# Phylogenetic and evolutionary analysis of the *PLUNC* gene family

COLIN D. BINGLE,[1] ELIZABETH E. LECLAIR,[2] SUZANNE HAVARD,[1] LYNNE BINGLE,[1] PAUL GILLINGHAM,[3] AND C. JEREMY CRAVEN[3]

[1]Academic Unit of Respiratory Medicine, Division of Genomic Medicine, University of Sheffield Medical School, Sheffield, UK
[2]DePaul University, Department of Biological Sciences, Chicago, Illinois 60614, USA
[3]Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK

(RECEIVED July 25, 2003; FINAL REVISION September 12, 2003; ACCEPTED October 5, 2003)

**Abstract**

The PLUNC family of human proteins are candidate host defense proteins expressed in the upper airways. The family subdivides into short (SPLUNC) and long (LPLUNC) proteins, which contain domains predicted to be structurally similar to one or both of the domains of bactericidal/permeability-increasing protein (BPI), respectively. In this article we use analysis of the human, mouse, and rat genomes and other sequence data to examine the relationships between the PLUNC family proteins from humans and other species, and between these proteins and members of the BPI family. We show that PLUNC family clusters exist in the mouse and rat, with the most significant diversification in the locus occurring for the short PLUNC family proteins. Clear orthologous relationships are established for the majority of the proteins, and ambiguities are identified. Completion of the prediction of the LPLUNC4 proteins reveals that these proteins contain approximately a 150-residue insertion encoded by an additional exon. This insertion, which is predicted to be largely unstructured, replaces the structure homologous to the 40s hairpin of BPI. We show that the exon encoding this region is anomalously variable in size across the LPLUNC proteins, suggesting that this region is key to functional specificity. We further show that the mouse and human PLUNC family orthologs are evolving rapidly, which supports the hypothesis that these proteins are involved in host defense. Intriguingly, this rapid evolution between the human and mouse sequences is replaced by intense purifying selection in a large portion of the N-terminal domain of LPLUNC4. Our data provide a basis for future functional studies of this novel protein family.

**Keywords:** PLUNC; BPI; rapid evolution; innate immunity

**Supplemental material:** See www.proteinscience.org

Reprint requests to: Colin D. Bingle, Academic Unit of Respiratory Medicine, Division of Genomic Medicine, University of Sheffield Medical School, Sheffield, S10 2JF, UK; e-mail: c.d.bingle@sheffield.ac.uk; fax: 00-44 (0) 114-272-1104; or C. Jeremy Craven, Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, S10 2TN, UK; e-mail: c.j.craven@sheffield.ac.uk.; fax: 00-44 (0) 114-272-8697.

*Abbreviations:* at, *Arabidopsis thaliana*; b, bovine; c, chicken; cg, *Crassostrea gigas* (pacific oyster); ci, *Ciona intestinalis*; t, Trout; cael, *Caenorhabditis elegans*; ec, Equus caballus (horse); h, human; m, mouse; r, rat.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03332704.

Epithelial surfaces are patrolled by numerous molecules that protect against pathogenic colonization, and that constitute components of the innate immune system (Martin 2000). The innate immune response not only provides the first-line barrier against infection but also determines to which antigens the acquired immune system responds and the nature of these responses (Fearon and Locksley 1996). Many of the molecules associated with the innate immune response specifically interact with and respond to the bacterial surface. In humans, two of the proteins critical to the mediation of

 

signals from Gram-negative bacteria are lipopolysaccharide (LPS) binding protein (LBP) and bactericidal/permeability-increasing protein (BPI; Elsbach and Weiss 1998; Fenton and Golenbock 1998). LBP and BPI, which are structurally similar (Beamer et al. 1997), both bind LPS from the outer envelope of Gram-negative bacteria, but are generally considered to have antagonistic functions. LBP serves to alarm the host to the presence of minute amounts of LPS, and can therefore be considered to be proinflammatory, whereas BPI has been shown to render LPS noninflammatory (Elsbach and Weiss 1998; Fenton and Golenbock 1998). The predominance of these pro- and anti-inflammatory pathways may ultimately determine the host's response to bacteria, and perturbation of this balance may result in severe sepsis with concomitant mortality (Elsbach and Weiss 1998; Martin 2000). The anti-inflammatory and antimicrobial activity of BPI has therapeutic applications that are currently under active development (Levy 2002).

We recently showed that the human *PLUNC* gene (Bingle and Bingle 2000) belongs to a family of at least seven genes located in a 300 kB locus on chromosome 20q11.2 (Bingle and Craven 2002), and we further showed that the PLUNC family proteins are predicted to demonstrate significant 3D similarity to BPI and LBP. We have defined two subgroups of PLUNC family proteins: the short (SPLUNC1, 2, 3, etc.) and the long (LPLUNC1, 2, 3, etc.) proteins. "Short" proteins have homology only to the N-terminal domain of BPI, whereas "long" proteins have homology to both the N- and C-terminal domains of BPI and LBP. Within this classification the protein originally designated hPLUNC becomes hSPLUNC1. In addition to the predicted structural similarity to LBP/BPI, there are a number of observations that suggest that PLUNC family proteins may function in host defense against bacteria. The rat protein, PSP, has been shown to interact with bacterial membranes (Robinson et al. 1997). hSPLUNC1 is present in the antimicrobial fraction of human nasal secretions (Cole et al. 2002), and has been shown to be increased in the sputum of patients with chronic obstructive pulmonary disease (Di et al. 2003). In addition, PLUNC family proteins are characteristically expressed in regions of the respiratory tract, oro- and nasopharynx (Weston et al. 1999; Bingle and Bingle 2000; Le Clair et al. 2001; Bingle and Craven 2002; Di et al. 2003), which are sites of significant bacterial load and locations where LBP and BPI are not significantly expressed, although certain cell lines from these tissues have been shown to express both proteins (Dentener et al. 2000; Canny et al. 2002). Host defense proteins in these regions may be directly bactericidal, they may be bacteriostatic, or they may play a role in preventing an inappropriate inflammatory response (Ganz 2002).

Our previous study (Bingle and Craven 2002) was restricted to the human PLUNC family proteins. In this article we survey the extent to which PLUNC family proteins exist in other species, and whether the molecular evolution of the proteins (both at the level of paralogs and orthologs) provides any clues to their function.

