Research

Bacterial α_2-macroglobulins: colonization factors acquired by horizontal gene transfer from the metazoan genome?

Aidan Budd*, Stephanie Blandin*, Elena A Levashina† and Toby J Gibson*

Addresses: *European Molecular Biology Laboratory, 69012 Heidelberg, Germany. †UPR 9022 du CNRS, IBMC, rue René Descartes, F-67087 Strasbourg CEDEX, France.

Correspondence: Toby J Gibson. E-mail: toby.gibson@embl.de

Published: 26 May 2004

Genome ***Biology*** 2004, **5:**R38

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/6/R38

Received: 20 February 2004
Revised: 2 April 2004
Accepted: 8 April 2004

Abstract

Background: Invasive bacteria are known to have captured and adapted eukaryotic host genes. They also readily acquire colonizing genes from other bacteria by horizontal gene transfer. Closely related species such as *Helicobacter pylori* and *Helicobacter hepaticus*, which exploit different host tissues, share almost none of their colonization genes. The protease inhibitor α_2-macroglobulin provides a major metazoan defense against invasive bacteria, trapping attacking proteases required by parasites for successful invasion.

Results: Database searches with metazoan α_2-macroglobulin sequences revealed homologous sequences in bacterial proteomes. The bacterial α_2-macroglobulin phylogenetic distribution is patchy and violates the vertical descent model. Bacterial α_2-macroglobulin genes are found in diverse clades, including purple bacteria (proteobacteria), fusobacteria, spirochetes, bacteroidetes, deinococcids, cyanobacteria, planctomycetes and thermotogae. Most bacterial species with bacterial α_2-macroglobulin genes exploit higher eukaryotes (multicellular plants and animals) as hosts. Both pathogenically invasive and saprophytically colonizing species possess bacterial α_2-macroglobulins, indicating that bacterial α_2-macroglobulin is a colonization rather than a virulence factor.

Conclusions: Metazoan α_2-macroglobulins inhibit proteases of pathogens. The bacterial homologs may function in reverse to block host antimicrobial defenses. α_2-macroglobulin was probably acquired one or more times from metazoan hosts and has then spread widely through other colonizing bacterial species by more than 10 independent horizontal gene transfers. *yfhM*-like bacterial α_2-macroglobulin genes are often found tightly linked with *pbpC*, encoding an atypical peptidoglycan transglycosylase, PBP1C, that does not function in vegetative peptidoglycan synthesis. We suggest that YfhM and PBP1C are coupled together as a periplasmic defense and repair system. Bacterial α_2-macroglobulins might provide useful targets for enhancing vaccine efficacy in combating infections.

Background

The broad-spectrum protease inhibitor α_2-macroglobulin (α_2M) and the complement factors C3, C4 and C5 belong to a gene family present in all metazoans ranging from corals to humans. These large (approximately 1,500 residue) proteins all undergo proteolytic processing and structural rearrangement as part of their role in host defense. The family is characterized by a unique thioester motif (CxEQ; single-letter amino-acid code), and a propensity for multiple conformationally sensitive binding interactions [1], which define their functional properties. The highly reactive thioester bond is buried inside the molecule in the native protein, protected from precocious inactivation [2]. Upon proteolytic cleavage, the thioester bond becomes exposed and can then mediate covalent attachment to activating self and non-self surfaces, in the case of complement factors, or covalent or noncovalent crosslinking to the attacking proteases in the case of α_2Ms [3]. The proteolytic activation of these proteins also mediates interactions with receptors.

In contrast to complement factors, which are activated by specific 'convertase' protease complexes, α_2Ms have an accessible 'bait' region with target sites for many proteases. The rearrangement of α_2M that follows cleavage of the bait region entraps the attacking protease in a cage-like structure, hindering protein substrates from reaching the protease active site [4]. In this way, exported proteases that are essential for parasitic infections can be rendered ineffective by α_2M entrapment [5-7]. Protease-reacted α_2M is then cleared from circulation by binding to the receptor CD91, triggering endocytosis. In addition, α_2Ms bind cytokines and growth factors and regulate their clearance and activity [8,9].

Vertebrate complement factors C3, C4 and C5 are part of an activation cascade that leads to the assembly of the membrane-attack complex and lysis of the pathogen. Binding of C3 also targets pathogens for phagocytosis. Proteolytic activation of all three complement proteins yields anaphylatoxins (cleaved amino-terminal fragments) which are recognized by specific receptors and activate the inflammatory response at the site of infection. In contrast to α_2Ms, complement factors also possess a carboxy-terminal domain extension, the netrin or NTR module (PFAM:PF01759) [10]. Some members of the complement/α_2M family (for example, C5 and ovostatin) have lost the thioester motif.

No α_2M-related proteins have been found in any eukaryotes outside metazoans. Within the Metazoa, representatives have been found in all species examined, with a so-called 'C3-like' protein sequenced from the cnidarian *Swiftia exserta* (SWISS-PROT acc:Q8IYP1). There is no information from sponges as yet. We may speculate that the gene family evolved in an early metazoan in response to challenge from invasive microorganisms exploiting the new niche provided by the interstitial spaces and body cavities. The more derived role of the complement factors, together with their extra netrin domain, suggests that they arose by gene duplication from an ancestral α_2M-like gene. Apart from vertebrates, α_2M-group proteins have been most actively studied in arthropods. The horseshoe crab *Limulus* has a plasma α_2M that is a component of an ancient invertebrate defense system; it is able to inhibit a wide range of proteases as well as to modulate plasma cytolytic activity [11]. *Limulus* α_2M forms tetramers, binding covalently across the multimers rather than to the attacking proteases, but still traps these in a cage-like structure after proteolytic activation [12]. In dipteran insects, there are multiple α_2M homologs, the thioester-containing proteins (TEPs). The *TEP* genes have been amplified by a process of tandem duplication into linked multigene families. *Drosophila melanogaster* has six *TEP* genes, whereas the mosquito *Anopheles gambiae* has 15 [13]. It is thought that the impressive expansion of *TEP* genes in the mosquito might be linked to the parasitic challenge provided by its blood-sucking lifestyle [13]. The first characterized TEP in mosquitoes, TEP1, binds to and promotes phagocytosis of bacteria [14]. TEP1 also binds to *Plasmodium berghei* and mediates its killing [15]. Thus the complement/α_2M protein family is part of an innate immune system in metazoans that long predates the immunoglobulin-based immune system of vertebrates, yet remains vital for combating parasites in all animal lineages examined.

While reviewing the distribution of α_2M/TEP proteins from invertebrates [16], we conducted BLAST searches of the protein databases and were surprised to discover a number of bacterial sequences with BLAST E-values indicating homology with α_2M. Given the absence of α_2Ms in all non-metazoan eukaryotic lineages, it immediately seemed clear that horizontal gene transfer (HGT) of α_2Ms must have occurred between metazoans and bacteria. But which way? Here we summarize the evidence for numerous horizontal transfers between bacterial lineages and discuss some biochemical and medical implications of the finding.

