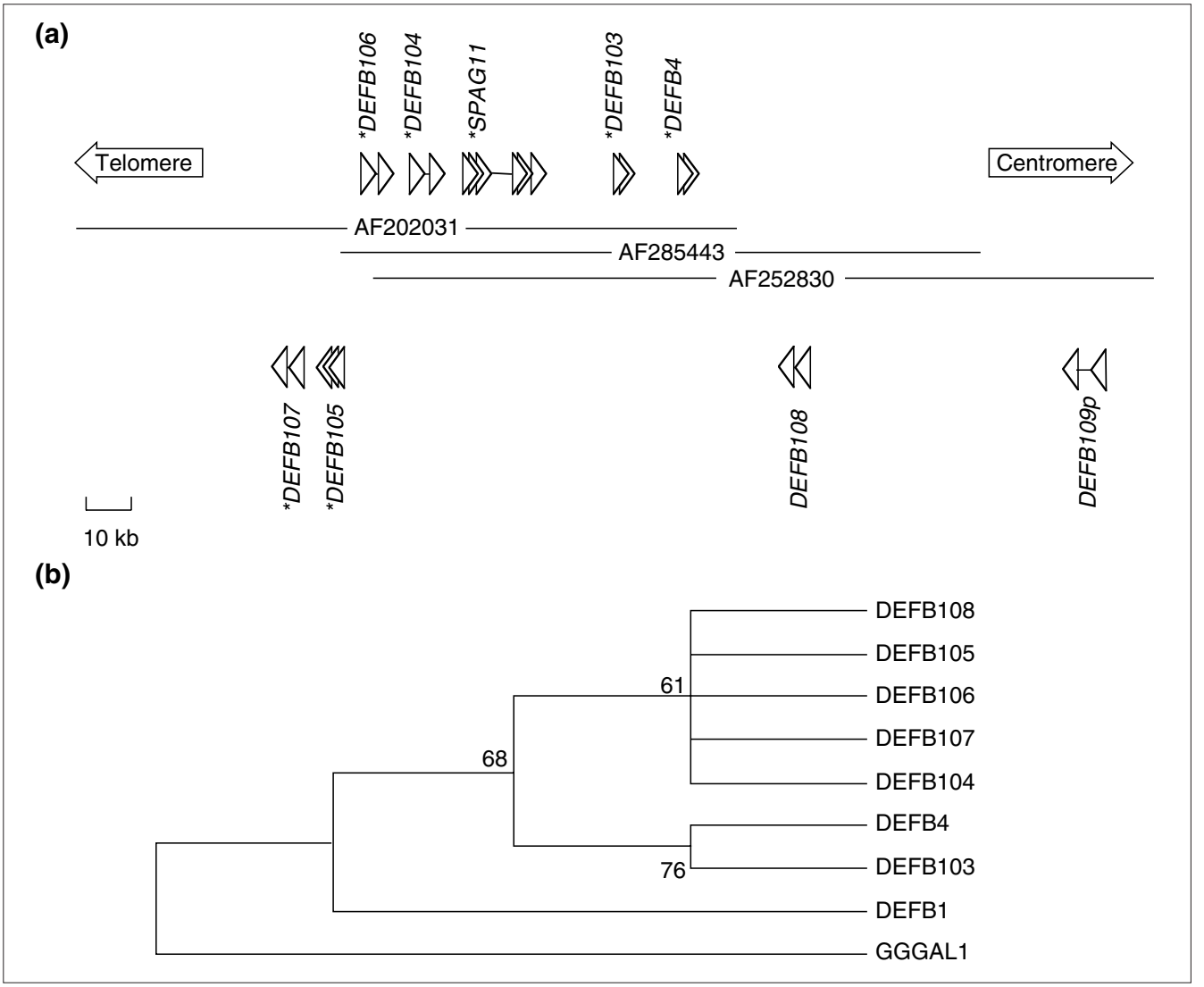

Figure 1

Genomic organization of novel human β-defensin genes. **(a)** The genomic organization of novel human β-defensin genes *DEFB105*, *DEFB106*, *DEFB107*, *DEFB108* and *DEFB109p* on 8p22-p23. The horizontal lines represent the three BAC clones in which all novel genes were found. Exons are represented as triangles with the vertical side representing the position of the exon. Exons above the horizontal lines are transcribed from the strand represented by the original BAC clone sequence entries whereas those below are transcribed in the opposite direction from the complementary strand. The region depicted is about 350 kb centromeric of *DEFB1*. Those genes marked with an asterisk were found to have orthologs in baboon genomic sequences (AC116558 and AC116559). **(b)** A phylogenetic tree of functional human β-defensins using the prepropeptide sequences encoded by the genes shown in (a). The phylogenetic tree was rooted with chicken gallinacin 1 (GGGAL1; P46156) and the reliability of each branch was assessed using 1,000 bootstrap replications.

DEFB104 and between *DEFB103* and *DEFB107*, with similar effects seen among the second exons of orthologous baboon genes (Table 1) using the method of Nei and Gojobori [20].