## Results

### *Sequence data*

PLUNC family and BPI-related genes in the public databases were identified via BLAST searches using established PLUNC family and BPI-related genes as queries. On the basis of analysis of interspecies sequence alignments, gene structure, computerized gene predictions, and EST databases it was possible to make additional revisions of several deposited sequences (hLPLUNC3, hSPLUNC3, mLPLUNC1, mLPLUNC4, rLPLUNC3, rLPLUNC4, rSPLUNC5, rBPI), predictions of novel genes (mLPLUNC3, mLPLUNC6, rLPLUNC1, rLPLUNC2, rLPLUNC6, rSPLUNC3), and prediction of genes that have also been confirmed in recent or patent database depositions (mBPI, hLPLUNC4). The relationships among these sequences are summarized in a phylogenetic tree in Figure 1. Protein sequences, accession codes, alternative nomenclature, and details of the extent to which sequences are novel and/or predicted/experimentally confirmed are available as Supplemental Material. We identified 12 homologs of the human PLUNC family proteins in

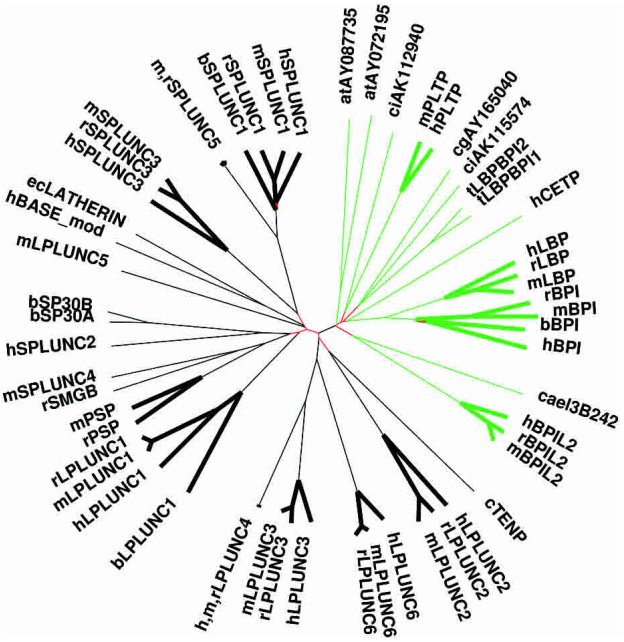


**Figure 1.** Unrooted phylogenetic tree of BPI and PLUNC family members. The PLUNC family branch of the tree is colored black, and the BPI branch is in green. Branches with bootstrap values <85% are colored red; the detailed branching order is uncertain in these regions. Thickened branches delineate groups of orthologous proteins, as defined in the text. See Materials and Methods for details of the regions of sequence used to construct the tree.

both the mouse and the rat. The coverage in the databases for other species is much less complete; however, the databases contain four bovine homologs, and one homolog each for horse and chicken. For the cow proteins, we have included those proteins previously identified in the literature (Wheeler et al. 2002), but ESTs also exist for proteins similar to LPLUNC2 and LPLUNC4. A number of the plunc family members that we identified have previously been described separately, with a varied nomenclature (Poulsen et al. 1986; Dear et al. 1991; Mirels and Ball 1992; Wheeler et al. 2002; Andrault et al. 2003). Where the orthology to the human PLUNC family proteins is clear (see below) we use our nomenclature for these proteins (Bingle and Craven 2002). Where the orthology is unclear, we maintain the original names in the literature. A list of alternative nomenclatures is provided in Supplemental Material.

In the analyses below we have also included the human, mouse, rat, and bovine orthologs of the BPI/LBP/CETP/PLTP proteins, where they exist in the databases, or where we have been able to make predictions/annotations. To place our data in the context of the widest possible view of the BPI family, the collection also includes a range of BPI-related proteins from diverse species obtained from BLAST searches. Mouse and rat orthologs of human BPI were not annotated in the databases prior to our analysis. Using the human BPI sequence, knowledge of the patterns of exon sizes in the BPI proteins (see below), the mouse and rat genome sequences, and the EST databases, it was straightforward to predict a novel sequence in mouse and rat that shared 53% and 55% pairwise amino acid identity, respectively, with human BPI. We confirmed the expression of *BPI* in mouse testes and alveolar neutrophils by RT-PCR (results not shown). The classification of these sequences as the rat and mouse orthologs of human BPI is discussed further below.

LPLUNC4 was unique among the PLUNC family proteins that we previously described (Bingle and Craven 2002), as it was not complete at the N-terminal end. To complete the sequence we identified a series of mouse ESTs that extended the mouse sequence in the 5′ direction. These sequences were subsequently mapped onto the mouse and rat genome sequences to identify the intron/exon arrangement. This information was used to generate a complete open reading frame for both the mouse and rat genes. Both predicted proteins contain a signal peptide. This information was then used to predict the complete human *LPLUNC4* gene. The *LPLUNC4* gene contains a very small third exon (63bp) compared to all of the other *PLUNC* family genes as well as a large additional exon, which we term exon 3b. Subsequently, the sequence of human *LPLUNC4* was confirmed by a complete cDNA sequence (AX283507) within the patent divison of Genbank. A further alternatively spliced human *LPLUNC4* sequence with an extended third exon is also found in the patent division (AX283509).

*Identification of orthologous relationships*

The completeness of the mouse genome sequence allowed the definition of the *PLUNC* family gene locus (Fig. 2). For the rat, the genomic data are slightly less complete, and therefore, the exact size of the locus remains to be completely defined. It appears to be highly similar to that seen in the mouse. In all three species the PLUNC family proteins have therefore remained in a compact cluster.

