

Todd E. Scheetz
Jennifer A. Bartlett
Jesse D. Walters
Brian C. Schutte
Thomas L. Casavant
Paul B. McCray, Jr

Genomics-based approaches to gene discovery in innate immunity

Authors' addresses

Todd Scheetz³, Jennifer A. Bartlett^{1,2}, Jesse D. Walters³,
Brian C. Schutte^{1,2}, Thomas L. Casavant^{2,3}, Paul B.
McCray, Jr^{1,2},

¹Department of Pediatrics, ²Interdisciplinary
Genetics PhD Program, ³Electrical and
Computer Engineering, The University of
Iowa, Iowa City, IA, USA.

Correspondence to:

Paul B. McCray, Jr
Department of Pediatrics
University of Iowa College of Medicine
Iowa City, IA 52242
USA
Tel.: 319 356 4866
Fax: 319 356 7171
e-mail: paul-mccray@uiowa.edu

Acknowledgements

We thank Margaret Malik, Andrea Penisten,
and Autumn Bradley for technical assistance.
We appreciate the comments and
contributions from our colleagues Michael
Welsh, Hong Peng Jia and Joseph Mitros. We
thank Stanley Perlman for critically
reviewing the manuscript. We acknowledge
the support of NIH NHLBI HL-61234 (P.B.M
and B.C.S) and NIH NICHD P30-HD27748
(B.C.S). T.L.C. is supported by NIH NCI CA-
85188, and a Functional Genomics Resource
Center Grant from the Cystic Fibrosis
Foundation.

Immunological Reviews 2002

Vol. 190: 137–145

Printed in Denmark. All rights reserved

Copyright © Blackwell Munksgaard 2002

Immunological Reviews
0105-2896

Summary: The completion of draft sequences of the human and mouse genomes offers many opportunities for gene discovery in the field of immunology through the application of the methods of computational genomics. One arm of the innate immune system includes the antimicrobial peptides that protect multicellular organisms from a diverse spectrum of microorganisms. The β -defensins comprise an important family of mammalian antimicrobial peptides. To better define the β -defensin gene family, we developed an approach to search genomic databases for conserved motifs present in the β -defensin family using HMMER, a computational search tool based on hidden Markov models (HMMs), in combination with the basic local alignment search tool. The approach was first used to identify candidate second-exon coding regions, and later applied to finding associated first exons. This strategy discovered 28 new human and 43 new mouse β -defensin genes in five syntenic chromosomal regions. Within each syntenic cluster, the gene sequences and organization were similar, suggesting that each cluster pair arose from a common ancestor and was retained because of conserved functions. These findings demonstrate an important proof-of-principle for a genome-wide search strategy to identify genes with conserved structural motifs. Such an approach may be readily adopted to address other questions of relevance to immunology.

Introduction

Multicellular organisms have been obliged to develop multifaceted innate and adaptive immune systems to cope with the challenges to survival originating from microorganisms and their products. The diversity and breadth of innate immune mechanisms are in large part conserved in all multicellular organisms (1). Some basic principles of microbial recognition and response are emerging, and recently, the application of computational genomics has played an important role in extending such observations from model organisms, such as *Drosophila*, to higher vertebrates, including humans.

For use in this review, we consider computational genomics as the application of computer-based algorithms/techniques

to address genome-scale problems of biological interest. Computational approaches are now established, and evolving laboratory technologies allow investigators to ask questions on a genome-wide scale. Fundamental observations leading to the identification of receptor families involved in the recognition of key components of microorganisms had their genesis in studies of *Drosophila* (2–4). The discovery that signaling cascades activated via the *Drosophila* Toll gene product induce the expression of antimicrobial peptides stimulated the search for similar pathways in mammals (5). The application of basic local alignment search tool (BLAST)-based searches in expressed sequence tag (EST) databases yielded a variety of mammalian Toll-like receptors (TLRs) that are now recognized as key first-response components of the innate immune system (6). More recently, the identification of mutations in *Drosophila* causing loss-of-function phenotypes characterized inactivation of specific inducible antimicrobial and antifungal peptides and led to the identification of a novel family of peptidoglycan recognition proteins (PGRPs) (7–9). Again, application of the BLAST technique identified human homologs of this new family of microbial pattern recognition receptors (10, 11). In our current studies of the innate immune system, we applied techniques of computational genomics to extend our understanding of the β -defensin family of antimicrobial peptides through gene discovery (12).

The defensins comprise a family of peptides whose broad-spectrum antimicrobial activity contributes to innate immune defenses across a wide range of species (13). Their expression in groups as diverse as plants, insects and vertebrates suggests that defensins represent an ancient and conserved mechanism by which organisms combat infection (14). Three classes of defensins have been identified in vertebrates. The α -defensins are expressed in neutrophils as well as macrophages and Paneth cells of the intestine (15). The θ -defensins, which feature a unique cyclic structure, have only been identified in leukocytes of rhesus macaques to date (16). The third group, the β -defensins, are found in many mammalian species as secreted products of leukocytes and epithelial cells in several tissues, including the skin (17, 18), tongue (19, 20), gingiva (19, 20), cornea (21), salivary glands (19), esophagus (20), intestine (22), kidney (23, 24), urogenital tract (23), and the respiratory epithelium (25–29; reviewed in 30). β -defensins are a component of the first line of defense against microorganisms in the mucosal surfaces of these tissues, sites where the body interfaces with its environment. In this innate immune role, β -defensins are either constitutively expressed or induced in response to microbial components or other pro-inflammatory stimuli (13). In addition to their microbicidal

properties, increasing evidence suggests a role for some β -defensins as chemokines for immature dendritic cells and memory T cells (31), thus serving as a link between the innate and adaptive immune systems.