Moreover, in these comparisons d_S tends to be rather low relative to the rest of the data set (mean d_S = 0.464). Similar results are obtained using this method either modified to take account of the transition-to-transversion ratio R [21] or using the Jukes-Cantor correction, although the unmodified method is thought to be a more reliable basis for the detection of positive selection [22]. If we assume that synonymous substitutions (which are selectively neutral or nearly so) have accumulated regularly with time, such a pattern of substitution suggests that duplication was followed by rapid nonsynonymous change that subsequently decelerated. Most comparisons involving the second exons of *DEFB103*, *DEFB104* and *DEFB107* with the second exons of other genes in the dataset show excesses of d_N over d_S, but these excesses fail to reach significance by either of the tests used (data not shown). Although some of

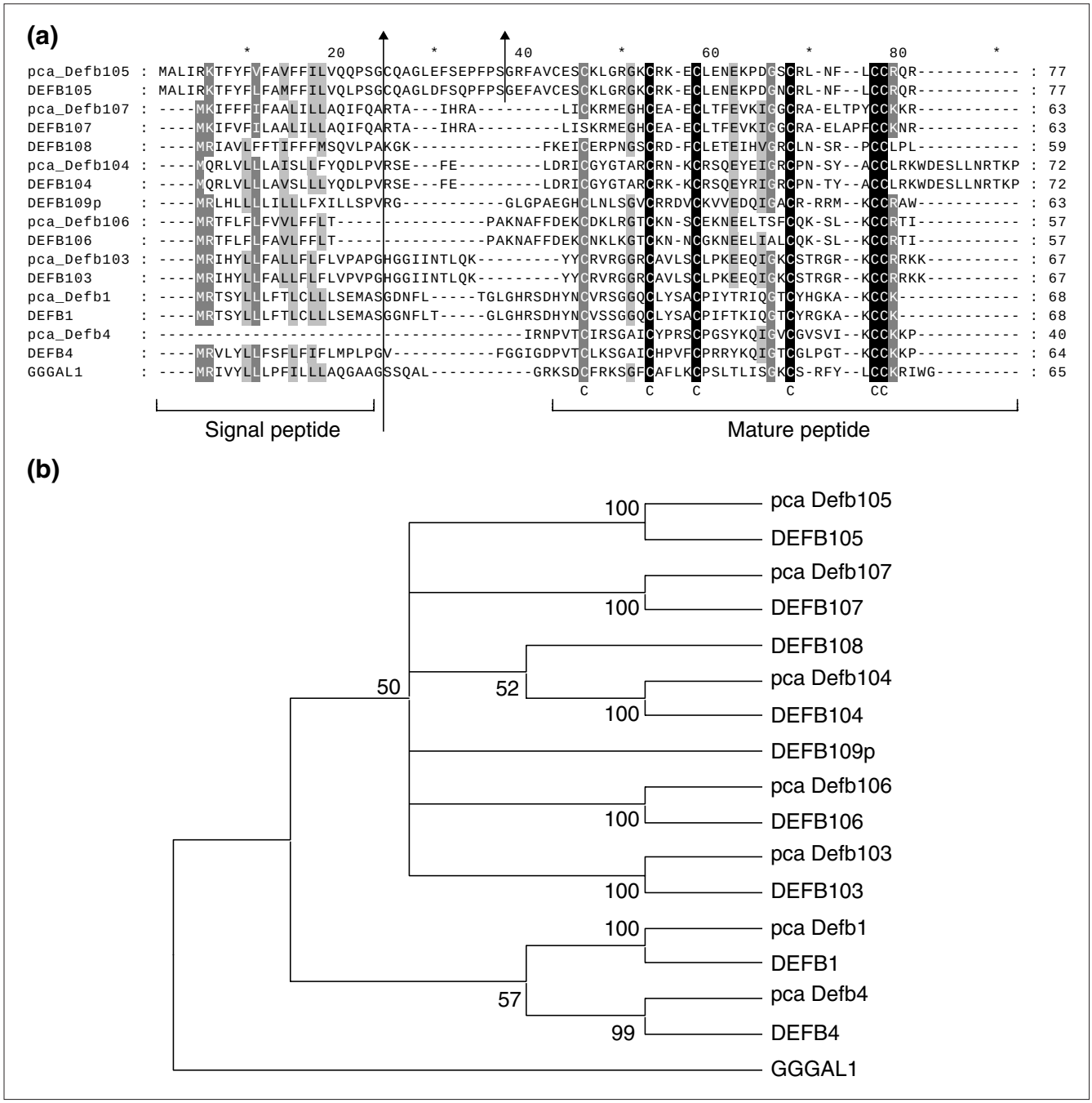

Figure 2
(a) Alignment and **(b)** phylogenetic tree of human and baboon β-defensin protein sequences. The tree was rooted with chicken gallinacin 1 (GGGAL1; P46156) and the reliability of each branch was assessed using 1,000 bootstrap replications. The alignment shows the same sequences with the estimated locations of the signal peptide and mature peptide regions; the intervening region is the propiece. The long arrow indicates the position of first introns: in each case except *DEFB105* the intron splits the codon that encodes the residue immediately before the arrow. The short arrow indicates the second intron, found only in *DEFB105*. The shading represents the degree of conservation at each position in the alignment, taking into account similar physicochemical properties of residues. The six canonical cysteines are indicated under the appropriate alignment positions. X at residue 14 denotes the location of the premature stop codon in *DEFB109p*.

the comparisons shown in Figure 4b indicate an excess of d_S over d_N, this is never significant by either of the two statistical tests used.