From analysis of the phylogenetic tree (Fig. 1), genomic locus (Fig. 2), and pairwise identity matrices (Supplemental Material) it was possible to define clear 1:1 orthologs (Waterston et al. 2002) in humans, mice, and rats for LPLUNC1, LPLUNC2, LPLUNC3, LPLUNC4, LPLUNC6, SPLUNC1, and SPLUNC3. These proteins satisfy the requirements of reciprocal best matches, and the genes are present in closely similar locations in the loci, taking into account the insertion of a small number of nonorthologous genes. These defini-

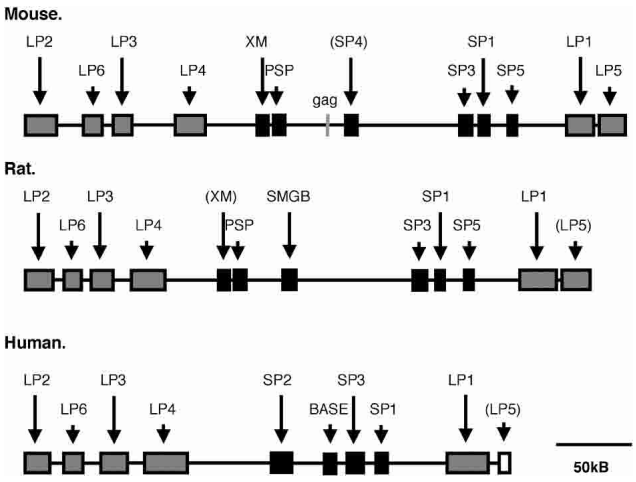


**Figure 2.** The organization of the mouse, rat, and human *PLUNC* family gene loci. The stippled and black boxes represent "long" (LP) and "short" *PLUNC* family proteins (SP). Due to the less complete nature of the rat genomic sequences in the htgs database (as of December 22, 2002) the absolute size of the rat locus may be underestimated. The updated human locus is also presented showing the position of *LPLUNC6, XM* and *BASE* and the partial sequence of the ortholog of *LPLUNC5* (shown as the white box), which may represent a pseudogene. The sequence of *rLPLUNC5* is a near complete prediction. The gene designated *XM* represents a predicted gene (XM_206487), which has definite *PLUNC* family characteristics but which we are presently unable to assign to either the Long or Short PLUNC family proteins due to the lack of expression information and the lack of a definitive prediction in GENSCAN (Burge and Karlin 1997). Genes contained within brackets are those for which there is no expression information in any form. Immediately 5′ of the mouse *SPLUNC4* gene is a viral *gag* gene (gag), which due to its position may block transcription of the otherwise completely predicted gene. The size bar represents 50 kb. During our study the cloning and sequencing of human *LPLUNC2* and *LPLUNC6* was reported (Mulero et al. 2002). In this article the authors designated these proteins as BPI-like proteins (BPIL) 1 and 3, respectively. The position of the human *BASE* gene (Egland et al. 2003) is also shown adjacent to *SPLUNC3*.

tions are further strengthened by the fact that the pairwise amino acid identity between these orthologs is always greater than 55%, which contrasts clearly with the typically much lower similarity observed between paralogs in the PLUNC family. Orthologs can be identified for SPLUNC5 and LPLUNC5 in the mouse and rat only. From a more subjective assessment of their position in the phylogenetic tree, bovine members of the LPLUNC1 and SPLUNC1 groups were also identified. For the other members of the PLUNC family the orthology is less clear. mPSP and rPSP satisfy the criteria for 1:1 orthologs, although they are the most dissimilar mouse/rat pair in the family, with only 69% pairwise identity. The human ortholog of these proteins is potentially hSPLUNC2 (Emes et al. 2003), but this is not a clear 1:1 orthology, as hSPLUNC2 is equally similar to rSMGB and mSPLUNC4. The similarity of rSMGB and mSPLUNC4 is even lower than for mPSP and rPSP, and rSMGB is only weakly more similar to mSPLUNC4 (44%) than it is to mPSP (38%). The bovine proteins bSP30a and bSP30b (Wheeler et al. 2002) are very weakly more similar to hSPLUNC2 than to any other protein, but at a level too low to propose a close orthologous relationship. Likewise, cTENP is potentially a weak ortholog of the LPLUNC2 proteins. The protein hBASE is discussed further below. The limited published expression data is consistent with the orthologous relationships outlined above (Poulsen et al. 1986; Mirels and Ball 1992; Le Clair et al. 2001; Bingle and Craven 2002; Andrault et al. 2003).

In our current analysis there are two exceptions to the rule of low pairwise identity between paralogous proteins in the PLUNC family. The bovine proteins bSP30a and bSP30b share 83% pairwise amino acid identity, and the proteins SPLUNC1 and SPLUNC5 share 62% pairwise identity in the mouse, and 59% pairwise in the rat. Along with the restricted species distribution of the *SPLUNC1* and *SPLUNC5* genes and their close proximity in the locus, it appears that *SPLUNC1* and *SPLUNC5* have arisen by a comparatively recent gene duplication that postdates the divergence of the human–rodent lineages.

Using the criteria we describe above for the definition of the PLUNC family orthologs we would classify mBPI and rBPI as the true orthologs of human BPI. First, mBPI and human BPI share 53% amino acid identity, whereas the next closest human protein is human LBP with which it shares only 38% identity. The reverse relationship also holds, that mBPI is the closest murine homolog of human BPI. The figures are similar for rBPI, and these relationships are reflected by the location of mBPI and rBPI in the same branch of the phylogenetic tree as hBPI. Second, the *mBPI* and *rBPI* genes are located in a region synteneic to that of *hBPI,* between the *TGM2* and *LBP* genes (Gray et al. 1993). cDNA clones of *mBPI* (AK033770) have recently appeared in the public databases, along with an automated prediction for *rBPI* (XM_230800) of a sequence similar but not identical to our prediction. Neither of these sequences are annotated as orthologs of human *BPI*, presumably due to the low sequence similarity, although the genomic location of *mBPI* was recently also proposed by Andrault et al. (2003). Proof that these sequences are the functional counterparts of hBPI will require further studies.