Results

Our BLAST2SRS server provides the species in the BLAST output page: this is useful for quick visual surveys of the taxonomic distribution of a protein family. A BLAST2SRS search with human α_2M unexpectedly listed an entry (SWISS-PROT accession number Q9X079) with E-value 2.3e-8 from *Thermotoga maritima*, a thermophilic eubacterium. With a length of 1,538 residues, a signal sequence and a matching CxEQ motif, there was no doubt that this was a genuine α_2M homolog. Numerous other bacterial sequences with lower E-values but obvious topological equivalence were also listed: for example, *Escherichia coli* YfhM (P76578) at 5.8e-5; *Pseudomonas putida* AAN66197 at 1.3e-4; *Rhizobium meliloti* Q92VA6 at 5.0e-3. Profile searches with a metazoan α_2M alignment and subsequently with an alignment of the stronger bacterial hits revealed a number of additional, highly diverged homologs, some lacking the CxEQ. For example, *E.*

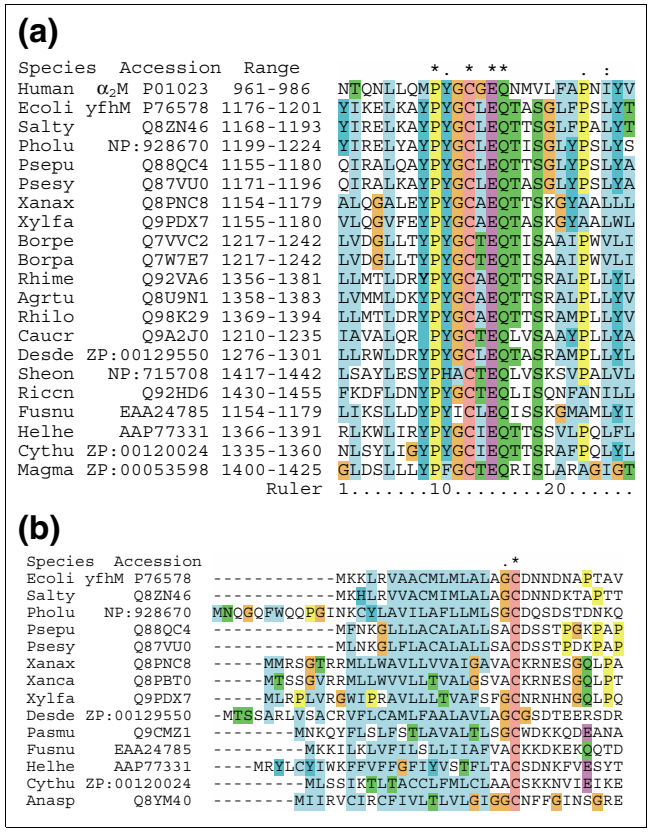

Figure 1
Sequence alignments. **(a)** Alignment detail of YfhM group bacterial α_2-macroglobulin sequences from bacterial proteomes plus human α_2-macroglobulin (α_2M), centred on the conserved CxEQ thioester motif. **(b)** Alignment of selected bacterial α_2-macroglobulin signal peptides possessing the conserved cysteine (C) residue. Signal peptides require a run of hydrophobic residues preceded by a positively charged residue. Cleavage is at the small (glycine (G)/alanine (A)) residue terminating the signal peptide (marked by a dot). Aminoacylation of lipoproteins occurs in the inner membrane at a C (marked by *) directly following the signal peptide. An aspartate residue (D) after the C acts as a retention signal to the inner membrane in *E. coli*, preventing lipoprotein transfer to the outer membrane [17,18]. Alignments are color-coded using the Clustal X defaults [66]. Blue denotes conserved hydrophobicity, as in the signal peptide, while a strongly conserved C is colored pink. Accession numbers are SWISS-PROT or NCBI genomes (NP, finished genome; ZP, provisional assignment in unfinished genome). Species names follow the SWISS-PROT convention.

coli has a second divergent homolog, YfaS (P76464). It is noteworthy that not a single instance of an archaeal α_2M sequence could be found. Thus α_2M-like sequences are restricted to eubacteria and metazoans. No function has been experimentally ascribed to any of the bacterial α_2Ms (bact-α_2Ms).

Bacterial α_2-macroglobulin sequences

Figure 1a shows an alignment of the segment spanning the CxEQ motif for a representative set of bacterial α_2M homologs. Not all bact-α_2Ms possess the CxEQ motif. Using *E. coli* as the reference, YfhM is the archetype of a large group, mostly with the thioester motif, and YfaS is the archetype of a smaller, diverged group always lacking the motif. The sequences of the YfhM group are sufficiently divergent that accurate alignment proved time-consuming, but was achieved over almost the whole sequence length, other than the highly variable amino termini. We did not attempt to align together the YfhM and YfaS groups and the metazoan α_2Ms. This would only be useful if the trees would be informative, but the high divergence between the groups precludes accurate alignment, leading to unreliable tree calculation. (In future, given more YfaS sequences and α_2Ms from more metazoan lineages and a solved three-dimensional structure to guide alignment, this might be worth revisiting.) One feature apparent in many of the aligned YfhM sequences is a conserved cysteine directly following the signal peptide (Figure 1b), indicating palmitoylation. The presence of an aspartic acid residue following the palmitoylated cysteine has been shown in *E. coli* to dictate sorting to the inner membrane [17,18], in which case YfhM will be found in the periplasmic space, attached to the inner membrane. Given the CxEQ motif, covalent trapping of proteases in the periplasmic space seems to be the most likely function (whether the covalent links are to the trapped protease or between the α_2M multimers, as in the horseshoe crab *Limulus* [12]). The YfaS group of bact-α_2Ms lack a palmitoylable cysteine, so may be secreted, while absence of the CxEQ motif indicates the molecular function must be different, at least in part, though this does not, of itself, rule out protease entrapment, as in chicken ovostatin which also lacks the reactive thioester motif [19].

Genomic context of bacterial α_2-macroglobulins

A survey of completely sequenced bacterial genomes was undertaken to establish which lineages possessed bact-α_2Ms and which did not. Representative results are summarized in Figure 2. It is clear that there is a highly inconsistent correlation of bact-α_2M possession and phylogenetic relationship, except for very closely related species.

Bact-α_2Ms are absent from the full proteomes of the following anciently diverged free-living species: the hyperthermophilic chemolithoautotroph *Aquifex aeolicus*, the thermophilic photolithoautotroph *Chlorobium tepidum*, the cyanobacteria *Synechocystis*, *Synechococcus* and *Prochlorococcus*, all firmicutes including *Bacillus subtilis*, all actinobacteria including *Streptomyces coelicolor*, the β-proteobacterium *Nitrosomonas europaea* and the δ-proteobacterium *Geobacter metallireducens*. Furthermore, possession of bact-α_2M is inconsistently represented within clades such as the proteobacteria, spirochetes and cyanobacteria. This is well illustrated by the two species of *Helicobacter*, one exploiting the acidic stomach and the other the very different environment of the liver: only the latter has a bact-α_2M. The *H. hepaticus* genome lacks essentially all the proposed *H. pylori* virulence factors and is believed to possess a quite different set, adapted to its hepatobiliary habitat [20]. The irregular phylogenetic correlation suggests that bact-α_2Ms are