The hallmarks of a β -defensin are small size, a high density of cationic charges and a characteristic six-cysteine motif (32). They are usually encoded by two exon genes in which the first exon encodes a hydrophobic leader sequence and the second exon contains the cysteine motif; the leader sequence is cleaved from this prepropeptide during processing to produce a final peptide of around 45 amino acids in length. The six-cysteine motif, which is present in the mature peptide sequence, maintains the tertiary structure of the peptide through three disulfide bonds, between C1–C5, C2–C4, and C3–C6 (32). The most typical spacing for this motif is C–X₆–C–X₄–C–X₉–C–X₆–CC, although this pattern may vary slightly at some positions (with the exception of C5 and C6, which are invariably adjacent) (12). While this motif is recognizable in every β -defensin family member, there is little sequence conservation beyond the six cysteines.

How is it that the β -defensins may have such divergent primary structures and retain their microbicidal activity? It appears that the characteristic pattern of disulfide bonds, as determined by the six-cysteine motif, causes the molecule to assume a conformation in which hydrophobic residues are concentrated on one surface and cationic charges are concentrated on the other (33, 34). The emerging picture suggests that the cationic surface of this polar β -defensin molecule facilitates electrostatic interactions with anionic bacterial cell wall components, while the hydrophobic surface anchors the peptide in the lipid layers of the cell membrane. Once this interaction with the bacterial cell wall is established, the peptide may then permeabilize the cell by oligomerizing to form channel-like pores or through some other mechanism that disrupts or interferes with the integrity of the cell wall (35–37). Thus, based on this model, a high degree of primary sequence conservation seems to be unnecessary for antimicrobial function as long as the three-dimensional structure of the peptide is preserved.

To better understand the role that β -defensins play in immunity, it is important to identify all members of the β -defensin family. Initially, β -defensin discovery efforts focused on protein isolation methods to identify novel defensin gene products. Such techniques, including acid extraction, gel filtration and reverse-phase high performance liquid chromatography (HPLC), are well suited for isolation of the generally small and cationic defensins. Using such an approach, bovine tracheal antimicrobial peptide (TAP) was first purified from

bovine airway tissue (25), followed by the identification of 13 defensin-like proteins from bovine neutrophils (BNBDs) (32). Together with TAP, these neutrophil peptides became the founding members of the β -defensin family, distinguished from the previously recognized α -defensins by their cysteine residue spacing and resulting pattern of disulfide bonds. Similar biochemical approaches were later used to isolate human β -defensin-1 (HBD-1, gene name *DEFB101*) from dialysate hemofiltrate (38) and HBD-2 (*DEFB102*) and HBD-3 (*DEFB103*) from skin (17, 18). While several β -defensins were subsequently isolated from other vertebrate species using similar protein-based methods, the rate of expansion of the gene family was accelerated by the introduction of nucleic acid-based strategies that identified new β -defensin genes through sequence homology. Numerous studies described the amplification of new β -defensin genes with polymerase chain reaction (PCR), using cDNA sequences of known β -defensins to design PCR primers (39–41). Interestingly, this method was most frequently successful when the primers used to amplify a new gene were derived from a β -defensin in a different species rather than from the same species. This finding suggests that such an approach is more suited to the discovery of β -defensin orthologs (genes that perform the same function in different species and would therefore be expected to share more conserved sequence) than to paralogous β -defensin genes, which perform similar functions within a species and consequently may show more sequence divergence from one another. Along similar lines, several groups reported using nucleic acid hybridization screens at low stringency to identify new β -defensins, using sequence from known β -defensins across and within species as cDNA probes (39, 42).

The search for new β -defensin genes received a further boost when it became possible to perform sequence similarity-based searches of protein, cDNA and genomic databases. A new mouse (43) and two rat (44) β -defensin genes were identified using this approach, through BLAST searches of the EST database (dbEST). However, a limitation of this strategy is that the EST database contains only sequences from genes that are expressed at the time of sample collection. Because expression of many β -defensins is induced upon stimulation by infection or pro-inflammatory agents, β -defensin family members are likely to be under-represented in cDNA databases. In addition, the rapidly growing genomic databases provide extraordinary resources for efforts to identify new β -defensin genes. As our gene discovery project began, the β -defensins were known to colocalize with the α -defensins at chromosome 8p22-p23 in the human (45). β -Defensin clusters were also identified on mouse chromosome 8 (43) and

rat chromosome 16 (44), in regions syntenic to the human defensin cluster. This genomic organization suggested that the defensin clusters may have arisen and expanded by duplication events (45), leading us to hypothesize that analysis of the genomic sequence in the vicinity of these clusters would yield additional β -defensin gene family members. This approach was used to identify human *DEFB103* (20), an inducible gene that is not represented in dbEST. Because the gene is absent from the EST database and is induced under specific conditions, its discovery illustrates the utility of a genomic DNA search strategy for identifying those β -defensin genes that are unlikely to be found using expression libraries or other expression-based methods.