### *Patterns of exon sizes*

Exon sizes were collated for all the complete mouse, rat, and human sequences. The sizes of exons for one representative of each paralog is shown in Figure 3. Between orthologous genes, the exon sizes are generally well conserved (data not shown). As has been previously observed (Gray et al. 1993), the sizes of exons are well conserved across the *BPI/CETP/PLTP/LBP* family, with a small degree of variation in the final exon, which in the case of CETP may have functional consequences (Bruce et al. 1998). The long *PLUNC* family genes generally display a very similar pattern of exon sizes to those of the *BPI/CETP/PLTP/LBP* genes, especially for exons 4 to 15. In contrast, exon 3 is highly variable in size, and *LPLUNC4* includes the additional exon, exon 3b. The short *PLUNC* family genes

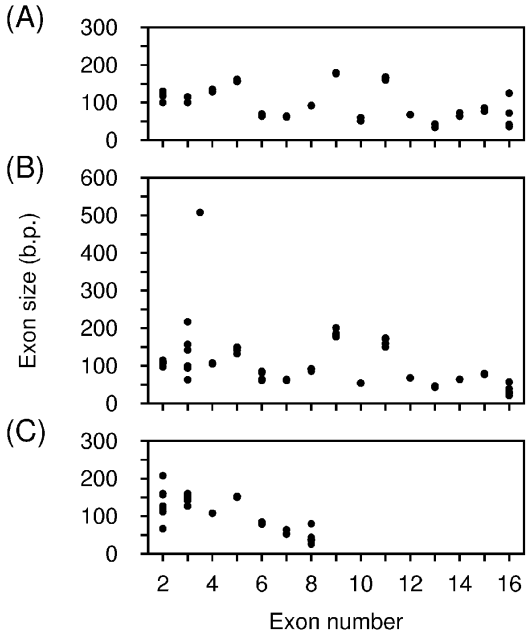


**Figure 3.** Conservation of exon sizes in BPI and PLUNC family proteins. The number of coding nucleotides in the coding exons of BPI and PLUNC family proteins (exons 2–8 in one domain proteins and exons 2–16 in two domain proteins) for (*A*) hBPI, hLBP, hPLTP, and hCETP; (*B*) hLPLUNC1, hLPLUNC2, hLPLUNC3, hLPLUNC4, mLPLUNC5, and hLPLUNC6; (*C*) hSPLUNC1, hSPLUNC2, hSPLUNC3, mPSP, rSMGB, mSPLUNC4, mSPLUNC5. For LPLUNC4, the large additional exon 3b is shown positioned between exons 3 and 4. For CETP, the two exons that correspond to exon 5 of BPI have been shown with a combined size of 159 nucleotides.

preserve the exons sizes for exons 4 to 7 found in the BPI/CETP/PLTP/LBP proteins and the long PLUNC family proteins. The greatest variation in the short *PLUNC* family genes occurs in the first coding exon. The structural consequences of these facts are discussed further below.

### *Mapping sequences onto BPI*

BPI contains two structurally similar, yet highly sequence dissimilar domains (Beamer et al. 1997). The LPS binding and bactericidal activity of BPI is conferred by the N-terminal domain, whereas the C-terminal domain is involved in the opsonic role of the protein (Elsbach and Weiss 1998; Levy 2002). The long PLUNC family proteins contain both domains, whereas the short PLUNC family proteins are predicted to be most similar to the N-terminal domain (Bingle and Craven 2002).

With the exception of BASE (see below), analysis using the 3DPSSM 3D-threading server (Kelley et al. 2000) confidently predicted all proteins in this study to share a similar fold to at least one domain of BPI (data not shown). Furthermore, all proteins, with the exception of BASE and the two proteins from *Arabidopsis thaliana*, conserve the cysteine pair corresponding to the disulphide bond found in BPI.

Attempts to thread the complete LPLUNC4 sequence onto the BPI structure provided a firm alignment to BPI only beyond residue L227, corresponding to residue I58 of BPI. I58 lies very close to the boundary of sequence encoded by exons 3 and 4 in BPI (Fig. 4A). The 3DPSSM algorithm was unable to confidently map the residues between this point and the N terminus onto the BPI structure. However, the N-terminal portion of the sequence (once the signal peptide is excluded) was predicted to form a β-strand and α-helical segment as found at the N-terminal end of the BPI structure (Beamer et al. 1997). In addition, the sequence corresponding to the extra exon 3b was not predicted to form stable secondary structure. We hypothesized that exons 3 and 3b might encode a largely unstructured insertion of approximately 150 residues that replaces the 40s hairpin (in the conventional BPI numbering scheme; see description of the BPI structure below). Therefore, we used the 3DPSSM algorithm to predict the structure of a modified form of the hLPLUNC4 sequence in which exon 3b was deleted. This created a threading that successfully mapped residues 25–44 onto the N-terminal 20 residues of BPI, at a similarly high confidence level as the rest of the mapping of the structure (data not shown). Clear mapping to the BPI structure resumes at M56 (Fig. 4A). Consistent with the substantial disruption of structure in the 40s hairpin, the mapping also predicts the replacement of the 90–103 hairpin of BPI by a four-residue loop, as previously also found for the incomplete LPLUNC4 sequence (Bingle and Craven 2002).

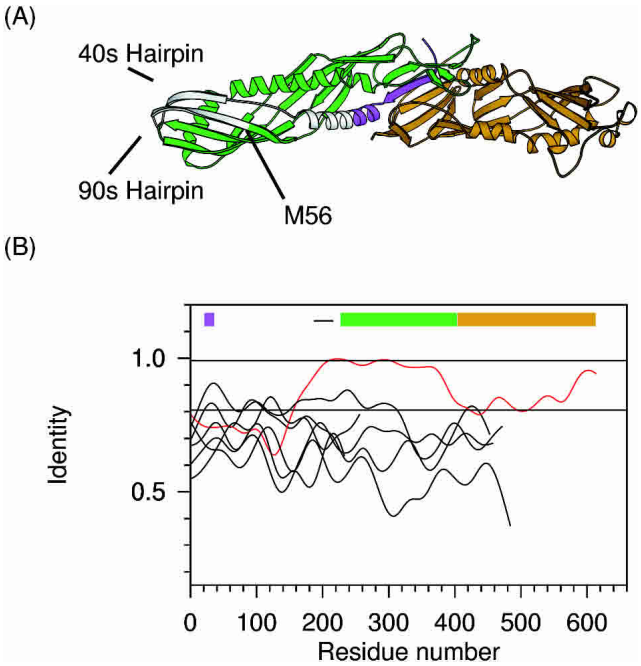


**Figure 4.** Sequence comparison of mouse and human LPLUNC proteins. (*A*) The 3D structure of human BPI colored to show the region encoded by exon 2 (purple) and that encoded by exon 3 (white), the remainder of the N-terminal domain (green), and the C-terminal domain (gold). (*B*) Locally averaged human/mouse sequence identity traces overlaid for SPLUNC1, SPLUNC3, LPLUNC1, LPLUNC2, LPLUNC3, LPLUNC4, and LPLUNC6. The trace for LPLUNC4 is in red; traces for the other proteins are in black. The average at each sequence position was calculated as a Gaussian weighted average of neighboring positions. The standard deviation of the Gaussian function used was 15 residues. The colored bars mark the regions corresponding to the regions colored in (*A*) mapped to the residue numbers of LPLUNC4. The thin bar shows the position of the GL-rich region in the sequence of LPLUNC4. The two horizontal lines mark the 16th to 83rd percentile ranges of amino acid identity observed in domain-containing protein regions in a survey of the human and mouse genomes (Waterston et al. 2002) . Individual traces are available as Supplemental Material.