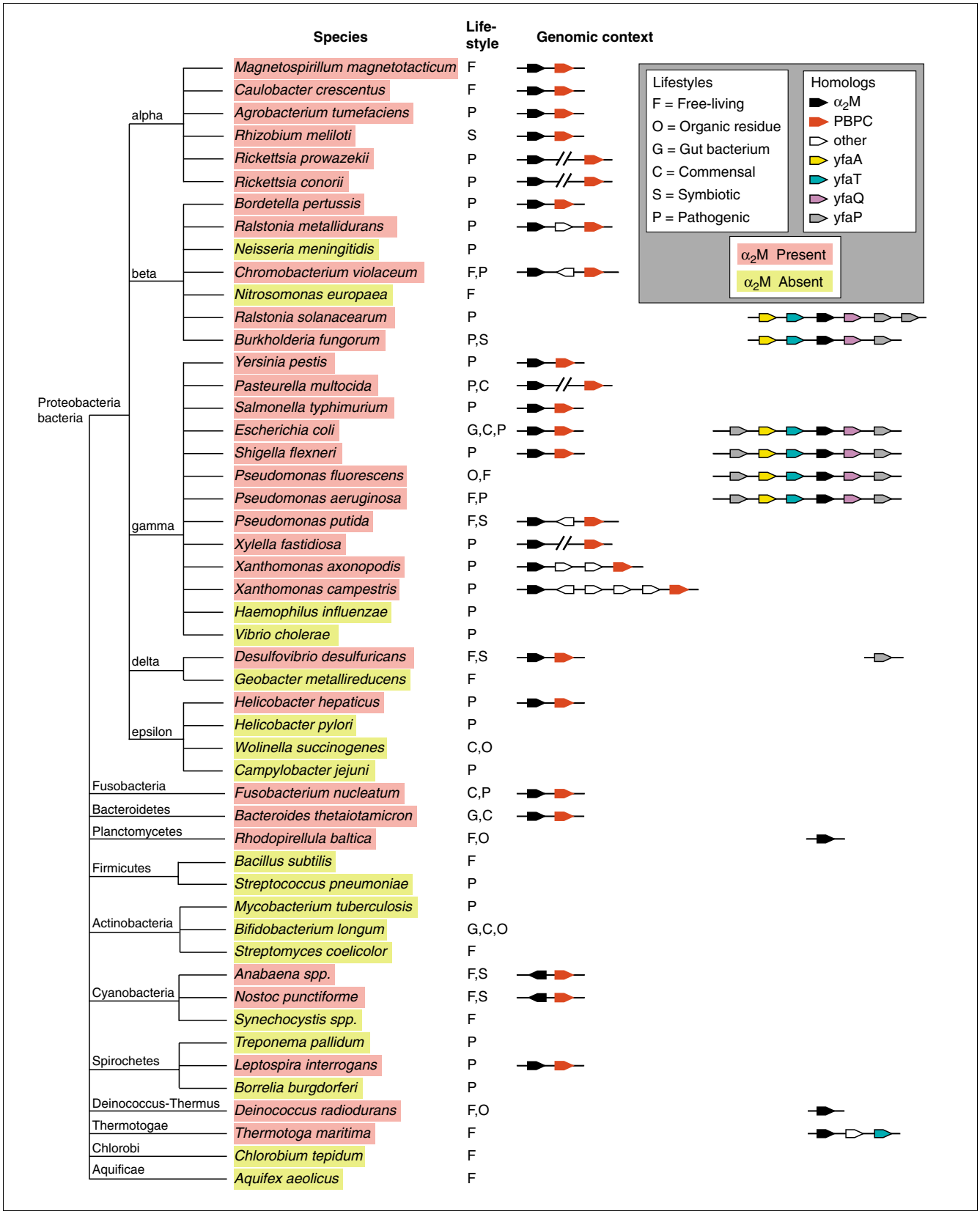

Figure 2 *(see legend on next page)*

Figure 2 *(see previous page)*
Phylogenetic distribution of bacterial α_2-macroglobulin homologs (α_2M). Pink, species that possess bacterial α_2-macroglobulin genes; yellow, species without bacterial α_2-macroglobulin genes. Shared genomic context is indicated for genes found to co-occur with bacterial α_2-macroglobulin genes. Because bacterial phylogeny has many uncertainties, the tree is simplified into multiple nodes representing three levels of divergence. There is little phylogenetic consistency for bacterial α_2-macroglobulin possession. Colonizing proteobacteria are overwhelmingly expected to have a bacterial α_2-macroglobulin gene, although exceptions occur, notably *Helicobacter pylori*, *Vibrio cholerae* and *Neisseria meningitidis*. No examples of bacterial α_2-macroglobulin genes have been found in colonizing Gram-positives in the Firmicutes or Actinobacteria, which include such major infectious clades as streptococci and mycobacteria. *Anabaena* is a facultative plant symbiont, while other free-living cyanobacteria (here represented by *Synechocystis*) lack bacterial α_2-macroglobulin. *Thermotoga maritima*, *Magnetospirillum magnetotacticum* and *Caulobacter crescentus* are the only species possessing bacterial α_2-macroglobulin for which no apparent connection exists with niches linked to exploitation of higher eukaryotes. Genome context of bacterial α_2Ms is based on automated STRING annotation [21], supplemented by re-analysis of individual genomes. Double slanted bars between genes indicate that they are not tightly linked. Bacterial α_2-macroglobulins make up two distinct groups typified by the *E. coli* genes *yfhM* and *yfaS*. The members of the *yfhM* group (on the left side of the figure) almost always co-occur with *pbpC* and are often, but not always, found adjacent to and on the same strand as one another in an operon configuration. Members of the *yfaS* group (grouped on the right side of the figure), when present in β- or γ-proteobacteria, are linked to four other gene families. All their predicted gene products also possess signal peptides, but are otherwise of unknown function. In other taxa, members of the *yfaS* group of bacterial α_2-macroglobulins are either unassociated with any of these gene families (planctomycetes and deinococci), or linked to a member of just one of the families (thermotogae).

'lifestyle' genes, affecting which niches a bacterium is able to exploit. Although an association with colonization seems clear (Figure 2), there is a strong bias in bacterial genome sequencing in favor of pathogenic species: this currently precludes a statistical assessment and might create a misleading phylogenetic perspective.

The STRING server [21] was used to check for neighboring genes that persistently co-occur with bact-α_2Ms. Using either *yfhM* or *yfaS* as seed, STRING reported two conserved gene sets that are widely found with bact-α_2Ms. The results are summarized in Figure 2. The *yfhM* group always co-occurs with *pbpC*, which encodes penicillin-binding protein 1C (PBP1C). The gene topology is almost always consistent with *pbpC* and *yfhM* being in the same operon (or co-transcribed from a bidirectional promoter, as in *Anabaena*). The more strongly an operon structure is conserved across species, the more likely are the encoded proteins to have associated functions [22]. Moreover, products of conserved gene pairs very often associate physically [23]. Therefore, if YfhM is involved in colonizing or pathogenic lifestyles, so should be its partner. PBP1C is a paralog of the periplasmic cell-wall biosynthesis proteins PBP1A and PBP1B, though with the addition of a carboxy-terminal non-enzymatic domain of approximately 100 residues (PFAM:PF06832). The PBP1A and PBP1B peptidoglycan synthases each have two enzymatic domains, an amino-terminal transglycosylase and a carboxy-terminal transpeptidase (reviewed in [24]). Although it possesses the two enzymatic domains, studies have shown that PBP1C does not substitute for these proteins in cell-wall biosynthesis during vegetative growth [25]: indeed deletion of *pbpC* has a weak phenotype not affecting cell viability in the laboratory, although the number of peptide crosslinks is increased [25]. The transpeptidase domain in PBP1C is thought not to bind to most of the β-lactams that inhibit the paralogous enzymes, nor to be a functional transpeptidase [25]. One curious finding is that, *in vitro*, PBP1C accounts for 75% of transglycosylase activity, yet is responsible for only 3% of *de novo* peptidoglycan biosynthesis in the cell [25]. As PBP1C does not substitute for the biosynthetic enzymes, a possible role would be in emergency repairs to the peptidoglycan, where its efficient transglycosylase activity would be appropriate.

The *yfaS* group of bact-α_2Ms is likewise usually found in a candidate operon, at least within the proteobacteria (Figure 2), in this case with four other gene families, defined by the *E. coli yfaA*, *yfaQ*, *yfaP* and *yfaT* genes. All these genes have signal sequences and their encoded proteins are expected to be secreted or periplasmic, but, otherwise, sequence analysis has yielded no clues to their function. It is possible that all the encoded proteins function to disrupt or resist host defenses. The YfaS-like bact-α_2Ms of the free-living and highly divergent *Thermotoga*, *Deinococcus* and *Rhodopirellula* (none of which is known to be invasive) are not found associated with most of these other genes.

Microarray expression data

The STRING server was also used to check for any significant coexpression of *yfhM*, *yfaS* and other members of the two candidate operons, using *E. coli* data from the Stanford microarray database [26]. All the genes associated with those for bact-α_2Ms are present in the experiments included in the STRING database, and are expressed at levels significantly above background. However, none of the genes exhibits coordinated variation in expression levels either with each other or with any other genes in the *E. coli* genome under the conditions investigated.