There is no detectable similarity between the introns of the genes under study except between the five putatively orthologous pairs of human and baboon genes where intronic

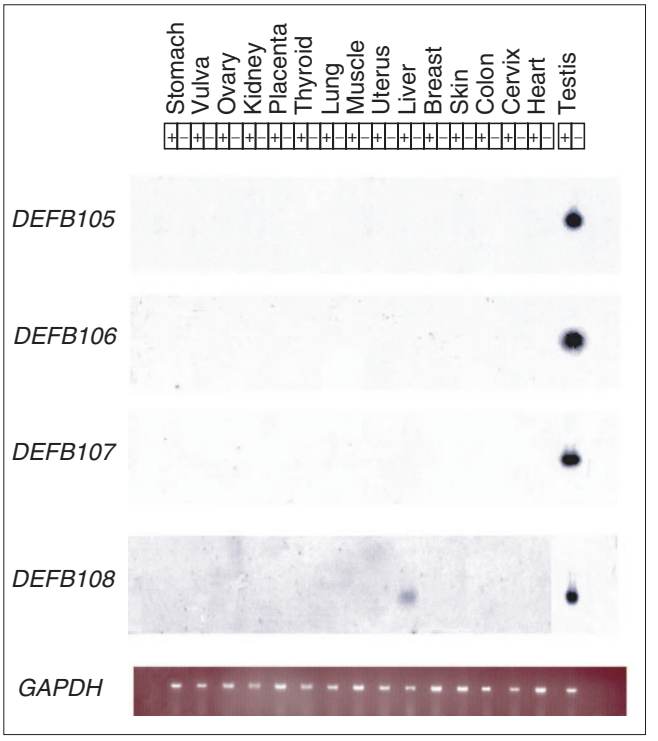

Figure 3
Expression patterns of novel human β-defensin genes. RT-PCR analysis of novel human gene expression carried out on a panel of human RNA samples. The tissues are indicated with a plus (+) and a minus (-) reverse transcriptase reaction shown for each sample. GAPDH RT-PCR was carried out as a control.

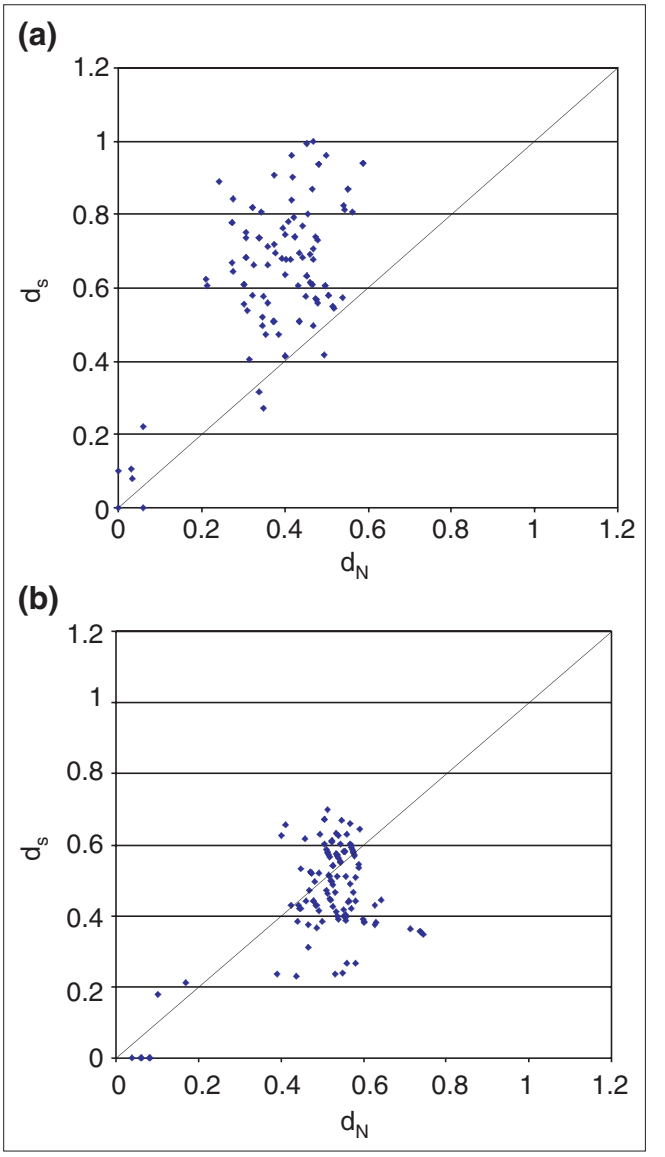

Figure 4
The rates of synonymous and nonsynonymous substitutions within first and second exons. Graphs of the number of synonymous substitutions per synonymous site (d_S) against the number of nonsynonymous substitutions per nonsynonymous site (d_N) for comparisons between **(a)** the first exons and **(b)** second exons of all genes in the dataset. In each case the diagonal line represents $d_N = d_S$.

sequence is available (*DEFB103*, *DEFB104*, *DEFB105*, *DEFB106* and *DEFB107*): each pair shares 74-91% identity over 80-99% of their lengths in spite of many small indel events. In every case these pairs of orthologous introns show a substitution rate that is not significantly different from the value of d_S obtained for the coding sequences. In particular, the orthologous intron sequence comparisons for *DEFB103*, *DEFB104* and *DEFB107* give substitution-rate estimates of 0.106±0.010, 0.081±0.005 and 0.066±0.004 respectively, which are consistent with the d_S estimates for the coding sequences of these pairs: 0.166±0.077, 0.085±0.058 and 0.053±0.049. Thus, the excess of nonsynonymous substitutions observed for the second exons of these genes is not attributable to artificially low d_S estimates as a result of sampling error, but is caused by a real increase in d_N relative to d_S, which is the pattern expected to be generated by positive selection.