Given the success of this computational approach and the observed frequency of duplication within defensin clusters, we hypothesized that additional β -defensins could be identified using a similar strategy. We therefore pursued a genome-wide, iterative computational search strategy to test this hypothesis. Our approach combined the use of two strategies, BLAST (46) and the hidden Markov model (HMM) (47). BLAST-based searches enable efficient searching of large data-sets while incorporating knowledge of amino acid substitutions. While useful in identifying the presence of specific genes, or closely related genes (i.e. close relatives in the same family), the primary drawback of a BLAST-based strategy is the pairwise nature of the search. At any time, only a single known sequence can be used as a query. In contrast, HMM-based searches do not provide incorporation of amino acid substitutions. However, they offer the ability to perform model-based searches based upon HMMs derived from multiple sequence alignments. This enables a 'broader' space to be searched. Typically, the multiple sequence alignment is refined iteratively using a multiple sequence alignment program such as ClustalW (48). To define better the β -defensin gene family, we adopted a novel genomics approach using HMMER, a computational search tool based on HMMs (47) in combination with BLAST.

BLAST-based searches

The initial strategy focused on identifying genomic sequences coding for possible novel β -defensin second exons. Genomic search strategies for human gene discovery were applied to the GenBank NR, HTGS and EST databases using the *blastp* and *tblastn* programs (46), using the tools available on the National Center for Biotechnology (NCBI) website (<http://www.ncbi.nih.gov/BLAST/>). A similar approach was used to query the Celera mouse genome assembly ([Immunological Reviews 190/2002](http://www.ce-</p>
</div>
<div data-bbox=)

lera.com). The queries for the initial search used the amino acid sequences for the known human defensins (*DEFB101*, *DEFB102*, *DEFB103*, *DEFB104*) (17, 18, 20, 38, 49), and two *SPAG11* sequences (50, 51) and the known mouse β -defensins (*Defb1*, *Defb2*, *Defb3*, *Defb4*, *Defb5*, *Defb6*) (40, 42, 43, 52, 53). The default search parameters were used in the searches and any potential hits were curated manually.

Construction of hidden Markov models for the β -defensin exon motifs

In addition to BLAST-based approaches, a complementary strategy employed sequence analysis using HMMs (47). The core human and mouse β -defensin amino acid sequences containing the six-cysteine motif were sorted according to their alignments in HMMs built from the set of known defensin motifs. Twelve second-exon, six-cysteine motifs derived from human and mouse β -defensin sequences were defined by manual inspection of full-length β -defensin domain sequences. These motifs were aligned using ClustalW (48) and trimmed of extra positions extending on both sides of a 38 amino acid core. These 12 aligned sequences were used as input for the HMMER 2.1.1 suite of software at Washington University (<http://HMMER.wustl.edu>) (47) to build the first of several HMM β -defensin models. We constructed the first model using the program *hmmbuild* and then used *hmmcalibrate* to calibrate E-value scores. Three models were built: (i) human second exon sequences, (ii) mouse second exons, and (iii) a combined model of human and mouse second exons. These models were used to search all the contigs of the University of California Santa Cruz Golden Path assembly (<http://genome.cse.ucsc.edu/goldenPath/01apr2001>), and the July 30, 2001 NCBI assembly (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens) using the *hmmsearch* program. The assemblies, once downloaded to a local file server, were translated into all six reading frames for searching, as HMMER does not currently have the native capability to search nucleotide databases. The same procedure was followed for all contigs from the mouse genomic sequence that contained β -defensin genes as revealed by BLAST analysis. A Linux cluster consisting of 16 dual-Pentium-III (550MHz) systems, each with 2 Gbytes of RAM, was used to perform the analysis in parallel. Default cutoff values were used for filtering the results of these queries, and we performed postprocessing of the results (using custom Perl programs) to remove motifs that did not contain the six-cysteine target motif. For each novel β -defensin gene identified using the *hmmsearch* program, additional iterative BLAST searches were performed against

the human and mouse databases to identify additional related sequences and search for expressed sequence tags to confirm that the sequences are transcribed.

Discovery of novel genomic sequences encoding β -defensin second exons

An HMM (47) developed from the mature peptide sequences of known β -defensin genes (HMMER software (47)) was first used to screen ~ 4 Mb of genomic DNA sequence around the β -defensin locus on human chromosome 8p23-p22 (12). This search found 11 genes, including the 5 known β -defensin genes (*DEFB101–104* and *SPAG11*) plus 6 β -defensin homologs (*DEFB105–109* and *DEFB132P*) (12). When these novel sequences were used as queries in a BLAST search of the available draft human genomic sequence, another β -defensin gene, *DEFB110*, was identified. Interestingly, *DEFB110* mapped to chromosome 6p12, whereas all previously identified human defensin genes localized to chromosome 8p23-p22 (45, 54). We rebuilt the HMM with a comprehensive set of predicted peptide sequences from the new genes and used it to analyze the genomic DNA sequence derived from the bacterial artificial chromosome (BAC) clone that contains *DEFB10* at the 6p21 locus (GenBank accession number NT_007402). This analysis revealed four more β -defensin genes (*DEFB111–114*) (12). Therefore, the *DEFB110–114* genes represented a second β -defensin gene cluster. Subsequent iterations of the BLAST/HMMER process identified 15 additional human β -defensins (*DEFB115–129*) (12). These genes were mapped to two sequence contigs, one on chromosome 20q11.1 and one on chromosome 20p13, and represent two more β -defensin gene clusters.