### *BASE*

The *BASE* gene was recently discovered in a screen for genes upregulated in breast cancer (Egland et al. 2003). In normal tissue it is predominantly found in the salivary gland. The BASE protein is only 179 residues, compared to a typical size of approximately 250 residues for short PLUNC family proteins. Analysis of the genomic sequence indicates that the truncation of this protein arises from a single nucleotide deletion in exon 6. The resultant frameshift generates a stop codon that deletes the whole of the second major helix, including one of the cysteines in the otherwise highly conserved disulphide bond. If an extra nucleotide is inserted into the 5′ end of exon 6, a full-length *BASE_mod* gene can be constructed using the remaining two coding exons, which has approximately 41% identity over its entire length to equine Latherin. We have cloned

*BASE* from salivary gland, confirming that the reported sequence is correct (results not shown). The *BASE_mod* gene would encode a protein that is predicted by 3DPSSM to form a fold similar to the other single domain proteins. It is unlikely, although not inconceivable, that BASE is a folded protein, because an integral part of the structure has been deleted. Our hypothesis is that *BASE* represents a "dying gene" that is still expressed despite not producing a viable folded protein product.

### *Human–mouse sequence comparisons*

The identification of orthologous pairs of human and mouse PLUNC family proteins allowed us to assess the degree and patterns of sequence conservation in the PLUNC and BPI families. Two measures were used: (1) sequence wide Ka/Ks, and (2) locally averaged sequence identity.

#### *Ka/Ks*

Ka/Ks (Hughes and Nei 1988) data for the PLUNC family and BPI/LBP/PLTP proteins is shown in Table 1. For domain containing protein regions, the median value for Ka/Ks in the human/mouse genomes is 0.061, with a range of 0.015–0.178 at the 16th to 83rd percentiles (Waterston et al. 2002). Therefore, with the exception of LPLUNC4, for which the sequence conservation is further discussed below, all the PLUNC family proteins show substantially elevated rates of molecular evolution. Similarly elevated values are observed for BPI and LBP, and also for the functionally unclassified protein BPIL2 (Mulero et al. 2002).

#### *Locally averaged sequence identity*

To determine whether sequence conservation was uniform across the sequences, we calculated local pairwise identity smoothed using a gaussian (standard deviation = 15) moving window for each of the mouse human ortholog pairs (Fig. 4B).

**Table 1.** *Ka/Ks data for the PLUNC or BPI family proteins for which mouse/human orthologs exist*

| | Ka | Ks | Ka/Ks |
|---|---|---|---|
| SPLUNC1 | 0.18 | 0.40 | 0.46 |
| SPLUNC3 | 0.25 | 0.50 | 0.49 |
| LPLUNC1 | 0.28 | 0.87 | 0.32 |
| LPLUNC2 | 0.21 | 0.49 | 0.43 |
| LPLUNC3 | 0.13 | 0.51 | 0.26 |
| LPLUNC4 | 0.08 | 0.63 | 0.13 |
| LPLUNC6 | 0.16 | 0.72 | 0.23 |
| BPI | 0.36 | 0.76 | 0.47 |
| LBP | 0.20 | 0.76 | 0.26 |
| PLTP | 0.10 | 0.70 | 0.15 |
| BPIL2 | 0.19 | 0.44 | 0.44 |

In most sequences the pattern is not signficantly different from random. In LPLUNC4, however, the pattern is quite different, as it contains an approximately 200-residue stretch that is almost totally conserved. This region corresponds to a GL rich region that we have previously commented upon (Bingle and Craven 2002) which is encoded by the 3′ end of exon 3b, and to the remainder of the N-terminal domain encoded by exons 4 to 8 and part of exon 9. In this 200-residue stretch there are only three amino acid substitutions; thus, this region is under intense purifying selection (i.e., selection acting to purge mutations).

It was also investigated whether the sequence variability between human/mouse orthologous pairs mapped to equivalent positions in the 3D structure of BPI for groups of PLUNC family proteins. No statistically significant clustering could be detected (data not shown).

## Discussion

### *Structure of the PLUNC/BPI superfamily*

Previously (Bingle and Craven 2002) we showed that the human PLUNC family proteins form a subfamily of a BPI/PLUNC superfamily. This division was based upon the genomic colocalization of the *PLUNC* family genes into a compact cluster of approximately 300 kB, and upon the observation that the degree of structural conservation is predicted to be significantly different in the two families.

Our results here show that clusters of genes orthologous to the *PLUNC* family genes are also present in the mouse and rat genomes, and that a phylogenetic tree robustly classifies the PLUNC family proteins into a separate branch to the BPI/LBP/CETP/PLTP proteins (Fig. 1). The proteins falling into the PLUNC family branch are exclusively from air-breathing vertebrates, whereas the BPI branch also contains proteins from invertebrates and plants. Definitive proof that the PLUNC family proteins are restricted to air-breathing vertebrates will only be possible with the completion of genome sequences from more diverse species.

In turn, the PLUNC family branch approximately subdivides into two subbranches of long and short proteins (see below regarding LPLUNC1 and LPLUNC5). Because we restricted comparison to a portion of the N-terminal domain, we can infer that this division in the phylogenetic tree is a true reflection of detailed amino acid sequence, rather than a consequence of differences necessitated by the one- or two-domain nature of the proteins involved. The division of the tree into short and long branches strongly argues against the hypothesis that the short and long proteins might form agonist/antagonist pairs. A similar grouping of the short proteins is also observed in the genomic locus (Fig. 2).