Calculation of sequence trees

An initial rough tree calculated from an alignment of *yfhM* family sequences gave strong indications that several horizontal transfers had occurred among the available set. As *yfhM* is always found together with *pbpC*, indicating that the paired genes should have a shared phylogenetic history, a quick check of the PBP1C tree was also done. The two trees, which provide controls for each other's topologies, were very similar, indicating that the apparent HGTs were unlikely to be artifacts. Therefore, we undertook a more careful

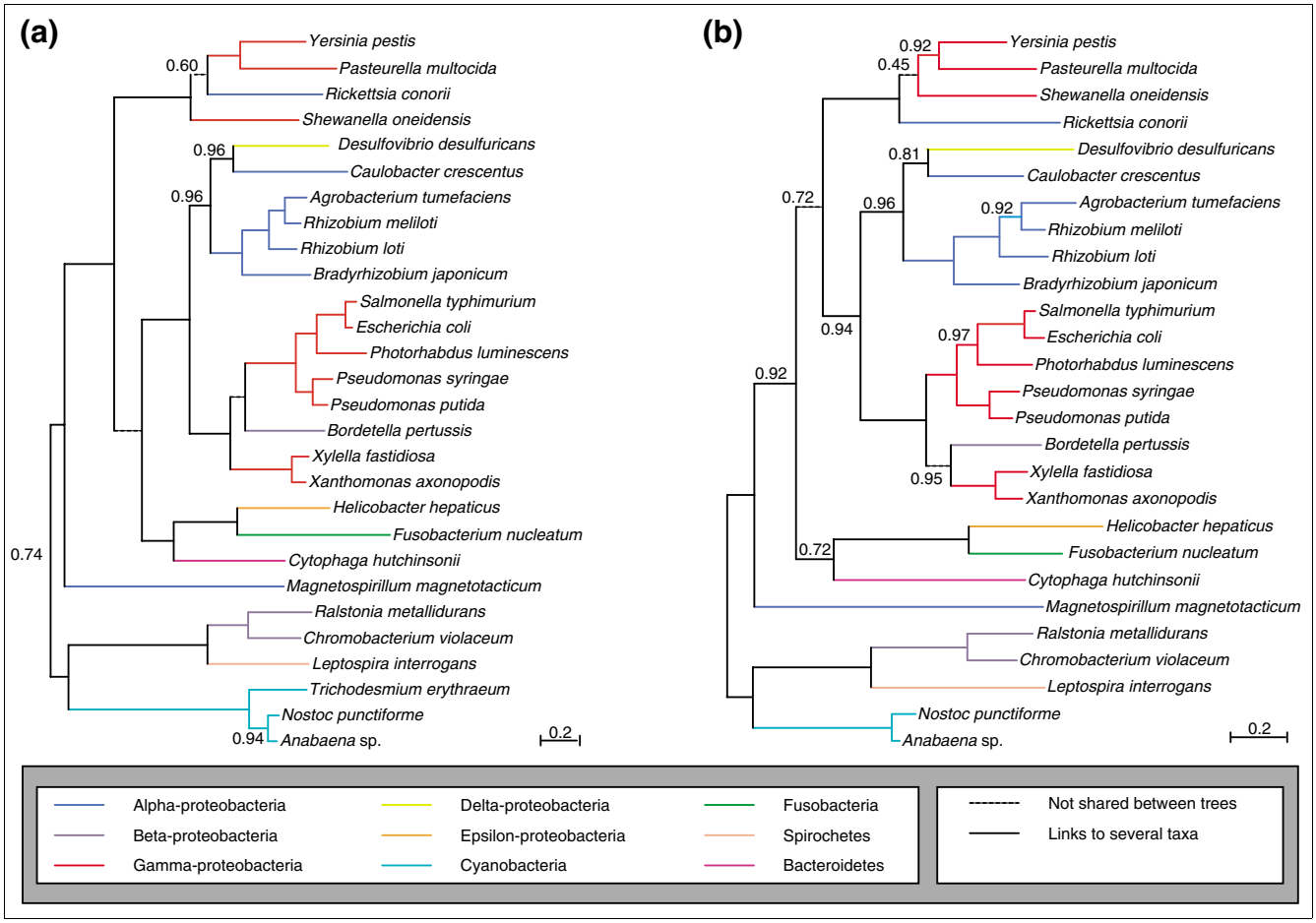

Figure 3
Trees calculated from amino-acid sequence alignments. **(a)** The YfhM group of bacterial α_2-macroglobulins; **(b)** the PBP1Csthat always co-occur and are usually found adjacent in the same operon. As shown by the key, branches are color-coded by taxon for easy visualization of phylogenetic inconsistencies. All branches have Bayesian posterior probabilities of 1.0 (that is, are completely stable during resampling) unless otherwise indicated. Three branches not shared between the trees are indicated by dotted lines: all other branches are congruent. The roots of the trees are not known, so the time vector of deep internal branches is not clear. See Materials and methods for details of the tree calculation.

phylogenetic analysis with a view to improving the phylogenetic signal-to-noise ratio and using a method that is less prone to rate variation artifacts than neighbor-joining.

Alignments were reviewed and edited by hand, then processed to remove especially noisy segments, as outlined in Materials and methods. Trees were calculated with MrBayes, a Bayesian resampling protocol that is now widely adopted [27]: MrBayes approaches the quality of maximum-likelihood methods while being quicker to calculate (though still computationally demanding). Results of the tree calculations are presented in Figure 3. The two trees differ by only three branch placements, indicating that the topologies are mostly sound, except for a few branches with low support (low posterior probabilities). As the calculated trees are unrooted, the ordering of the deepest branches cannot be mapped onto time.

Fitting the observed tree topologies to the vertical descent model

The number of ancestral genes required to explain an observed tree topology can be determined by embedding the sequence tree within a species tree. We prepared a species tree for the bacterial species in Figure 3 such that currently uncertain affinities were assigned in favor of the observed trees: this will provide a minimum estimate of ancestral gene number. The sequence tree topology was embedded into the bacterial species tree using GeneTree [28]. The reconciled tree required six gene-duplication events and 29 lineage-specific deletions. The last common ancestor (LCA) of the full set had a minimum of three genes, the LCA of the proteobacteria had four genes, while the LCA of the α/β-proteobacteria had six genes. The tree reveals a tendency for increasing gene number over time when vertical descent has strictly occurred.

The problems of the vertical descent model are manifold. First, all sequenced extant genomes have single copies of the *yfhM/pbpC* genes, yet vertical descent shows a progression toward increasing gene number over time. This requires late but fully independent massive gene loss to have occurred in all lineages. Second, the observed robust sequence tree topologies would require a clear affinity between cyanobacteria and spirochetes, an affinity that has hitherto gone entirely unnoticed in the field of bacterial phylogeny. Third, the number of events (gene duplications and deletions) found to be required under a model of vertical descent is based on a species tree chosen to minimize this number (see Materials and methods.) As the species tree used is unlikely to be accurate in places where bacterial phylogeny is unresolved, the number of such events required under a vertical descent model is probably greater than described (and hence, correspondingly less likely.)

Although bizarre evolutionary scenarios can always be invoked, the given tree topologies are difficult to explain solely by vertical descent from a common ancestral eubacterium.

Horizontal transfers of the *yfhM* and *pbpC* gene couplet

Difficulties in accounting for the observed YfhM and PBP1C trees disappear if it is assumed that a number of horizontal gene transfers have occurred. Vertical transmission then only occurred among some sets of quite closely related bacteria. There are four deeply diverged sets within the tree, which will be discussed in turn.

The major proteobacterial grouping

Of the 22 proteobacterial species sampled, 18 are exclusively grouped together in the two trees. The species are all plant or animal pathogens and symbionts - even the anaerobic sulfate-reducing *Desulfovibrio desulfuricans* is a symbiont of deep-sea hydrothermal vent polychete worms [29]. Sub-branches compatible with vertical descent are present for five α-proteobacteria including *Agrobacterium tumefaciens* and for seven γ-proteobacteria including *E. coli*. For bact-α_2M and PBP1C to have existed in proteobacteria before the α/γ split, these gene sequences would have to be evolving more slowly than in other parts of the tree. It is more likely that the genes spread via HGT through these groups some time ago and then have been vertically inherited (at least in part). The remainder of the grouping consists of unambiguous HGT, although the direction of transfer is not always clear-cut. The β-proteobacterium *Bordetella pertussis* has acquired the genes from a γ-proteobacterium. The δ-proteobacterium *D. desulfuricans* has acquired the genes from an α-proteobacterium. An outlier set of α- and γ-proteobacteria, including *Rickettsia conorii* and *Yersinia pestis*, indicate two further transfers, but in this case the order of the transfers is not determined. Therefore to create the topology of this grouping, a minimum of four unique horizontal transfers has occurred.