To analyze the entire human genome with HMMER, all 31 human α -defensin genes were used for the HMM. This search revealed 40 more sequences with the potential to encode the six-cysteine motif. However, in only two cases were the position of the cysteines and the cationic charge density consistent with β -defensin genes, *DEFB130* and *DEFB131* (1). In addition, these two genes were located on the same BAC clone that mapped ambiguously to multiple regions of the genome including chromosome 8p23-p22. As the sequence of this clone is not contiguous with the 8p23-p22 contigs containing the known β -defensin locus, it may represent a fifth cluster in the human genome. Significantly, this HMMER search detected only 13 of the 31 previously identified β -defensin genes. These results suggest that, like BLAST searches, genome-wide searches with HMMER alone are not sufficient for

identifying all β -defensin genes. Further BLAST and HMMER analyses did not detect additional sequences in the human genome. In total, this iterative, computational method discovered 28 novel putative human β -defensin genes. An outline of the general application of BLAST and HMMER to identify putative β -defensin coding regions is shown in Fig. 1.

To search for novel β -defensin second-exon motifs in the mouse genome, this hybrid process was also used to screen assembled mouse sequence contigs in the Celera database (55) (Fig. 1). This approach identified 43 new homologs (12) (Fig. 2) in 5 clusters, on chromosomes 1, 8, 14, and two separate clusters on chromosome 2. Of note, these regions are syntenic to the regions in the human genome that contain the β -defensin clusters (<http://www.ncbi.nlm.nih.gov/>

homology), indicating that the syntenic clusters have a common origin. Supporting this hypothesis, we found the highest sequence similarity between gene products from syntenic clusters. Moreover, the order and orientation for many genes in the syntenic clusters are conserved (Fig. 2). This syntenic relationship between the human and mouse β -defensin clusters and the conservation of genes within them suggests that each pair originated from a common ancestral gene cluster (40, 45). To look for evidence of transcription for these predicted genes, we performed BLAST searches against the six-frame translation of the dbEST or performed PCR analysis of cDNA libraries (J.L.B., B.C.S., P.B.M., unpublished data). At least one EST was found for 24 human and 23 mouse predicted genes (Fig. 2), representing all clusters.

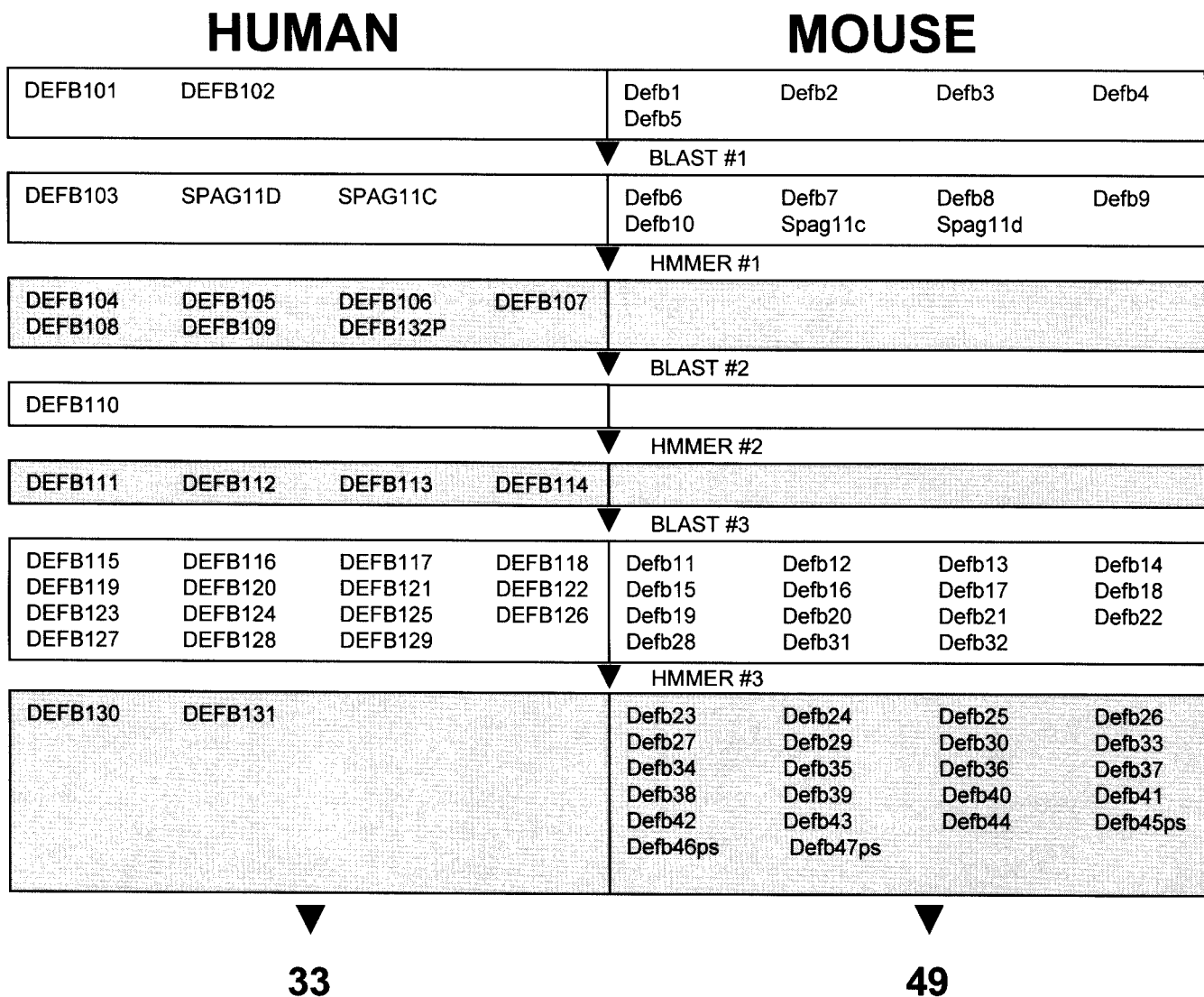


Fig. 1. A schematic summary of the iterative BLAST/HMMER search strategy used for β -defensin gene discovery in the human and mouse genomes.