The structural similarity of the two domains of the long proteins strongly suggests they arose from a duplication

event in a one-domain protein. As PLUNC family proteins are apparently restricted to air-breathing vertebrates, whereas two domain BPI like proteins in general are found in a wide range of species (including distant homologs in plants and worms), it appears the short PLUNC family proteins must have arisen from a comparatively recent deletion of the C-terminal domain. If this deletion occurred following a duplication of a precursor of the *LPLUNC1/5* genes, this would explain the positioning of the *SPLUNC* genes between the *LPLUNC2/3/4/6* and *LPLUNC1/5* genes in the genomic locus, and the positioning of *LPLUNC1/5* proteins slightly closer to the SPLUNC proteins in the phylogenetic tree. There is also evidence for a link between *LPLUNC1/5* and the *SPLUNC* genes in the detailed sizes of exons 4–7 (see Supplemental Material).

### *Rapid molecular evolution of the PLUNC family proteins supports a role in innate immunity*

Very recently it has become possible to perform whole genome-wide analyses of the evolution of human and mouse orthologs (Waterston et al. 2002; Emes et al. 2003). Such analysis has shown that, for instance, 80% of proteins have a 1:1 ortholog in the corresponding genome, and that domain containing orthologs will share on average 81%–99% (16th to 83rd percentiles) sequence identity (Waterston et al. 2002). The levels of sequence identity between the orthologs in the PLUNC family are much lower than this, being in the range 45%–76% (with the notable exception of LPLUNC4 as discussed above). An alternative measure of the evolutionary pressure on proteins is the Ka/Ks ratio (Hughes and Nei 1988; Hurst 2002). It has been observed that an elevated Ka/Ks ratio, and hence rapid molecular evolution, is a frequent characteristic of proteins involved in host defense (Hurst and Smith 1999; Waterston et al. 2002; Emes et al. 2003). The data for the PLUNC family proteins (Table 1) indicate that this family is evolving more rapidly than the eight families with the highest median Ka/Ks described in Table 13 of Waterston et al. (2002). Six of these families are involved in host defense and immunity. Thus, the data are consistent with a role in innate immunity; however, proteins involved in other functions also exhibit rapid evolution. For instance, the p450 family of proteins exhibit elevated Ka/Ks values, and are involved in the metabolism of toxic compounds. Such a function therefore appears to necessitate rapid adaptive change (Emes et al. 2003), and a clearance or transport role is an alternative possible function for the PLUNC family. More detailed analysis of whether the elevated Ka/Ks values observed in the PLUNC family are the result of reduced purifying selection, or of positive selection at a few sites must await the availability of many more vertebrate genome sequences.

The clustering of the *PLUNC* family genes also probably reflects the rapid evolution of the PLUNC protein family as a whole, with gene duplication occurring more rapidly than gene dispersion. Similarly, the conservation of exon sizes in the presence of very low paralogous similarities may reflect the effect of rapid evolution following rather recent paralogous duplications.

### *LPLUNC4 displays unique features: An extreme case, or a new function?*

To discuss the unique features of the LPLUNC4 proteins, it is helpful first to consider some features of the BPI fold (Fig. 4A). The BPI fold can be described as two distorted "barrels," connected by an interdomain β sheet (Beamer et al. 1997). Each barrel is akin to a long thin barrel where two of the strands are replaced by long α-helices. The break in the side of the barrel produced by these helices generates the opening to the lipid binding pockets. The β-sheet connecting the two barrels comprises a number of strands from each domain. In each domain, these strands form the first and the last secondary structure elements in the sequences. The first two coding exons, exons 2 and 3, encode residues 1–55. These residues comprise the signal peptide, the first strand in the interdomain sheet, the first long α-helix, and the 40s hairpin at the end of the barrel distal to the interdomain sheet. This hairpin (which we refer to throughout as the "40s hairpin") contains a positively charged loop that has been suggested to be important for the disruption of the phospholipid/divalent cation interactions in the bacterial outer membrane. The role of this loop, however, is not definitively established, because it has also been proposed to be involved in the LPS neutralizing and antimicrobial roles of BPI, yet the sequence is moderately well conserved in LBP, which is neither LPS neutralizing nor directly antimicrobial. The sequence encoded by exon 4 onwards forms the four long runs of twisted and somewhat irregular structure that form the pseudo β-barrel before the second long helix that then runs into the final strands of the N-terminal domain that are part of the interdomain β-sheet. The topology of the C-terminal domain is similar, although somewhat less regular.

We previously noted (Bingle and Craven 2002) that although the PLUNC family is predicted to be structurally similar to BPI, there is predicted to be a diversity of structure, especially in the tip of the N-terminal domain distal to the interdomain region. Such diversity is not predicted for LBP, PLTP, and CETP. The variability in the structure in this region, coupled with the low sequence identity between sequences, made it very difficult to predict a clear structural basis to the variability between sequences. Analysis was further complicated by the fact that a complete sequence was not available at that time for hLPLUNC4.

Starting from the complete sequence of the LPLUNC4 proteins, and the associated exon structure, it is now possible to refine the picture of the pattern of variability. As we

have described above, it appears that a consistent model for LPLUNC4 can be constructed by assuming that exons 3 and 3b encode a large unstructured insertion that substantially replaces the 40s hairpin. Across the LPLUNC family, exon 3 is by far the most variable fully coding exon in terms of size (Fig. 3B), suggesting that some aspects of the model of LPLUNC4 can be extended to the other LPLUNC proteins, and that the predominant features of variability in the LPLUNC proteins reside in the nature of the structure in the region homologous to the 40s hairpin.

In the short PLUNC family proteins the variability appears to reside instead in exon 2, which encodes the N-terminal part of the protein. Further analysis is considerably more difficult, as the nature of the structure in this redundant interdomain β-sheet region cannot be modeled on BPI. In SPLUNC1, this variability is so great that there is a significant difference in the lengths of the human, mouse, and rat proteins (see Supplemental Material).

This interpretation of LPLUNC4 as a paradigm for the rest of the LPLUNC proteins is, however, challenged by the observation that the approximately 200-residue portion of sequence encoded by the 3′ end of exon 3b and extending through most of the rest of the N-terminal domain is >97% conserved between human and mouse, in complete contrast to all the other sequences in the PLUNC or BPI families (Fig. 4B). If the enhanced molecular evolution in the PLUNC family proteins and BPI and LBP proteins is indeed due to interaction with a rapidly evolving pathogen, then this effect appears to be completely replaced by extreme purifying selection in a large part of one domain of LPLUNC4.