Averages for the ratio of radical to conservative amino-acid changes, p_R/p_C calculated over 47 mammalian genes were reported as 0.81 and 0.49 for charge and the Miyata-Yasunaga (polarity and volume) classification respectively [23]. The equivalent values from Table 1, for comparisons between the second exons of genes showing evidence of positive selection, are all greater than these averages. Furthermore Table 1 shows that there has been a higher rate of change with respect to charge than with respect to polarity and volume (Miyata-Yasunaga amino-acid classification). That is, where there is evidence of positive selection, most nonsynonymous changes have tended to change the charges of the residues encoded but have tended to conserve the polarities and volumes of those residues.

Likelihood ratio tests (LRTs), as implemented by the PAML package [24] also indicate the operation of positive selection

Table 1

Results of comparisons between the second exons of human and baboon genes demonstrating positive selection

	DEFB1 vs DEFB104		*DEFB103 vs DEFB107*	
	Homo sapiens	*P. cynocephalus anubis*	*H. sapiens*	*P. cynocephalus anubis*
S*	17.917±1.180	17.167± 1.169	16.917±1.165	16.833±1.124
N	48.083±1.144	48.833±1.181	49.083±1.156	49.167±1.124
s	6.25±1.954	6.25±1.954	4.5±1.796	4±1.767
n	35.75±3.482	34.75±3.559	27.5±3.382	27±3.373
d_S	0.349±0.103	0.364±0.106	0.266±0.104	0.238±0.105
d_N	0.744±0.062	0.712±0.063	0.560±0.064	0.549±0.067
Z-test†	0.001	0.005	0.020	0.008
Fisher's‡	0.012	0.019	0.060	0.023
Charge§				
p_C	0.498±0.088	0.535±0.102	0.447±0.094	0.409±0.094
p_R	0.784±0.077	0.830±0.086	0.571±0.087	0.575±0.085
p_R/p_C	1.57¶	1.55¶	1.28	1.41
M-Y#				
p_C	0.577±0.116	0.708±0.108	0.598±0.124	0.581±0.124
p_R	0.618±0.092	0.634±0.100	0.467±0.080	0.446±0.082
p_R/p_C	1.0	0.90	0.78	0.77

*Estimates (±SE) of the number of synonymous sites (S), number of nonsynonymous sites, numbers of synonymous substitutions (s), numbers of nonsynonymous substitutions (n), the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N). †The result of a two-tailed Z-test of $d_N - d_S = 0$. ‡The result of a Fisher's exact test. §Rates of radical (p_R) and conservative (p_C) changes in amino-acid properties, with the ratio of radical to conservative changes (p_R/p_C) for residues categorized in terms of their charges. ¶p_R is significantly greater than p_C. #Rates of radical (p_R) and conservative (p_C) changes in amino-acid properties, with the ratio of radical to conservative changes (p_R/p_C) for residues categorized in terms of the Miyata-Yasunaga classification (M-Y; a combination of polarity and volume).

at sites within second exons. These tests indicate whether data (the substitutions inferred from an alignment) are best explained by one of two models of $\omega = d_N/d_S$. Since ω is a measure of selective pressure on proteins, these models can be used to assess the evidence for variable selective pressures among sites. In a test for positive selection (the presence of sites at which $\omega > 1$) two statistical distributions are compared: a null model that uses a distribution that does not allow for sites with $\omega > 1$ and another model which does allow such sites. Three pairs of site-specific likelihood models were compared that assume variable selective pressure (as determined by the value of ω) among sites but no variation among sequences in the dataset: M0 (one-ratio) and M3 (discrete), M1 (neutral) and M2 (selection), and M7 (beta) and M8 (beta+ω) [25]. The second exon of the pseudogene *DEFB109p* was omitted from the analysis. The discrete model (M3) with two site classes suggested that 41% of second-exon sites are under positive selection with ω = 1.67 and identified nine amino-acid sites under positive selection at the 95% cutoff. M3 was a significantly better fit to the data than the one-ratio model M0; the LRT statistic is $2\Delta l = -1061.79-(-1004.12) = 115.32$, and $p < 0.001$ with two degrees of freedom. However the M0-M3 comparison is essentially a test of variability in the ω ratio among sites and does not constitute a rigorous test of positive selection. Model M1 (neutral) assumes two site classes with $\omega = 0$ and $\omega_1 = 1$ fixed and with the proportions p and p_1 estimated. Model M2 (selection) adds a third site class with the ratio ω_2 estimated, it suggests that about 47% of sites are under positive selection with $\omega_2 = 53.17$ and identified 12 amino-acid sites under positive selection at the 95% cutoff. The two models can be compared using an LRT as follows, $2\Delta l = -1030.51-(-1022.03) = 16.97$; $p < 0.001$ with 2 df. So model M2 is significantly better than M1. Model M7 (beta) assumes a beta distribution for ω over sites. The beta distribution is limited to values between 0 and 1, providing the most flexible null hypothesis for testing positive selection. Model M8 (beta+ω) adds another site class to M7 (beta), with the ω ratio estimated from the data. However, the difference between M7 and M8 is not statistically significant, as indicated by the LRT: $2\Delta l = -1006.44-(-1006.44) = 0$. Nine particular sites were implicated (in both M2 and M3 models) as being under positive selection with greater than 95% confidence: positions 43, 44, 48, 52, 56, 57, 69, 70 and 73 in Figure 2a. All these positions are close or adjacent to a conserved cysteine residue and so it is possible they are important

in determining β-defensin structure. In summary, the PAML analysis indicates that ω varies significantly between sites, and in two separate LRTs the parameters estimated suggest a substantial proportion of sites are under positive selection. In spite of this, the most stringent test (M7 versus M8) does not indicate a significant difference from neutrality. This result may be attributable to the extremely short lengths (33 codons aligned omitting positions with gaps) of the sequences aligned, such an effect was seen when analyzing short (130 codons) lysozyme sequences in the same way [25].