The bacteroidete/fusobacteria/ε-proteobacteria grouping

This group consists of three unrelated taxa which exploit niches related to the animal digestive system. The ε-proteobacterium *Helicobacter hepatica* colonizes mouse liver ducts, *Fusobacterium* species colonize the teeth, *Bacteroides thetaiotamicron* (not shown on the tree owing to an incomplete bact-α_2M sequence) is a major gut bacterium, while a second bacteroidete, *Cytophaga hutchinsonii*, exploits cellulose-rich animal waste. Horizontal transfer into the ε-proteobacterium *H. hepaticus* is clear-cut, as it is isolated on the trees from all other proteobacteria, whereas other *Helicobacter* lack these genes. Another transfer has occurred between fusobacterial and bacteroidete lineages, but the direction is not clear. A third HGT is likely to have originally introduced the genes into these lineages but cannot be formally assigned without a root.

The isolated Magnetospirillum *α-proteobacteria branch*

Magnetospirillum magnetotacticum bact-α_2M and PBP1C are deeply diverged from all other species, including other α-proteobacteria. This positioning away from its relatives indicates that HGT occurred into the *Magnetospirillum* lineage. The strong divergence from other sequences may indicate that the sequence has undergone rapid evolution. This latter point may be addressed in future if the branch becomes populated by some closer relatives.

The cyanobacteria/spirochete/β-proteobacteria grouping

This branch consists of three very unrelated taxa: cyanobacteria facultatively symbiotic with plants, spirochetes pathogenic to metazoans and a pair of closely related genera of β-proteobacteria that each include free-living, symbiotic and pathogenic forms. The deepest diverged in the group are the *Anabaena*-like symbiotic cyanobacteria. The economically significant *Anabaena-Azolla* symbiosis provides the nitrogen fixation that fertilizes paddy fields [30]. As other free-living cyanobacteria, such as *Synechococcus*, lack these genes, HGT into this lineage is very likely. The isolation of the *Ralstonia* and *Chromobacterium* clade from other proteobacteria also indicates HGT into their lineage. HGT for *Leptospira* (the causal agent of leptospirosis) is also indicated, as other spirochetes such as *Borrelia burgdorferi* (the causal agent of Lyme disease) and *Treponema pallidum* (the causal agent of syphilis) lack these genes. Thus, this set of genes that are clearly grouped together by molecular phylogeny, yet are found within very diverse taxa, appear to have been transmitted three times.

Discussion

Sifting the evidence for bacterial HGT

There is increasing evidence that HGT has had - and continues to have - a major role in the adaptation of organisms, especially prokaryotes, to exploiting new environments. Nevertheless, it is often hard to demonstrate HGT, and there is considerable confusion about how to do so. The default

hypothesis should remain vertical transmission unless there is good evidence for HGT. The over-hasty assignment of recent bacterial-to-vertebrate gene transfers, solely on the basis of BLAST E-values [31], has been firmly refuted [32,33]. Such premature HGT assignments have been surveyed and used to provide guidelines for evaluating HGT [34,35]. Sometimes the evidence is clear-cut, as when adaptive genes are carried on phage, plasmid or transposon. Inconsistent phylogenetic distribution may be evidence for HGT but must be carefully balanced against gene-loss models, recognizing that the two processes are not mutually exclusive. Phylogenetic trees only provide good evidence for HGT when branching is robust and clearly delimited by appropriate outgroups: the HGT must carry a diagnostic molecular evolutionary signal.

One of the best paradigms for investigating recent and ongoing HGT in parasitic prokaryotes is the γ-proteobacterium *Vibrio cholerae*, which acquired pathogenicity late in recorded history. Free-living *Vibrio* species are common, harmless aquatic microorganisms. The first recorded cholera pandemic occurred in 1817, the sixth and seventh occurred recently enough to be investigated with modern molecular techniques, and the eighth is probably underway now (see [36] for details). The basic pathogenicity genes *ctxAB*, which encode cholera toxin, lie within the genome of the filamentous phage CTXϕ [37]. Other pathogenicity gene 'islands' include the toxin-co-regulated pilus, needed for colonization, and the VSP-1 and VSP-2 islands, which appeared in strains of the seventh pandemic and are suggested to have been integral to that event [38]. The recent O139 serotype arose by wholesale replacement of the pre-existing gene cluster encoding lipopolysaccharide O side-chain synthesis, yielding an outer surface with a different architecture, less susceptible to pre-existing immunity [39]. Thus, pathogenic *V. cholerae* continues to adapt to the invasive lifestyle, to a large extent through HGT-mediated acquisition of new capabilities, including, but not limited to, better avoidance of host defenses. Although many of the functions encoded by the genes within pathogenic islands are not understood, their absence from the free-living *Vibrio* species is good evidence that they have been incorporated, and then conserved, because of a direct or indirect role in enhancing virulence. Even though it is a γ-proteobacterium, the genomic sequence data show that *V. cholerae* has not (re-)acquired a bact-α_2M gene. At least, not yet.

HGT of α_2-macroglobulin among colonizing bacteria

Our unexpected finding that α_2-macroglobulins, hitherto only known from metazoans, are widely present in eubacterial genomes has provided one of the most clear-cut examples of widespread HGT between extremely divergent bacterial taxa that can be monitored by molecular phylogenetic approaches. We have been able to infer a minimum of 11 independent HGTs for the major *yfhM* group among 27 sequences tested. Because this group always coexists with a second gene, *pbpC*, shared evolutionary history means the trees are controlled for topological consistency, so that the assignment of HGT is not in doubt. This work does not address an earlier evolutionary history preceding the link-up of this gene pair.

It is striking that all four deeply diverged groups in the trees include proteobacterial species. This alone clearly indicates that HGT has occurred. Because this is the most heavily researched bacterial taxon and provides most of the sequenced genomes, it is not yet clear whether other taxa will also show multiple independent acquisitions of bact-α_2M and *pbpC*. Currently, the trees show a minimum of 11 independent HGT events, even if the originating (but unknown) taxon were represented here. A twelfth HGT is indicated if bact-α_2M was originally captured from a metazoan (or vice versa). Extensive gene loss is also likely to have contributed to the phylogenetic distributions in Figure 2, particularly amongst the α-,β-, and γ-proteobacteria, where possession seems the default yet both vertical and horizontal transmission occur. Quite possibly, a cycle of gain-loss-gain has repeatedly occurred as strains adapt between colonization and free-living environments. The role of gene loss cannot be quantified with current data, but this may become possible in the future with more comprehensive genome coverage.

Where pathogenic bacteria and their eukaryotic hosts share related genes that appear to be transferred from one to the other, it is believed that the direction is overwhelmingly from the eukaryote to the bacterium. The failure to find phylogenetic evidence for bacterium-to-vertebrate gene transfers is consistent with this direction [32,33]. We expect that bact-α_2M was transferred from a metazoan host to a pathogenic bacterium, but this is not yet demonstrable and remains supposition. Given a simple early metazoan, where the germ cells would not be physically isolated from any bacterial infection, one can see how selection could act to fix a bact-α_2M gene transferred in the opposite direction, if bact-α_2M was originally bacterial. This issue may become resolvable in future given much more extensive phylogenetic coverage.

Bacterial α_2-macroglobulin in apparently free-living bacteria

Many bacterial taxa contain a plethora of strains adapted for free-living, symbiotic and pathogenic lifestyles. Examples include the *Ralstonia* and *Anabaena* genera adapted to plants, *Escherichia* and *Treponema* adapted to animals and pseudomonads adapted to both. Many free-living bacterial strains are also facultative colonizers. This creates some difficulty in cataloguing genes that are adapted to colonizing niches versus free-living: it is rarely certain whether an apparently free-living species never colonizes a higher organism, or is not part of a continuum of strains frequently exchanging lifestyle genes. Given this caveat, we reviewed all the currently completed genomes of bacteria that are not in any way known to have close associations with higher eukaryotes. The available set of Gram-positive bacterial genomes stand out as never possessing a bact-α_2M gene (see below).