A second noteworthy feature in the LPLUNC4 sequence is the GL-rich sequence GLLGGGGLLGDGGLLGGGGVL, which is encoded by the 3′ end of exon 3b, and is therefore, positioned just to the C-terminal end of the putative greatly expanded and unstructured 40s loop. This curious sequence is not unique to LPLUNC4, as a highly similar sequence (GLLGSGGLLGGGGLLGHGGVF) is found in hLPLUNC3, encoded at the end of exon 3. Two smaller GL-rich sequences are also found in the SPLUNC1 proteins (Bingle and Craven 2002) and are also encoded by exon 3. The cDNA for the alternatively spliced variant of hLPLUNC4 encodes an extra GL-rich repeated peptide N-terminal to the sequence encoded by exon 3b. In this variant, therefore, the 180-residue inserted segment is flanked at either side by GL-rich regions.

These observations therefore raise the question as to the functional homogeneity of the PLUNC family proteins. Does LPLUNC4 simply represent a somewhat extreme case in a broad spectrum of structural diversity in proteins with an underlying common function, or has LPLUNC4 evolved a completely new function? The large insertion and the unique pattern of conservation might suggest a new function, whereas the conserved GL-rich region that is also found in LPLUNC3 suggests a retained functional commonality.

### *Conclusions*

In this study we have shown that PLUNC family proteins exist in a number of species, and that they appear to be restricted to air-breathing vertebrates. Comparison of the human and mouse orthologs shows that the family is very rapidly evolving, which is consistent with involvement in host defense. Comparison of paralogous proteins shows that, in the LPLUNC proteins, the size of the second coding exon (exon 3) is much more highly variable than is the size of the other exons, and that LPLUNC4 contains an extra exon (exon 3b) that encodes a putatively unfolded insert of approximately 150 residues. Exon 3 encodes sequence corresponding to the 40s hairpin of BPI, and the variability of this region may be key to differences in specificity between the members of the family. In the SPLUNC proteins, the variability appears to instead reside in the N-terminal sequence. LPLUNC4 is also anomalous compared to the other PLUNC family proteins, in that the rapid evolution between the human and mouse sequences is replaced by intense purifying selection in the majority of the part of the N-terminal domain that is predicted to be structured. These results provide a rational framework upon which to base further functional studies of PLUNC family proteins and their potential role in innate immunity.

## Materials and methods

Sequence searches utilized the NCBI blast server (http://www.ncbi.nlm.nih.gov/blast), the ensembl server (http://www.ensembl.org), and the BLAT server (http://genome.ucsc.edu). Gene predictions were refinements of inital predictions made using the GENSCAN server (http://genes.mit.edu/; Burge and Karlin 1997). The phylogenetic analysis was performed on the amino acid sequences using CLUSTAL W (Thompson et al. 1994), which implements the neighbor-joining method of Saitou and Nei (1987). A portion of the N-terminal domain was used for the phylogenetic analysis, comprising residues corresponding to the central residues 56–182 of BPI in pairwise alignments obtained from the threading analysis performed using 3DPSSM. This choice of residues allowed the comparison of long and short proteins, because it comprises the core structure of the domain, while excluding those parts of the domain that are intimately involved in interdomain contacts. Furthermore, it excludes the sequence in the extreme N terminus of the domain that is predicted to be structurally very variable (Bingle and Craven 2002; and see below), for which definitive alignments of paralogs are difficult. The essential details of the tree are insensitive to this choice. For instance, the division between the BPI and PLUNC family branches, and the apparent closer relationship between LPLUNC1 and the SPLUNC proteins is maintained even if the full length of all sequences is used. The representation of the tree was created using PHYLIP (Felsenstein 1989) and adjusted using in-house software. No correction of distances for multiple substitution was made. Bootstrap values were obtained from 1000 random samples. Sequence alignments were made us-

ing CLUSTAL W. For calculation of mouse/human Ka/Ks ratios, orthologous amino acid sequences were aligned using CLUSTAL W, and the alignment obtained was transferred to the cDNA sequences. Ka/Ks values were then calculated using the yn00 program of the PAML package, which implements the method of Yang and Nielsen (2000). Figure 4A was created using Molscript (Kraulis 1991), using the coordinates 1ewf.pdb (Beamer et al 1997).

## Electronic supplemental material

Supporting information is available, comprising a PDF file containing a table of accession codes and extra annotation; alignments of amino acid sequences where our sequence deviates from the published sequence; protein amino acid sequences; individual traces as overlaid in Figure 4B; a figure of exon sizes for individual proteins for exons 4–7.

## Acknowledgments

This work was funded by the National Heart, Blood and Lung Institute (R15-HL067220 to E.E.L.), a joint Research Grant from the American Lung Association and the American Lung Association of Metropolitan Chicago (to E.E.L.), and the British Lung Foundation (to L.B. and C.D.B.). The Krebs Institute is a designated BBSRC center and a member of NESBIC. We thank Chris Ponting for helpful discussions, and Rosie Staniforth, Laszlo Hosszu, Martin Parker, Martina Daly, and David Dockrell for comments on the manuscript. Paul Frossell wrote the software to efficiently perform multiple submissions to 3DPSSM.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## References

Andrault, J.-B., Gaillard, I., Giorgi, D., and Rouquier, S. 2003. Expansion of the BPI family by duplication on human chromosome 20: Characterization of the RY gene cluster in 20q11.21 encoding olfactory transporters/antimicrobial-like peptides. *Genomics* **82:** 172–184.

Beamer, L.J., Carroll, S.F., and Eisenberg, D. 1997. Crystal structure of human BPI and two bound phospholipids at 2.4 Å resolution. *Science* **276:** 1861–1864.

Bingle, C.D. and Bingle, L. 2000. Characterisation of the human plunc gene, a gene product with an upper airways and nasopharyngeal restricted expression pattern. *Biochim. Biophys. Acta* **1493:** 363–367.

Bingle, C.D. and Craven, C.J. 2002. PLUNC: A novel family of candidate host defence proteins expressed in the upper airways and nasopharynx. *Hum. Mol. Genet.* **11:** 937–943.