Discussion

This study represents the most detailed study of a human β-defensin cluster to date, including the full-length sequences of four novel genes (*DEFB105*, *DEFB106*, *DEFB107*, *DEFB108*), and a novel pseudogene (*DEFB109p*), their expression patterns and sequences for the baboon orthologs of six genes from this cluster. The 8p22-23 defensin locus appears to have evolved by successive rounds of duplication followed by substantial divergence, to produce a diverse cluster of paralogous genes defined by these four novel genes and four known β-defensin genes (*DEFB4*, *DEFB103* and *DEFB104*). Divergence has been most rapid within the second exons of these genes, which encode the mature β-defensin peptide, with many comparisons between paralogous genes showing an excess of nonsynonymous over synonymous substitutions. Statistically significant evidence of elevated nonsynonymous change is seen by two methods in the second exons, indicating the action of positive selection. By contrast, comparisons between the first exons of genes from this cluster, which encode a signal peptide, show an excess of synonymous substitutions consistent with neutral evolution or weak purifying selection. The duplication and subsequent positive selection of these genes predates human-baboon divergence more than 23 million years ago and is consistent with observations that *DEFB1* has undergone very little change during the evolution of primates [11]. The positive selection observed has tended to change the charges of residues encoded more than other qualities such as residue polarity or volume. As seems to be the case with other antimicrobial peptides, such as MHC receptors, immunoglobulins and α-defensins [6-8], this selection may be a response to the rapid evolution of pathogens.

In this study the computationally predicted gene structures were found by laboratory work to deviate from the actual structures in three out of five novel genes. These errors in the predictions arose in spite of the fact that the predictions were based on all the β-defensin protein-sequence data available and involved three completed BAC sequences rather than unfinished, gapped sequence. This has implications for purely computational approaches to novel gene discovery such as that of Schutte *et al*. [4].

DEFB104 and *DEFB108* have no detectable orthologs in the mouse genome and therefore appear to have arisen by duplication since the divergence of rodents and primates; alternatively, there could have been a loss of these genes in the rodent lineage. They are also the best candidates for primate-specific β-defensins as they lack orthologs within all other known mammalian defensins, although our knowledge of mammalian defensins is currently incomplete. As we have shown, the evolution of the *DEFB104* mature peptide has been driven by positive selection since its emergence, which is consistent with its novel antimicrobial properties [26]. All defensins were thought to exist as monomers stabilized by three disulfide bridges between their three pairs of conserved cysteines [19]. However, when compared with other known human β-defensins (*DEFB1*, *DEFB4* and *DEFB103*), *DEFB104* was found to have a different number of residues between its second and third, and between its fourth and fifth cysteine residues. Furthermore, *DEFB104* was found to have bactericidal activity against *Pseudomonas aeruginosa* that was more than sixfold stronger than for any other β-defensin [26]. Three of the novel genes described here (*DEFB105*, *DEFB106* and *DEFB108*) encode mature peptides exhibiting the same spacing of conserved cysteine residues as seen in *DEFB104* as well as sharing similar expression patterns, with highest expression in the testes. The functional divergence of β-defensin genes appears to have continued following human-baboon divergence, as exemplified by *DEFB107* which displays a serine residue instead of the first canonical cysteine seen in the baboon ortholog. It is notable that a novel mouse β-defensin gene (*Defr1*), which also lacks the first canonical cysteine, has potent antimicrobial activity against a spectrum of pathogens [27]. In addition a polymorphism in the *DEFB1* gene which alters the first canonical cysteine to a serine residue has been shown to produce a peptide which is as active against the microorganisms tested as the usual form [28]. The unusual amino-acid composition of the proteins encoded by the novel genes presented here suggests that they may possess novel functions, indeed the cationic nature of β-defensins is lost in DEFB107 and DEFB108. As shown recently, there may be more subtle consequences of variation in β-defensin protein sequences, affecting dimerization as well as net charge and disulfide bridges [29]. It is worth noting that recent research has identified additional functions for β-defensins that link the innate and adaptive immune response. Both human and mouse β-defensins have been shown to be chemotactic for immature dendritic cells and memory T cells via the CCR6 chemokine receptor [30]. The mouse β-defensin *Defb2* has been shown to act directly on immature dendritic cells as an endogenous ligand for Toll-like receptor 4 (TLR-4), inducing upregulation of co-stimulatory molecules and dendritic cell maturation. These events, in turn, trigger robust, type-1 polarized adaptive immune responses *in vivo*, which suggests that β-defensins may have an important role in immunosurveillance against pathogens [31].

The expression of the novel antimicrobial peptides reported here in the human male reproductive tract is also of interest.