Only three apparently free-living Gram-negatives (*Magnetospirillum*, *Caulobacter* and *Thermotoga*) have bact-α_2Ms while seven (*Aquifex*, *Chlorobium*, *Synechocystis*, *Synechococcus*, *Prochlorococcus*, *Nitrosomonas* and *Geobacter*) do not. Thus this crude estimate would suggest that possession of a bact-α_2M gene is associated with colonization, not as a core colonization factor, but as an accessory that enhances fitness for the colonization environment. Further, it may imply that the three 'free-living' species possessing a bact-α_2M gene have undocumented facultative symbiotic capabilities with higher eukaryotes.

Usage of host α_2-macroglobulin by invasive Gram-positive bacteria

The Gram-positive firmicutes and actinobacteria stand out as always lacking bact-α_2M genes (Figure 2). However, certain Gram-positives have found a more direct way to take advantage of α_2M proteins. Pathogenic *Streptococcus pyogenes* directly co-opt host α_2M for defense against host proteases through the cell-surface proteins GRAB and protein G [40,41]. As Gram-positive bacteria do not possess an outer membrane, defensive strategies are likely to differ from those of Gram-negatives. Invasive Gram-positives are found to coat themselves in a selected set of host proteins to obstruct host defenses. Streptococcal GRAB mutants that are unable to bind α_2M have attenuated virulence [40]. It seems remarkable that prokaryotes have evolved two totally independent strategies to take advantage of α_2M. On the one hand, Gram-positives are able to use the host's own protein, on the other, Gram-negatives have acquired their own gene. The clear implication is that α_2M functionality has a wide and general significance spanning many bacterial taxa.

Bacterial α_2-macroglobulin YfhM/PBP1C: a second line of defense?

The lipopolysaccharide (LPS) layer of the outer membrane of Gram-negative bacteria provides a first line of defense. The outer membrane barrier is sufficient to prevent the enzyme lysozyme from lysing Gram-negative bacteria in culture [42]. Under attack from host immunity and antimicrobial peptides [43], LPS can be disrupted or stripped away - for example, when released into the circulation, it can lead to septic shock [44] - leaving the peptidoglycan cell wall and inner membrane exposed. There is current interest in antibacterial strategies that endeavor to enhance lysozyme activity by co-administration with agents that disrupt the outer membrane, such as EDTA [42].

The following assumptions lead us to a hypothesis for YfhM bact-α_2M/PBP1C as a periplasmic defense system. First, bact-α_2M and PBP1C form a complex, probably through the carboxy-terminal non-enzymatic domain of PBP1C. Second, the complex resides in the periplasmic space, attached by acylation to the inner membrane. Third, bact-α_2M functions to entrap attacking proteases. Fourth, PBP1C is a transglycosylase that polymerizes glycan chains. Fifth, a periplasmic defense is only needed when the outer membrane has been breached and peptidoglycan is under attack.

The role of the bact-α_2M/PBP1C system is then perceived to be defense at, and repair of, peptidoglycan breaches induced by the host (Figure 4). PBP1C provides 75% of the transglycosylase activity *in vitro*, but only 3% of peptidoglycan biosynthesis *in vivo* [25]: it is a fast linear transglycosylase, ideal for traversing and repairing a breach. During repair it will, however, be exposed to attacking proteases and may be rapidly rendered dysfunctional. The role of bact-α_2M will be to entrap attacking proteases, protecting PBP1C and other periplasmic proteins such as the high-affinity lysozyme inhibitor Ivy in *E. coli* [45]. In this way, the fate of the invading bacterial cell will depend on the relative balance of the host's attacking forces versus the bacterial defense systems. Under an optimized host attack, such defenses would be rapidly overwhelmed but when (or where) the host is not well prepared, these defenses may serve to prolong colonization.

Potential experimental and medical applications

The *yfhM/pbpC* gene pair in bacteria not only suggests experimental research strategies, but may have medical potential to help combat pathogenic organisms. Predicted periplasmic location and complexing of bact-α_2M and PBP1C with each other (and any other periplasmic proteins) should be straightforward to investigate biochemically. Elucidation of the host proteases entrapped by bact-α_2Ms should reveal which host defense proteases are targeted at which parasites, leading to enhanced understanding of host defense mechanisms. Bact-α_2M-inhibited proteases should be directly active against pathogen proteins - or else act indirectly as, for example, do the proteases of the complement cascade. *PbpC* deletions should show increased sensitivity to lysozyme treatments and *pbpC/ivy* double mutants, yet more so.

The bact-α_2M/PBP1C proteins also provide targets for medical intervention, for example by training host immunity, the administration of anti-bact-α_2M monoclonal antibody or in combination therapies. Antibodies to bact-α_2Ms should act not just by promoting immune clearance but also to block the bact-α_2M activity, so that the host antibacterial proteases are unhindered. This dual effect may provide an enhanced prophylactic efficacy for vaccines that are augmented with extra bact-α_2M protein (probably as an inactive variant) or be directly invoked by targeted anti-bact-α_2M antibody administration for combating acute infection. PBP1C should also be rendered dysfunctional by specific antibodies, perhaps in combination with transglycosylase inhibitors such as the antibiotic moenomycin.

Conclusions

Bact-α_2Ms are spread widely amongst symbiotic and pathogenic bacteria. The implication is that protease inhibition is often an aid to colonizing higher eukaryotes. The major form