Bruce, C., Beamer, L.J., and Tall, A.R. 1998. The implications of the structure of the bactericidal/permeability increasing protein on the lipid-transfer function of the cholesteryl ester transfer protein. *Curr. Opin. Struct. Biol.* **8:** 426–434.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Canny, G., Levy, O., Furuta, G.T., Narravula-Alipati, S., Sisson, R.B., Serhan, C.N., and Colgan, S.P. 2002. Lipid mediator-induced expression of bactericidal/permeability-increasing protein (BPI) in human mucosal epithelia. *Proc. Natl. Acad. Sci.* **99:** 3902–3907.

Cole, A.M., Liao, H., Stuchlik, O., Tilan, J., Pohl, J., and Ganz, T. 2002. Cationic polypeptides are required for antibacterial activity of human airway fluid. *J. Immunol.* **169:** 6985–6991.

Dear, T.N., Boehm, T., Keverne, E.B., and Rabbitts, T.H. 1991. Novel genes for potential ligand-binding proteins in subregions of the olfactory mucosa. *EMBO. J.* **10:** 2813–2829.

Dentener, M.A., Vreugdenhil, A.C., Hoet, P.H., Vernooy, J.H., Nieman, F.H., Heumann, D., Janssen, Y.M., Buurman, W.A., and Wouters, E.F. 2000. Production of the acute-phase protein lipopolysaccharide-binding protein by respiratory type II epithelial cells: Implications for local defense to bacterial endotoxins. *Am. J. Respir. Cell. Mol. Biol.* **23:** 146–153.

Di, Y.-P., Harper, R., Zhao, Y., Pahlavan, N., Finkbeiner, W., and Wu, R. 2003. Molecular cloning and characterization of spurt, a human novel gene that is retinoic acid-inducible and encodes a secretory protein specific in upper respiratory tracts. *J. Biol. Chem.* **278:** 1165–1173.

Egland, K.A., Vincent, J.J., Strausberg, R., Lee, B., and Pastan, I. 2003. Discovery of the breast cancer gene BASE using a molecular approach to enrich for genes encoding membrane and secreted proteins. *Proc. Natl. Acad. Sci.* **100:** 1099–1104.

Elsbach, P. and Weiss, J. 1998. Role of the bactericidal/permeability-increasing protein in host defence. *Curr. Opin. Immunol.* **10:** 45–49.

Emes, R.D., Goodstadt, L., Winter, E.E., and Ponting, C.P. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12:** 701–709

Fearon, D. and Locksley, R. 1996. The instructive role of innate immunity in the acquired immune response. *Science* **272:** 50–54.

Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* **5:** 164–166.

Fenton, M.J. and Golenbock, D.T. 1998. LPS-binding proteins and receptors. *J. Leukoc. Biol.* **64:** 25–32.

Ganz, T. 2002. Epithelia: Not just physical barriers. *Proc. Natl. Acad.Sci.* **99:** 3357–3358.

Gray, P.W., Corcorran, A.E., Eddy, R.L., Byers, M.G., and Shows, T.B. 1993. The genes for the lipopolysaccharide binding protein (LBP) and the bactericidal permeability increasing protein (BPI) are encoded in the same region of human chromosome 20. *Genomics* **15:** 188–190.

Hughes, A.L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335:** 167–170.

Hurst, L.D. 2002. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **18:** 486–487.

Hurst, L.D. and Smith, N.G. 1999. Do essential genes evolve slowly? *Curr. Biol.* **9:** 747–750.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299:** 499–520.

Kraulis, P.J. 1991. Molscript—A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24:** 946–950.

LeClair, E.E., Nguyen, L., Bingle, L., MacGowan, A., Singleton, V., Ward, S.J., and Bingle, C.D. 2001. Genomic organization of the mouse plunc gene and expression in the developing airways and thymus. *Biochem. Biophys. Res. Commun.* **284:** 792–797.

Levy, O. 2002. Therapeutic potential of the bactericidal/permeability-increasing protein. *Expert Opin. Investig. Drugs* **11:** 159–167.

Martin, T.R. 2000. Recognition of bacterial endotoxin in the lungs. *Am. J. Respir. Cell. Mol. Biol.* **23:** 128–132.

Mirels, L. and Ball, W.D. 1992. Neonatal rat submandibular gland protein SMG-A and parotid secretory protein are alternatively regulated members of a salivary protein multigene family. *J. Biol. Chem.* **267:** 2679–2687.

Mulero, J.J., Boyle, B.J., Bradley, S., Bright, J.M., Nelken, S.T., Ho, T.T., Mize, N.K., Childs, J.D., Ballinger, D.G., Ford, J.E., et al. 2002. Three new human members of the lipid transfer/lipopolysaccharide binding protein family (LT/LBP). *Immunogenetics* **54:** 293–300.

Poulsen, K., Jakobsen, B.K., Mikkelsen, B.M., Harmark, K., Nielsen, J.T., and Hjorth, J.P. 1986. Coordination of murine parotid secretory protein and salivary amylase expression *EMBO. J.* **5:** 1891–1896.

Robinson, C.P., Bounous, D.I., Alford, C.E., Nguyen, K.H., Nanni, J.M., Peck, A.B., and Humphreys-Beher, M.G. 1997. PSP expression in murine lacrimal glands and function as a bacteria binding protein in exocrine secretions. *Am. J. Physiol.* **272:** G863–G871.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Waterston, R.H., Lindbald-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Weston, W.M., LeClair, E.E., Trzyna, W., McHugh, K.M., Nugent, P., Lafferty, C.M., Ma, L., Tuan, R.S., and Greene, R.M. 1999. Differential display identification of plunc, a novel gene expressed in embryonic palate, nasal epithelium, and adult lung. *J. Biol. Chem.* **274:**13698–13703.

Wheeler, T.T., Haigh, B.J., McCracken, J.Y., Wilkins, R.J., Morris, C.A., and Grigor, M.R. 2002. The BSP30 salivary proteins from cattle, LUNX/PLUNC and von Ebner's minor salivary gland protein are members of the PSP/LBP superfamily of proteins. *Biochim. Biophys. Acta.* **1579:** 92–100.

Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17:** 32–43.