Recent work has shown that the male urogenital tracts in mammals express a broad range of α- and β-defensins, with human *DEFB1* expressed in testicular biopsies, seminal plasma and ejaculated spermatozoa [32]. Together with our results demonstrating that *DEFB105*, *DEFB106*, *DEFB107* and *DEFB108* are predominantly expressed in the testes, these data suggest that the male reproductive tract has a complex innate defense mechanism.

Conclusions

The 8p22-23 β-defensin locus has evolved by duplication and subsequent divergence, to produce a diverse cluster of paralogous genes defined by four novel genes (*DEFB105*, *DEFB106*, *DEFB107*, *DEFB108*), a novel pseudogene (*DEFB109p*), and three known genes (*DEFB4*, *DEFB103* and *DEFB104*). We present full-length sequences for the four novel genes, their expression patterns and the predicted sequences for the baboon orthologs of six genes from this cluster. Although comparisons among first-exon sequences of these human genes show little variation, the second-exon sequences that encode the mature β-defensin peptides show substantial divergence. Evolutionary analyses suggest that the divergence seen in second exons has involved positive selection disproportionately favoring alterations in the charge of amino-acid residues.

Materials and methods

Identification of novel genes

The following mammalian β-defensin sequences were retrieved from GenBank. Mouse β-defensins: 1-11 (NP_031869, NP_034160, NP_038784, NP_062702, NP_109659, NP_473415, NP_631966, CAC44635, NP_631965, CAD26894, CAD26895), β-defensin 13 (NP_631969), β-defensin 15 (NP_631970), β-defensin 35 (NP_631970), defensin-related peptide (AJ344114). Human: *DEFB1* (Q09753), *DEFB4* (O15263), *DEFB103* (NP_061131), *DEFB104* (CAC85520). Rat: *Rattus norvegicus* β-defensin 1 (NP_113998), β-defensin 2 (O88514). Cow: *Bos taurus* β-defensin 1 (P46159), β-defensin 2 (P46160), β-defensin 3 (P46161), β-defensin 4 (P46162), β-defensin 5 (P46163), β-defensin 6 (P46164), β-defensin 7 (P46165), β-defensin 8 (P46166), β-defensin 9 (P46167), β-defensin 10 (P46168), β-defensin 11 (P46169), β-defensin 12 (P46170), β-defensin 13 (P46171) , tracheal antimicrobial peptide (P25068), lingual antimicrobial peptide (Q28880), enteric β-defensin (O02775), β-defensin C7 (O18815). Pig: *Sus scrofa* β-defensin 1 (O62697). Goat: *Capra hircus* β-defensin 1 (O97946), β-defensin 2 (CAA08905). Sheep: *Ovis aries* β-defensin 1 (O19038), β-defensin 2 (O19039). Rhesus monkey: *Macaca mulatta* β-defensin 1 (O18794), β-defensin 2 (AAK26259). Olive baboon: *Pabio cynocephalus anubis* β-defensin 1 (AAK61474). Chimpanzee: *Pan troglodytes* β-defensin 1 (AAF04110), β-defensin 2 (AAF20154), β-defensin 3 (AAK61549). These sequences were used as TBLASTX (version 2.1.1 with default settings [12]) queries against the HTG (high-throughput genomic) section of the EMBL database (15 July 2001 release). Hidden Markov models were constructed using HMMER (version 2.1.1 [33]) to process CLUSTALW (version 1.82 with default settings [34]) multiple sequence alignments. These models were searched against genomic sequence using WISE2 (version 2-1-20c with the human gene model option [35]). CLUSTALW alignments of diverse second-exon sequences were corrected using the patterns of gaps seen in the corresponding protein-sequence alignments that were more highly conserved.

Evolutionary analyses

All phylogenetic trees were constructed by the neighbour-joining method [36] based on the proportion of amino-acid sites at which sequences compared were different and omitting alignment gaps. The trees constructed were rooted with chicken gallinacin 1 (*GGGAL1*; P46156) and the reliability of each branch was assessed using 1,000 bootstrap replications. In pairwise comparisons between nucleotide sequences, the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) were estimated using the method of Nei and Gojobori [20], modified to take account of the transition-to-transversion ratio R [21]. R was estimated using the method of Kumar and Nei [22]. In addition, the Jukes-Cantor correction [37] was applied to account for multiple substitutions at the same site. Two codon-based tests of selection were used. Both tests are based on estimates of d_S and d_N. Standard errors for d_S and d_N were calculated using 1,000 bootstrap replicates. In the first test d_S and d_N and their respective variances are used in a two-tailed Z-test to test the null hypothesis that $d_N - d_S = 0$ [21]. In the second test, Fisher's exact test is used to test the null hypothesis that the proportions of synonymous and non-synonymous differences are the same [38]. Additional tests for the presence of sites under positive selection were carried out using the PAML package [24], which uses likelihood ratio tests (LRT) to compare models of the variation in d_N/d_S ratio between sites. The six models recommended by Anisimova *et al.* [39] were tested: M0 (one-ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (beta), and M8 (beta+ω). Intron sequences were aligned using DIALIGN (version 2.1 [40]) and the numbers of substitutions between them were estimated using Kimura's two-parameter method [41]. All phylogenetic trees, distance calculations and codon-based tests of selection were carried out using MEGA2 [42]. Estimates of the proportions of radical and conservative nonsynonymous substitutions, along with their standard errors, were made using the HON-NEW program [23] in an extension of earlier methods for the measurement of conservative and radical substitution rates [43]. The radical or conservative nature of nonsynonymous substitutions was assessed with respect to charge and to the polarity and volume of the amino acids (the Miyata-Yasunaga amino-acid classification [44]).