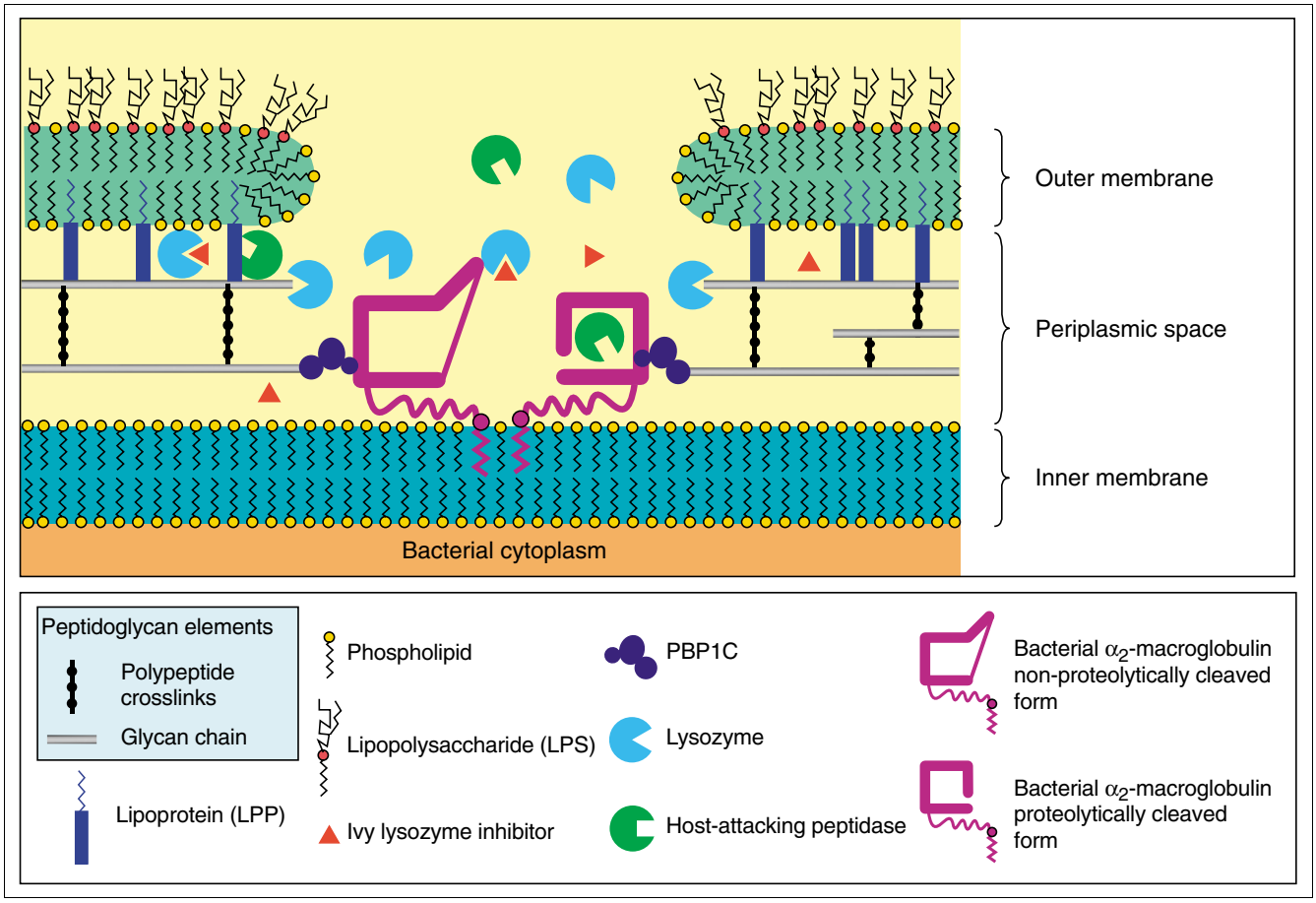

Figure 4
Schematic outline of the proposed defense of breaches of the bacterial outer membrane. Host systems (whether antimicrobial peptides, antibody and/or complement) have opened the outer membrane, allowing lysozyme and host proteases to attack periplasmic components, leading to a further breach of the peptidoglycan. Host attack is hampered by protease trapping (bacterial α_2-macroglobulin) and lysozyme inhibition (Ivy), giving PBP1C a chance to repair the glycan chains. The fate of the colonizing bacterial cell will now depend on whether the bacterial defenses are exhausted or the host attacking components are too limited to achieve cell lysis. Elements of the scheme are not drawn to scale.

of bact-α_2Ms is typified by *E. coli* YfhM and is a periplasmic protein that co-occurs with periplasmic PBP1C, a candidate peptidoglycan repair enzyme. The distribution of the *yfhM*/*pbpC* gene pair is inconsistent with the established bacterial phylogeny. Molecular trees calculated for each of the proteins are in good agreement with each other. Each tree provides a control for the other tree's topology, allowing confidence in the general topology. This allows us to state with high confidence that at least 11 separate gene transfers have occurred between highly diverged bacterial taxa. An additional gene transfer has occurred between bacteria and metazoans. We are not yet able to determine in which direction this transfer occurred, and therefore the title question is not yet answerable.

The known properties of α_2Ms and PBP1C point to a periplasmic line of defense at cell-wall breaches, mounted by the YfhM bact-α_2M and PBP1C. This defensive line should be sensitive to antibody-based therapeutic approaches, whether enhanced vaccine efficacy or direct administration of antibody.

Materials and methods

Sequence database searches

Bacterial α_2Ms were clearly revealed in a search of SWISSALL [46] using BLAST2SRS [47] in which the species names are included in the BLAST output [48]. Profile searches as described [49] using the EMBL Bioccelerators [50] supported and extended the findings and were used to retrieve a set of bacterial sequences. Reciprocal searches with bact-α_2M profiles reconfirmed the findings with good E-values (<1.e-25). The sets of proteomes provided by the BLAST server [51,52] at the National Center for Biotechnology Information (NCBI)

[53] were surveyed to determine the presence or absence of α_2Ms in bacteria and in non-metazoan eukaryotes.

Survey of genomic context

The STRING server [54] is a resource for exploring genome context (for example, identifying groups of genes found in close proximity in many different genomes [21]). Queries with bact-α_2Ms from *E. coli* or other bacteria yielded a recurring result: in most species the bact-α_2Ms cluster consistently with certain other gene families. This behavior is typical of gene sets belonging to the same operon. These families were retrieved and used for further database explorations, alignments and trees. To identify the location of these gene families in other genomes where linkage to bact-α_2Ms is less direct than those presented by STRING, we downloaded genomic database entries from the NCBI, converted the format of these files to EMBL using BioPerl [55], and assessed the location of the genes using Artemis [56]. In addition, linkage of these gene families was investigated in organisms not included in STRING using the same method.

Sequence alignment and editing

Sequences were aligned using Clustal X 1.83 [57]. Because many sequences are very dissimilar to each other, misaligned regions were to be expected. These were identified using the 'low scoring segments' check and either realigned using the 'realign selected range' option or were hand-edited in SeaView [58]. Corrections were assessed by both improvements to conserved hydrophobic columns (indicating structurally important residues) and with the 'low scoring segments' check. Sequences excluded because they were either too divergent to be aligned or may contain sequencing errors included *Deinococcus radiodurans* and *Bacteroides thetaiotamicron*.

Calculation of sequence trees

Preliminary trees were made by neighbor-joining [59] as implemented in Clustal X, excluding gaps and correcting for multiple substitutions with the Kimura PAM model. These initial trees indicated that HGT had occurred, warranting more careful assessment. Alignments were processed with the Gblocks server [60] (for the divergent bact-α_2Ms, the low stringency settings were used). Gblocks heuristically removes poorly conserved excessively divergent segments of alignments with low signal-to-noise ratio in order to enhance the phylogenetic signal [61]. Processed alignments were used to derive tree topologies using Bayesian inference of phylogeny as implemented by MrBayes v2.01 [27] with maximum-likelihood branch-length estimates provided by PUZZLE [62]. MrBayes was used with four heated chains over 250,000 generations, sampling every 20 trees. The likelihoods of these trees were examined to estimate the length of the burn-in phase, and all trees sampled 20,000 generations later than this point were used to create a consensus tree using the 50% majority rule. Both MrBayes and PUZZLE were used with the JTT model of amino-acid substitution [63], assuming the presence of invariant sites and using a gamma distribution approximated by four different rate categories to model rate variation between sites, estimating amino-acid frequencies from the alignment. Trees were displayed and rooted in Njplot [64].

Estimation of minimum *yfhM* gene number in the bacterial last common ancestor

The program GeneTree [28] was used to evaluate the cost of embedding the YfhM sequence tree in a bacterial species tree. To compute the minimum gene number required in the last common ancestor of the given bacterial set, we set the unresolved bacterial affinities to match the YfhM/PBP1C trees (that is, cyanobacteria and spirochetes form a clade, as do bacteroidetes and fusobacteria; within the proteobacteria, the subgroup affinities were allocated to minimize the number of duplications required in the observed trees). *Magnetospirillum* was excluded from the analysis as its position is not stable in the YfhM and PBP1C trees. Embedding the observed tree topology in this bacterial species tree yielded a reconciled tree requiring six duplication and 29 deletion events.

Microarray expression data

STRING was used to investigate the expression patterns of genes as detected by DNA microarray. The Stanford Microarray Database (SMD) [26,65] was used to verify that these genes were indeed spotted on the arrays used by STRING, and that the spots displayed intensities significantly higher than background levels.

Acknowledgements

We thank Christian von Mering and Lars Jensen for helpful discussions. We are grateful to Fotis Kafatos for his consistent support of the *Anopheles* TEP studies.