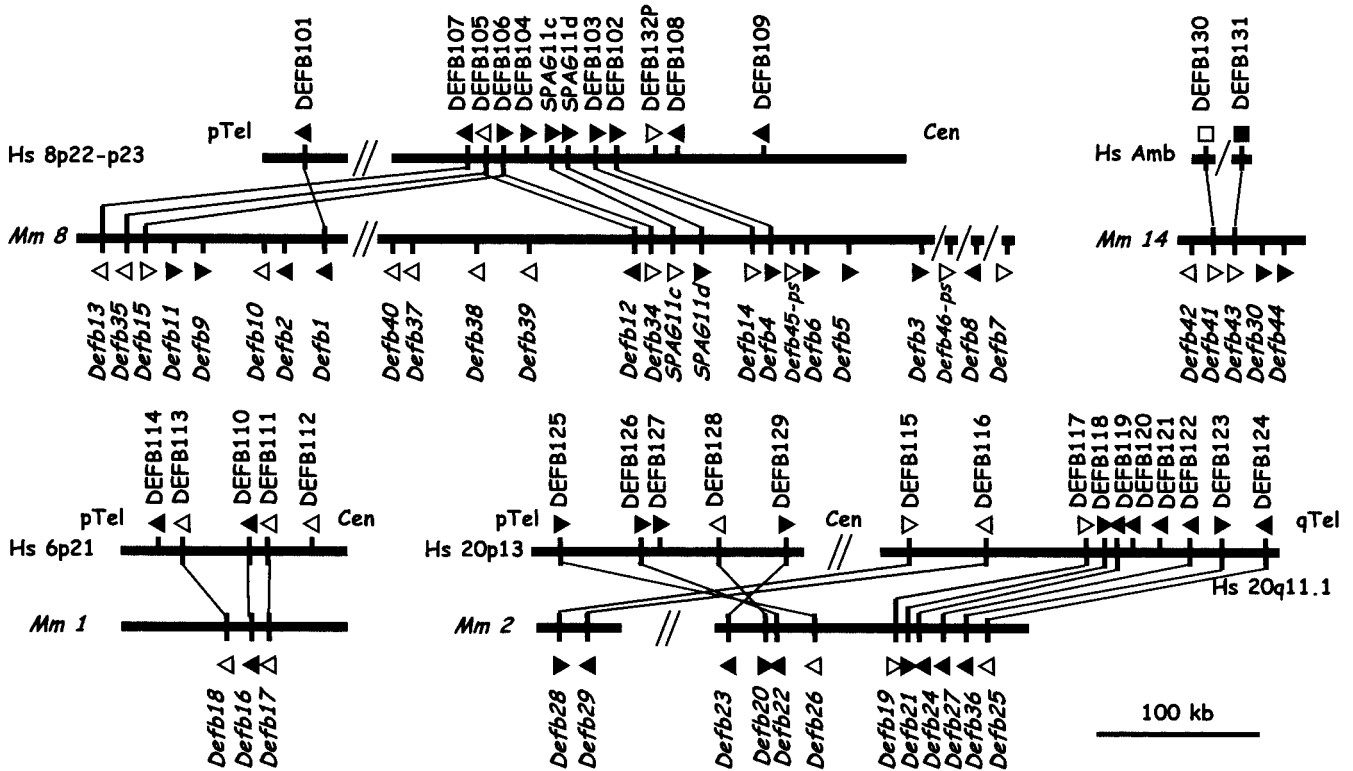


Fig. 2. Genomic structure of the five conserved human and mouse β-defensin gene clusters. Horizontal bars represent assembled genomic DNA sequence contigs [see Schutteet al. (12) for details] from the indicated human (Hs) and mouse (Mm) chromosomes. Slanted lines denote gaps in the genomic sequence of known (single) or unknown (double) length. The telomere (Tel) and centromere (Cen) orientation of human DNA sequence contigs was deduced from the position of genetic markers. The orientation of the mouse DNA sequence contigs was determined from the alignment of human and mouse gene homologs. Predicted direction of transcription is indicated by the orientation of the arrowheads. For those genes with confirmation of expression from the EST databases or PCR screening of cDNA libraries, the arrowheads are filled. Genes that mapped ambiguously are indicated by open squares. The thin lines connect human and mouse genes with highest sequence similarity.

Genomic search for β-defensin first exons

As one method of validating the second exons identified using the iterative search strategy just described, attempts were subsequently made to identify their respective first exons. This was done by building an HMM of the first-exon amino acid sequence for the known β-defensins. The search then proceeded in a manner similar to the strategy employed to identify the second exons. Candidate first-exons were then required to fall within a 10 000 amino acid region upstream of any second exon identified in the search described earlier.