RT-PCR analysis of gene expression

A range of human RNA samples was purchased from Stratagene (Amsterdam, The Netherlands) (stomach, vulva, ovary, kidney, placenta, thyroid, lung, skeletal, uterus, breast, liver, skin, colon, heart and cervix) and human testis RNA (BD Biosciences Clontech, Oxford). cDNA synthesis was carried out using a first-strand cDNA synthesis kit (Roche, Lewes, UK) according to the instructions, using random hexameric oligonucleotides. PCRs were carried out, using 5 μl of the resultant cDNA according to the following procedures: 94°C for 1 min followed by 35 cycles of 94°C for 30 sec, 55°C for 30 sec and 72°C for 1 min, and a final round of extension for 5 min. Products were analyzed on a 4% NuSIEVE agarose gel (FMC BioProducts, Rockland, ME, USA) by electrophoresis and also cloned using pGEM-T Easy Vector System I (Promega, Southampton). Amplification of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was carried out in parallel with conditions as for the other amplifications, but with an annealing temperature of 56°C. Reactions were verified for RNA amplification by including controls without reverse transcriptase. RT-PCR products were hybridized with a radiolabeled, internal oligonucleotide probe designed to each novel sequence to confirm the presence of the correctly amplified product. Purified plasmid DNA was sequenced from both strands with ABI Prism dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems, Warrington). Primers and internal probes were designed using the Primer3 primer design program from the Whitehead Institute, Center for Genome Research [45]. The sequences for the primers and internal probes used were as follows: *DEFB105*: 5′ primer: TCTATTTGCTATGTTCTTCATTTTGG, internal oligo: TTCAACTGCCATCAGGTGAG, 3′ primer: GCAGCAGAGAAAGTTCAGCC; *DEFB106*: 5′ primer: CGTGCTCTTCTTTCTGACCC, internal oligo: TACAGGGAAGGTGATCGGAG, 3′ primer: GTTCTTCATTTTTCCCGCAA; *DEFB107*: 5′ primer: TTTTGGCTGCTCTCATTCTTC, internal oligo: TCACTGTGAAGCCGAATGTC, 3′ primer: TGCAGCAAAATGGTGCTAAT; *DEFB108*: 5′ primer: TGCTGTCCTCTTCTTCACCA, internal oligo: GCCAAGTTCTACCAGCCAAG, 3′ primer: CGGCTATTTAAACATCTCCCA.

The novel human gene sequences for *DEFB105*, *DEFB106*, *DEFB107*, *DEFB108* and *DEFB109p*, as confirmed by RT-PCR, have been deposited in GenBank under sequence accession numbers AF540977, AF540978, AF540979, AF540980 and AF540981 respectively. The putative baboon orthologs of the human *DEFB4* (second exon only), *DEFB103*, *DEFB104*, *DEFB105*, *DEFB106* and *DEFB107* genes have been also been deposited in GenBank under sequence accession numbers BK000556, BK000557, BK000558, BK000559, BK000560 and BK000561 respectively.

Additional data files

The following files are available with this article: the DNA sequences for olive baboon β-defensins (Additional data file 1), the DNA sequences for human β-defensins (Additional data file 2), the aligned baboon and human DNA sequences (Additional data file 3), the aligned first-exon baboon and human DNA sequences (Additional data file 4), the aligned second-exon baboon and human DNA sequences (Additional data file 5) and the DNA sequences for baboon and human introns (Additional data file 6) and an Excel file listing the full d_N, d_S and d_N - d_S estimates for all first and second exon comparisons shown in Figures 4a and b (Additional data file 7). The alignments are in MSF format. The sequence and alignment files are also available from [46].

Acknowledgements

This work benefited from the financial support of the UK Medical Research Council. We would also like to thank Gillian Morrison for her initial involvement in this project.

References

1. Goldman MJ, Anderson GM, Stolzenberg ED, Kari UP, Zasloff M, Wilson JM: **Human beta-defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis.** *Cell* 1997, **88:**553-560.
2. Schroder JM, Harder J: **Human β-defensin-2.** *Int J Biochem Cell Biol* 1999, **31:**645-651.
3. Lehrer RI, Ganz T: **Defensins of vertebrate animals.** *Curr Opin Immunol* 2002, **14:**96-102.
4. Schutte BC, Mitros JP, Bartlett JA, Walters JD, Jia HP, Welsh MJ, Casavant TL, McCray PB Jr: **Discovery of five conserved β-defensin gene clusters using a computational search strategy.** *Proc Natl Acad Sci USA* 2002, **99:**2129-2133.
5. O'Neil DA, Porter EM, Elewaut D, Anderson GM, Eckmann L, Ganz T, Kagnoff MF: **Expression and regulation of the human beta-defensins *DEFB1* and *DEFB4* in intestinal epithelium.** *J Immunol* 1999, **163:**6718-6724.
6. Ota T, Sitnikova T, Nei M: **Evolution of vertebrate immunoglobulin variable gene segments.** *Curr Top Microbiol Immunol* 2000, **248:**221-245.
7. Hughes AL, Yeager M: **Natural selection at major histocompatibility complex loci of vertebrates.** *Annu Rev Genet* 1998, **32:**415-435.
8. Hughes AL, Yeager M: **Coordinated amino acid changes in the evolution of mammalian defensins.** *J Mol Evol* 1997, **44:**675-682.
9. Hughes AL: **Evolutionary diversification of the mammalian defensins.** *Cell Mol Life Sci* 1999, **56:**94-103.
10. Duda TF Jr, Vanhoye D, Nicolas P: **Roles of diversifying selection and coordinated evolution in the evolution of amphibian antimicrobial peptides.** *Mol Biol Evol* 2002, **19:**858-864.
11. Del Pero M, Boniotto M, Zuccon D, Cervella P, Spano A, Amoroso A, Crovella S: **Beta-defensin 1 gene variability among non-human primates.** *Immunogenetics* 2002, **53:**907-913.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
13. **Genome Sequencing Center, Washington University in St. Louis** [http://genome.wustl.edu]
14. **RepeatMasker** [http://ftp.genome.washington.edu/RM/RepeatMasker.html]
15. Frohlich O, Po C, Murphy T, Young LG: **Multiple promoter and splicing mRNA variants of the epididymis-specific gene EP2.** *J Androl* 2000, **21:**421-430.
16. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.
17. **HUGO Human Gene Nomenclature Committee** [http://www.gene.ucl.ac.uk/nomenclature]
18. Salvatore F, Scudiero O, Castaldo G: **Genotype-phenotype correlation in cystic fibrosis: The role of modifier genes.** *Am J Med Genet* 2002, **111:**88-95.