References

1. Chu CT, Pizzo SV: **alpha 2-macroglobulin, complement, and biologic defense: antigens, growth factors, microbial proteases, and receptor ligation.** *Lab Invest* 1994, **71:**792-812.
2. Salvesen GS, Sayers CA, Barrett AJ: **Further characterization of the covalent linking reaction of alpha 2-macroglobulin.** *Biochem J* 1981, **195:**453-461.
3. Law SK, Dodds AW: **The internal thioester and the covalent binding properties of the complement proteins C3 and C4.** *Protein Sci* 1997, **6:**263-274.
4. Sottrup-Jensen L: **Alpha-macroglobulins: structure, shape, and mechanism of proteinase complex formation.** *J Biol Chem* 1989, **264:**11539-11542.
5. Armstrong PB: **The contribution of proteinase inhibitors to immune defense.** *Trends Immunol* 2001, **22:**47-52.
6. Miyoshi S, Kawata K, Tomochika K, Shinoda S, Yamamoto S: **The C-terminal domain promotes the hemorrhagic damage caused by *Vibrio vulnificus* metalloprotease.** *Toxicon* 2001, **39:**1883-1886.
7. Woo PT: ***Cryptobia* (*Trypanoplasma*) *salmositica* and salmonid cryptobiosis.** *J Fish Dis* 2003, **26:**627-646.
8. Chu CT, Pizzo SV: **Interactions between cytokines and alpha 2-macroglobulin.** *Immunol Today* 1991, **12:**249.
9. Lysiak JJ, Hussaini IM, Webb DJ, Glass WF 2nd, Allietta M, Gonias SL: **Alpha 2-macroglobulin functions as a cytokine carrier to induce nitric oxide synthesis and cause nitric oxide-dependent cytotoxicity in the RAW 264.7 macrophage cell line.** *J Biol*

Chem 1995, **270:**21919-21927.
10. Banyai L, Patthy L: **The NTR module: domains of netrins, secreted frizzled related proteins, and type I procollagen C-proteinase enhancer protein are homologous with tissue inhibitors of metalloproteases.** *Protein Sci* 1999, **8:**1636-1642.
11. Armstrong PB, Quigley JP: **Alpha2-macroglobulin: an evolutionarily conserved arm of the innate immune system.** *Dev Comp Immunol* 1999, **23:**375-390.
12. Dolmer K, Husted LB, Armstrong PB, Sottrup-Jensen L: **Localisation of the major reactive lysine residue involved in the self-crosslinking of proteinase-activated *Limulus* alpha 2-macroglobulin.** *FEBS Lett* 1996, **393:**37-40.
13. Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, *et al.*: **Immunity-related genes and gene families in *Anopheles gambiae*.** *Science* 2002, **298:**159-165.
14. Levashina EA, Moita LF, Blandin S, Vriend G, Lagueux M, Kafatos FC: **Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, *Anopheles gambiae*.** *Cell* 2001, **104:**709-718.
15. Blandin S, Shiao S-H, Moita LF, Janse CJ, Waters AP, Kafatos FC, Levashina EA: **Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector *Anopheles gambiae*.** *Cell* 2004, **116:**661-670.
16. Blandin S, Levashina EA: **Thioester-containing proteins and insect immunity.** *Mol Immunol* 2004, **40:**903-908.
17. Fukuda A, Matsuyama S, Hara T, Nakayama J, Nagasawa H, Tokuda H: **Aminoacylation of the N-terminal cysteine is essential for Lol-dependent release of lipoproteins from membranes but does not depend on lipoprotein sorting signals.** *J Biol Chem* 2002, **277:**43512-43518.
18. Yamaguchi K, Yu F, Inouye M: **A single amino acid determinant of the membrane localization of lipoproteins in *E. coli*.** *Cell* 1988, **53:**423-432.
19. Nagase H, Harris ED Jr: **Ovostatin: a novel proteinase inhibitor from chicken egg white. II. Mechanism of inhibition studied with collagenase and thermolysin.** *J Biol Chem* 1983, **258:**7490-7498.
20. Suerbaum S, Josenhans C, Sterzenbach T, Drescher B, Brandt P, Bell M, Droge M, Fartmann B, Fischer HP, Ge Z, *et al.*: **The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*.** *Proc Natl Acad Sci USA* 2003, **100:**7901-7906.
21. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res* 2003, **31:**258-261.
22. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
23. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
24. Goffin C, Ghuysen JM: **Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs.** *Microbiol Mol Biol Rev* 1998, **62:**1079-1093.
25. Schiffer G, Holtje JV: **Cloning and characterization of PBP 1C, a third member of the multimodular class A penicillin-binding proteins of *Escherichia coli*.** *J Biol Chem* 1999, **274:**32031-32039.
26. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, *et al.*: **The Stanford Microarray Database: data access and quality assessment tools.** *Nucleic Acids Res* 2003, **31:**94-96.
27. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
28. Page RD: **GeneTree: comparing gene and species phylogenies using reconciled trees.** *Bioinformatics* 1998, **14:**819-820.
29. Cottrell MT, Cary SC: **Diversity of dissimilatory bisulfite reductase genes of bacteria associated with the deep-sea hydrothermal vent polychaete annelid *Alvinella pompejana*.** *Appl Environ Microbiol* 1999, **65:**1127-1132.
30. Hill DJ: **Pattern of development of *Anabaena* in *Azolla-Anabaena* symbiosis.** *Planta* 1975, **122:**179-184.
31. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
32. Salzberg SL, White O, Peterson J, Eisen JA: **Microbial genes in the human genome: lateral transfer or gene loss?** *Science* 2001, **292:**1903-1906.
33. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR: **Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates.** *Nature* 2001, **411:**940-944.
34. Genereux DP, Logsdon JM Jr: **Much ado about bacteria-to-vertebrate lateral gene transfer.** *Trends Genet* 2003, **19:**191-195.
35. Ragan MA: **Detection of lateral gene transfer among microbial genomes.** *Curr Opin Genet Dev* 2001, **11:**620-626.
36. Faruque SM, Mekalanos JJ: **Pathogenicity islands and phages in *Vibrio cholerae* evolution.** *Trends Microbiol* 2003, **11:**505-510.
37. Waldor MK, Mekalanos JJ: **Lysogenic conversion by a filamentous phage encoding cholera toxin.** *Science* 1996, **272:**1910-1914.
38. Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ: **Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease.** *Proc Natl Acad Sci USA* 2002, **99:**1556-1561.
39. Bik EM, Bunschoten AE, Gouw RD, Mooi FR: **Genesis of the novel epidemic Vibrio cholerae O139 strain: evidence for horizontal transfer of genes involved in polysaccharide synthesis.** *EMBO J* 1995, **14:**209-216.
40. Rasmussen M, Müller HP, Björck L: **Protein GRAB of *Streptococcus pyogenes* regulates proteolysis at the bacterial surface by binding alpha2-macroglobulin.** *J Biol Chem* 1999, **274:**15336-15344.
41. Sjobring U, Trojnar J, Grubb A, Akerstrom B, Björck L: **Ig-binding bacterial proteins also bind proteinase inhibitors.** *J Immunol* 1989, **143:**2948-2954.
42. Masschalck B, Michiels CW: **Antimicrobial properties of lysozyme in relation to foodborne vegetative bacteria.** *Crit Rev Microbiol* 2003, **29:**191-214.
43. Ganz T: **Antimicrobial polypeptides.** *J Leukoc Biol* 2003, **75:**34-38.
44. Wiese A, Gutsmann T, Seydel U: **Towards antibacterial strategies: studies on the mechanisms of interaction between antibacterial peptides and model membranes.** *J Endotoxin Res* 2003, **9:**67-84.
45. Monchois V, Abergel C, Sturgis J, Jeudy S, Claverie JM: ***Escherichia coli ykfE* ORFan gene encodes a potent inhibitor of C-type lysozyme.** *J Biol Chem* 2001, **276:**18437-18441.
46. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
47. **EMBL Blast sequence retrieval tool** [http://blast2srs.embl.de]
48. Bimpikis K, Budd A, Linding R, Gibson TJ: **BLAST2SRS, a web server for flexible retrieval of related protein sequences in the SWISS-PROT and SPTrEMBL databases.** *Nucleic Acids Res* 2003, **31:**3792-3794.
49. Thompson JD, Higgins DG, Gibson TJ: **Improved sensitivity of profile searches through the use of sequence weights and gap excision.** *Comput Appl Biosci* 1994, **10:**19-29.
50. **BIC web home page** [http://eta.embl-heidelberg.de:8000]
51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
52. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31:**28-33.
53. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/BLAST]
54. **STRING: functional association protein networks** [http://www.bork.embl-heidelberg.de/STRING]
55. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al.*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
56. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.
57. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31:**3497-3500.
58. Galtier N, Gouy M, Gautier C: **SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny.** *Comput Appl Biosci* 1996, **12:**543-548.
59. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.

60. **Gblocks server** [http://woody.embl-heidelberg.de/phylo/index.html]
61. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540-552.
62. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
63. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8:**275-282.
64. Perriere G, Gouy M: **WWW-query: an on-line retrieval system for biological sequence banks.** *Biochimie* 1996, **78:**364-369.
65. **SMD: home page** [http://genome-www.stanford.edu/microarray]
66. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.