This search proved significantly more problematic than the second-exon search due to the relative lack of specificity in the first-exon amino acid sequence. Table 1 presents a partial list of the sequences used to build the first-exon model. The sequences are presented in a multiple-sequence alignment format to indicate the positionally relevant data. A cursory examination shows that the information content of the first-exon sequences is low; they are rich in leucine (51) and

hydrophobic residues. Assuming equal codon frequencies throughout the genome, one would expect tens of thousands of similar subsequences within the human genome sequence. In agreement with this expectation, the initial first-exon search returned thousands of candidate first-exons throughout the genome. After applying the 10 000 amino acid restriction, this list of candidate first-exons was reduced to fewer than 100 candidate first-exons, averaging about two or three per second-exon. As a final filtering step, only the three best first-exon candidates were prioritized for validation using molecular methods.

There are 21 mouse β-defensin second exons for which

Table 1. Translated exon 1 sequences

Gene	Sequence
DEFB123	MKLLLLLTLTVL LLLS QLTP—
DEFB126	MKSLFFTLAVFM LLAQLVS—
SPAG11d	MRQRLLP SVTSL LLLVALL FPGSS—
DEFB104	MQRLLV LLLA I S LLLY QDLPG—
DEFB106	MRTFLFLF AVLFL FTTP—

the corresponding first-exon sequence is known, from EST database searches or from published literature. Of these, the HMMER-based search correctly identified 20. The only first exon not identified was located in an unfinished BAC. Only 200 bp were available 5' of the second exon, which was insufficient to identify the first exon. The molecular confirmation of the additional candidate human and mouse first exons identified using HMMER is ongoing.

The computational approaches outlined in this review provide a proof-of-principle for a novel genome-wide search strategy that identifies genes with conserved structural motifs. This strategy highlights the complementary nature of the BLAST and HMMER analysis tools and demonstrates their potential synergy for mining genomic databases and identifying new members of gene families. None of the novel sequences identified was included in past annotations of the human and mouse genomes, indicating that they were not previously identified.

Problems of computational gene discovery methods

Several pitfalls await the naïve application of this strategy utilizing genome-scale searches. These include the inadvertent selection of models that may span multiple exons and the use of weak models that detect many false-positive matches. Perhaps most significantly, however, are the potential ambiguities in the current assemblies of the genome.

Models that span multiple exons cause problems when HMMER is used with default parameters. By default, a global alignment is used to align the HMM model to the translated genome. This can be addressed by using the **-f** option during model construction. This allows a model to align partially, rather than requiring a global alignment to the entire model. In our analysis, the β -defensin six-cysteine motif was conveniently contained within a single exon. This does not necessarily occur for other recognizable protein domains or motifs. In such instances, it may be beneficial to implement a filtering procedure to weight short, strong-identity hits over longer hits with the same score.

Lack of specificity within a model is another issue that must be addressed. As shown in Table 1, the first exon model of β -defensins lacks a strong consensus protein motif. The defining features are a positively charged amino acid at the second residue and a leucine-rich stretch of 15–20 amino acids. Such lack of specificity invariably leads to many false-positive hits in the genome that will require further curation upon application of HMMER. This problem necessitated further modifications, including the addition of positional requirements for the first exon search, as described earlier.

The final issue to be addressed is the potential for problems arising from unfinished or improperly assembled genomic sequences. This issue is due in part to the inclusion of many unfinished BAC sequences, which has a severe impact on searching with multiexon domains, as regions of sequence across gaps may or may not be properly ordered or oriented. In addition, duplicate second exons were identified for several of the known β -defensins. These duplicates most probably result from assembly-derived duplications. Clearly, care must be exercised when using an unfinished sequence in a genome-wide search strategy. Another concern pertains to the assembly of the genomic sequences. During assembly, errors of various types can lead to inaccurate (or confusing) results. As an example, in the last few genome assemblies there has been a consistent duplication at the contig level across chromosomes 4, 8, 11 and 12. This region contains two putative defensins (*DEFB130* and *DEFB131*). Each successive genome release reduces the extent of the duplication. In previous releases, the duplication spanned an entire BAC. Currently, the duplication is approximately 5 kb and no longer includes chromosome 12.

Alternative search strategies

An alternative method to search for other β -defensin gene family members would be the use of the PSI-BLAST tool suite (<http://www.ncbi.nlm.nih.gov/blast/>). PSI-BLAST, a form of hybrid BLAST and probabilistic model, is used to search a database with a modified version of the BLAST search algorithm. PSI-BLAST is provided a query (consensus) sequence and an alignment of additional example sequences (presumably from the same domain or gene family) with which to create a search model. The alignment is used to bias a standard substitution matrix (e.g. PAM100, BLOSUM62, etc.) for each position in the consensus sequence. However, position-independent gap penalties are used. Hmmbuild (47), on the other hand, creates a sequential machine model of the alignment and, during the search process, is able to account very accurately for variations in sensitivities to gaps in alignments. These are very important for cross-species comparisons, in which gaps between motifs may vary significantly without altering the underlying primary functional role of the gene/protein. In addition to differences in the ways in which the probabilistic nature of the alignment is modeled in PSI-BLAST and HMMER, the overall procedure employed in our search used BLAST to selectively identify new members based on the members added to the 'family' in the previous HMMER search. While PSI-BLAST mirrors aspects of this iterative pro-

cess, the difference is significant, and in the case of the β -defensin family, we found the results to differ substantially. In particular, we were unable to identify all the defensin family members using PSI-BLAST alone, regardless of which family member was used as the query sequence.