19. White SH, Wimley WC, Selsted ME: **Structure, function, and membrane integration of defensins.** *Curr Opin Struct Biol* 1995, **5:**521-527.
20. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3:**418-426.
21. Zhang J, Rosenberg HF, Nei M: **Positive Darwinian selection after gene duplication in primate ribonuclease genes.** *Proc Natl Acad Sci USA* 1998, **95:**3708-3713.
22. Kumar S, Nei M: *Molecular Evolution and Phylogenetics*. New York: Oxford University Press; 2000.
23. Zhang J: **Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes.** *J Mol Evol* 2000, **50:**56-68.
24. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.
25. Yang Z, Nielsen R: **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.** *Mol Biol Evol* 2002, **19:**908-917.
26. Garcia JR, Krause A, Schulz S, Rodriguez-Jimenez FJ, Kluver E, Adermann K, Forssmann U, Frimpong-Boateng A, Bals R, Forssmann WG: **Human β-defensin 4: a novel inducible peptide with a specific salt-sensitive spectrum of antimicrobial activity.** *FASEB J* 2001, **15:**1819-1821.
27. Morrison G, Rofle M, Kilanowski F, Cross S, Dorin JR: **Identification and characterisation of a novel murine β-defensin related gene.** *Mamm Genome* 2002, **13:**445-451.
28. Circo B, Skerlavaj B, Gennaro R, Amoroso A, Zanetti M: **Structural and functional characterization of hBD-1(Ser35), a peptide deduced from a *DEFB1* polymorphism.** *Biochem Biophys Res Commun* 2002, **293:**586-592.
29. Schibli DJ, Hunter HN, Aseyev V, Starner TD, Wiencek JM, McCray PB Jr, Tack BF, Vogel HJ: **The solution structures of the human beta-defensins lead to a better understanding of the potent bactericidal activity of HBD3 against *Staphylococcus aureus*.** *J Biol Chem* 2002, **277:**8279-8289.
30. Yang D, Chertov O, Bykovskaia SN, Chen Q, Buffo MJ, Shogan J, Anderson M, Schroder JM, Wang JM, Howard OM, *et al.*: **Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6.** *Science* 1999, **286:**525-528.
31. Biragyn A, Ruffini PA, Leifer CA, Klyushnenkova E, Shakhov A, Chertov O, Shirakawa AK, Farber JM, Segal DM, Oppenheim JJ, *et al.*: **Toll-like receptor 4-dependent activation of dendritic cells by beta-defensin 2.** *Science* 2002, **298:**1025-1029.
32. Com E, Bourgeon F, Evrard B, Ganz T, Colleu D, Jegou B, Pineau C: **Expression of antimicrobial defensins in the male reproductive tract of rats, mice, and humans.** *Biol Reprod* 2003, **68:**95-104.
33. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
34. Higgins DG, Thompson JD, Gibson TJ: **Using CLUSTALW for multiple sequence alignments.** *Methods Enzymol* 1996, **266:**383-402.
35. **WISE2** [http://www.sanger.ac.uk/Software/Wise2]
36. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
37. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian Protein Metabolism*. Edited by Munro HN. New York: Academic Press; 1969: 21-132.
38. Zhang J, Kumar S, Nei M: **Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes.** *Mol Biol Evol* 1997, **14:**1335-1338.
39. Anisimova M, Bielawski JP, Yang Z: **Accuracy and power of bayes prediction of amino acid sites under positive selection.** *Mol Biol Evol* 2002, **19:**950-958.
40. Morgenstern B: **DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment.** *Bioinformatics* 1999, **15:**211-218.
41. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111-120.
42. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.
43. Hughes AL, Ota T, Nei M: **Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules.** *Mol Biol Evol* 1990, **7:**515-524.
44. Miyata T, Miyazawa S, Yasunaga T: **Two types of amino acid substitutions in protein evolution.** *J Mol Evol* 1979, **12:**219-236.
45. **Primer3** [http://www-genome.wi.mit.edu/genome_software/other/primer3.html]
46. **MRC HGU Semple lab: data/software** [http://www.hgu.mrc.ac.uk/Users/Colin.Semple/lab_data.html]