Subdivision for specificity

One strategy that was particularly successful in identifying β -defensins was the designing of several models around gene cluster-specific members. This strategy works, because the β -defensins within syntenic gene clusters are more similar to each other than to β -defensins in other chromosomal loci. Although this strategy typically does not identify members that could otherwise escape detection, it does score them differently. This approach allows rapid identification of probable novel β -defensins.

Future directions

The discovery of a much larger family of β -defensin genes provides a new starting point for investigating their role in

innate immunity. The success in the combined application of BLAST and HMMER to identify new β -defensin family members suggests that similar strategies could be applied to address other issues of interest to the field of immunology. The approach outlined is well suited to search for novel members of protein families distinguished by a common protein domain and for searches across species. We speculate that novel sequences in other genes families not identified using conventional strategies might be detected using this approach. For example, such a strategy could be readily applied to search on a genome-wide scale for novel α - or θ -defensin encoding motifs, additional members of pattern recognition protein families, or new chemokines (56). Many other tools of computational genomics are also available to address questions of interest to immunology (57, 58). One example is the use of expression profiling by microarray hybridization to investigate gene regulation (59, 60). Furthermore, the completion of the genome sequencing of multiple vertebrate species will facilitate computational comparative studies of the structures of key gene families involved in innate and adaptive immunity. Such an analysis may lead to new insights into the conserved elements in noncoding regions flanking known genes that are important in gene regulation.

References

1. Boman HG. Innate immunity and the normal microflora. *Immunol Rev* 2000;**173**:5–16.
2. Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA. The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in *Drosophila* adults. *Cell* 1996;**86**:973–983.
3. Lemaitre B, Reichhart J, Hoffmann JA. *Drosophila* host defense: differential induction of antimicrobial peptide genes after infection by various classes of microorganisms. *Proc Natl Acad Sci USA* 1997;**94**:14614–14619.
4. Lemaitre B, et al. A recessive mutation, immune deficiency (*imd*), defines two distinct control pathways in the *Drosophila* host defense. *Proc Natl Acad Sci USA* 1995;**92**:9465–9469.
5. Medzhitov R, Preston-Hurlburt P, Janeway CA. A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. *Nature* 1997;**388**:394–397.
6. Akira S, Takeda K, Kaisho T. Toll-like receptors: critical proteins linking innate and acquired immunity. *Nat Immunol* 2001;**2**:675–680.
7. Werner T, Liu G, Kang D, Ekengren S, Steiner H, Hultmark D. A family of peptidoglycan recognition proteins in the fruit fly *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 2000;**97**:13772–13777.
8. Gottar M, et al. The *Drosophila* immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* 2002;**416**:640–644.
9. Ramet M, Manfruelli P, Pearson A, Mathey-Prevot B, Ezekowitz RA. Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for *E. coli*. *Nature* 2002;**416**:644–648.
10. Kang D, Liu G, Lundstrom A, Gelius E, Steiner H. A peptidoglycan recognition protein in innate immunity conserved from insects to humans. *Proc Natl Acad Sci USA* 1998;**95**:10078–10082.
11. Liu C, Xu Z, Gupta D, Dziarski R. Peptidoglycan recognition proteins: a novel family of four human innate immunity pattern recognition molecules. *J Biol Chem* 2001;**276**:34686–34694.
12. Schutte BC, et al. Discovery of five conserved defensin gene clusters using a computational search strategy. *Proc Natl Acad Sci* 2002;**99**:2129–2133.
13. Lehrer RI, Ganz T. Defensins of vertebrate animals. *Curr Opin Immunol* 2002;**14**:96–102.
14. Zasloff M. Antimicrobial peptides of multicellular organisms. *Nature* 2002;**415**:389–395.
15. Ganz T, Weiss J. Antimicrobial peptides of phagocytes and epithelia. *Semin Hematol* 1997;**34**:343–354.
16. Tang YQ, et al. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. *Science* 1999;**286**:498–502.
17. Harder J, Bartels J, Christophers E, Schroder J-M. A peptide antibiotic from human skin. *Nature* 1997;**387**:861–862.
18. Harder J, Bartels J, Christophers E, Schroder JM. Isolation and characterization of human beta-defensin-3, a novel human inducible peptide antibiotic. *J Biol Chem* 2001;**276**:5707–5713.

19. Mathews MS, et al. Production of beta-defensin antimicrobial peptides by the oral mucosa and salivary glands. *Infect Immun* 1999;**67**:2740–2745.
20. Jia HP, et al. Discovery of new human beta-defensins using a genomics-based approach. *Gene* 2001;**263**:211–218.
21. McNamara NAR, Tuchin OS, Fleiszig SM. Ocular surface epithelia express mRNA for human beta defensin-2. *Exp Eye Res* 1999;**69**:483–490.
22. O'Neil DA, et al. Expression and regulation of the human beta-defensins hBD-1 and hBD-2 in intestinal epithelium. *J Immunol* 1999;**163**:6718–6724.
23. Valore EV, Park CH, Quayle AJ, Wiles KR, McCray PB Jr, Ganz T. Human β -defensin-1, an antimicrobial peptide of urogenital tissues. *J Clin Invest* 1998;**101**:1633–1642.
24. Zucht HD, et al. Human β -defensin-1: a urinary peptide present in variant molecular forms and its putative functional implication. *Eur J Med Res* 1998;**3**:315–323.
25. Diamond G, Zasloff M, Eck H, Brasseur M, Maloy WL, Bevins CL. Tracheal antimicrobial peptide, a cysteine-rich peptide from mammalian tracheal mucosa: peptide isolation and cloning of a cDNA. *Proc Natl Acad Sci USA* 1991;**88**:3952–3956.
26. Goldman MJ, Anderson MG, Stolzenberg ED, Kari PU, Zasloff M, Wilson JM. Human β -defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis. *Cell* 1997;**88**:1–9.
27. McCray PB Jr, Bentley L. Human airway epithelia express a β -defensin. *Am J Respir Cell Mol Biol* 1997;**16**:343–349.
28. Bals R, et al. Human β -defensin 2 is a salt-sensitive peptide antibiotic expressed in human lung. *J Clin Invest* 1998;**102**:874–880.
29. Singh PK, et al. Production of β -defensins by human airway epithelia. *Proc Natl Acad Sci USA* 1998;**95**:14961–14966.
30. Schutte BC, McCray PB Jr. Beta-defensins in lung host defense. *Annu Rev Physiol* 2002;**64**:709–748.
31. Yang D, et al. Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. *Science* 1999;**286**:525–528.
32. Selsted ME, et al. Purification, primary structures, and antibacterial activities of beta-defensins, a new family of antimicrobial peptides from bovine neutrophils. *J Biol Chem* 1993;**268**:6641–6648.
33. Hoover DM, et al. The structure of human beta-defensin-2 shows evidence of higher order oligomerization. *J Biol Chem* 2000;**275**:32911–32918.
34. Schibli DJ, et al. 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.
35. Lehrer RI, Barton A, Daher KA, Harwig SS, Ganz T, Selsted ME. Interaction of human defensins with *Escherichia coli*. Mechanism of bactericidal activity. *J Clin Invest* 1989;**84**:553–561.
36. Wimley WC, Selsted ME, White SH. Interactions between human defensins and lipid bilayers: evidence for formation of multimeric pores. *Protein Sci* 1994;**3**:1362–1373.
37. White SH, Wimley WC, Selsted ME. Structure, function, and membrane integration of defensins. *Curr Opin Struct Biol* 1995;**5**:521–527.
38. Bensch KW, Raida M, Magert HJ, Schulz-Knappe P, Forssmann WG. hBD-1: a novel beta-defensin from human plasma. *FEBS Lett* 1995;**368**:331–335.
39. Tarver AP, et al. Enteric β -defensin: molecular cloning and characterization of a gene with inducible intestinal epithelial cell expression associated with cryptosporidium parvum infection. *Infect Immun* 1998;**66**:1045–1056.
40. Jia HP, et al. A novel murine beta-defensin expressed in tongue, esophagus, and trachea. *J Biol Chem* 2000;**275**:33314–33320.
41. Zhao C, Nguyen T, Liu L, Sacco RE, Brogden KA, Lehrer RI. Gallinacin-3, an inducible epithelial beta-defensin in the chicken. *Infect Immun* 2001;**69**:2684–2691.
42. Morrison GM, Davidson DJ, Dorin JR. A novel mouse beta defensin, Defb2, which is upregulated in the airways by lipopolysaccharide. *FEBS Lett* 1999;**442**:112–116.
43. Huttner KM, Kozak CA, Bevins CL. The mouse genome encodes a single homolog of the antimicrobial peptide human β -defensin 1. *FEBS Lett* 1997;**413**:45–49.
44. Jia HP, et al. Molecular cloning and characterization of rat genes encoding homologues of human beta-defensins. *Infect Immun* 1999;**67**:4827–4833.
45. Liu L, Zhao C, Heng HHQ, Ganz T. The human-beta-defensin-1 and alpha-defensins are encoded by adjacent genes: two peptide families with differing disulfide topology share a common ancestry. *Genomics* 1997;**43**:316–320.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–410.
47. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–763.
48. Thompson JD, Higgins DG, Gibson TJ, Clustal W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**:4673–4680.
49. Garcia JR et al. Human beta-defensin 4: a novel inducible peptide with a specific salt-sensitive spectrum of antimicrobial activity. *FASEB J* 2001;**15**:1819–1821.
50. Kirchoff C, Osterhoff C, Habben I, Ivell R, Kirchoff C. Cloning and analysis of mRNAs expressed specifically in the human epididymis. *Int J Androl* 1990;**13**:155–167.
51. 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.
52. Bals R, et al. Mouse beta-defensin 3 is an inducible antimicrobial peptide expressed in the epithelia of multiple organs. *Infect Immun* 1999;**67**:3542–3547.
53. Yamaguchi Y, et al. A novel mouse β -defensin, mBD-6, predominantly expressed in skeletal muscle. *J Biol Chem* 2001;**276**:31510–31514.
54. Bevins CL, Jones DE, Dutra A, Schaffzin J, Muenke M. Human enteric defensin genes: chromosomal map position and a model for possible evolutionary relationships. *Genomics* 1996;**31**:95–106.
55. Marshall E. Genome sequencing. Celera assembles mouse genome; public labs plan new strategy. *Science* 2001;**292**:822–823.
56. Yoshie O, Imai T, Nomiya H. Chemokines in immunity. *Adv Immunol* 2001;**78**:57–110.
57. Fahrner AM, Bazan JF, Papanthasiou P, Nelms KA, Goodnow CC. A genomic view of immunology. *Nature* 2001;**409**:836–838.
58. Aderem A, Hood L. Immunology in the post-genomic era. *Nat Immunol* 2001;**2**:373–375.
59. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–470.
60. Derisi J, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;**14**:457–460.