Genome Analysis

The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system?

Mark J. Pallen

ESAT-6 is a small secreted protein of unknown function from *Mycobacterium tuberculosis* that is of fundamental importance in virulence and protective immunity. A PSI-BLAST search has identified distant homologues of ESAT-6 in more tractable bacteria, including *Bacillus subtilis*, *Bacillus anthracis*, *Staphylococcus aureus* and *Clostridium acetobutylicum*. The genes for ESAT-6-like proteins often cluster with genes encoding homologues of *B. subtilis* YukA. I speculate that the ESAT-6-like and YukA-like proteins form a novel Gram-positive secretion system potentially driven by the FtsK/SpoIIIE ATPase domains in the YukA-like proteins. The way is now open to investigate this hypothesis in organisms that are easier to manipulate than pathogenic mycobacteria.

Published online: 02 April 2002

More than one-third of the world's population has been infected with *Mycobacterium tuberculosis*, and this bacterium still kills more people than any other organism. ESAT-6 is a small protein (95 residues) that is found in short-term-culture filtrates of *M. tuberculosis* [1] and appears to be of fundamental importance in virulence and protective immunity. The ESAT-6 gene, *esx*, lies in the RD-1 region, which is deleted in all versions of the attenuated *Mycobacterium bovis* Bacille Calmette–Guérin (BCG) vaccine strain [2], and a defined *esx* mutant is attenuated in the guinea pig [3]. ESAT-6 is strongly immunodominant in the T-cell response to *M. tuberculosis* and this, together with its absence from BCG, has made it an excellent diagnostic reagent and a promising vaccine candidate [4,5].

The ESAT-6 family

Genome sequencing and other studies [6–9] have shown that ESAT-6 is a member of a family of ~100-residue proteins, 22 members of which are present in *M. tuberculosis*, some highly similar, others less well-conserved, with the sequence similarity focused on a central WXG motif. The genes encoding these proteins lie in tandem pairs within five large conserved clusters ('the ESAT-6 gene clusters') and six smaller clusters within the *M. tuberculosis* genome. Some clusters are deleted in attenuated strains (BCG and H37Ra) [2,10] and several ESAT-6 family members, like ESAT-6, are present in short-term-culture filtrates and immunodominant T-cell antigens [11]. A recent genome sequence analysis has shown that ESAT-6 gene clusters are also found in other mycobacteria, *Corynebacterium diphtheriae* and *Streptomyces coelicolor* [7].

Two unanswered questions

Despite the importance of the ESAT-6 family in *M. tuberculosis*, two questions remain unanswered: what is the function of these proteins and, given that they lack signal peptides, how are they exported from the cell? Progress on answering these questions has been hindered by the difficulty in manipulating mycobacteria and their close relatives and by a reported lack of ESAT-6 homologues in more tractable organisms. This prompted a search, using sensitive *in silico* methods, for ESAT-6-like proteins in other bacteria.

An iterative PSI-BLAST search [12] with ESAT-6 was performed on our ViruloGenome web site (http://www.vge.ac.uk) of a combined database of the NCBI's NR database and products predicted by GLIMMER (http://www.tigr.org/softlab/glimmer/glimmer.html) from unfinished genomes. (The settings were: BLOSUM80 matrix; expect value for inclusion in each iteration 0.05; filter off; and composition-based statistics on.) The genomic contexts of genes encoding new ESAT-6 homologues were scrutinized for relatives of other members of the ESAT-6 cluster and for other clues as to function. The domain structures of protein sequences derived from adjacent genes were analysed using PFAM (http://www.sanger.ac.uk/Pfam/). All ESAT-6 homologues were searched for signal peptides using SignalP (http://www.cbs.dtu.dk/services/SignalP-2.0/) and for coiled-coil domains using COILS (http://www.ch.embnet.org/software/COILS_form.html).

The ESAT-6 superfamily of WXG100 proteins

In line with previous reports [7–9], the PSI-BLAST search with ESAT-6 identified several dozen homologues within the actinobacteria. Surprisingly, however, the search, which continued for six iterations, also uncovered dozens of related proteins (Fig. 1) in various species of Gram-positive bacteria in the low-G+C group, including two proteins, YukE and YfjA, from the model organism *Bacillus subtilis*. Although reported as significant by PSI-BLAST, the levels of sequence similarity to ESAT-6 were very low (e.g. expect = 6e-09, 13% identity, 33% similarity for the ESAT-6/YfjA comparison). However, conservation of the WXG motif, of the protein length at ~100 residues and of a tendency for the associated genes to cluster with genes for other ESAT-6 family members lends credence to the inclusion of these proteins in an ESAT-6 superfamily of 'WXG100 proteins' (so-called after these conserved features). Furthermore, PSI-BLAST searches starting with some of the new WXG100 proteins (e.g. SA0271) linked ESAT-6 and YukE in as few as three iterations.

The *yukE* gene from *B. subtilis* encodes an 80-residue-protein and lies in the *yukABCDE* operon (Fig. 2), where *yukA* encodes a membrane-bound ATPase (as predicted by PFAM searches). *Bacillus halodurans* also possesses a *yukA–yukE*-like gene cluster (BH0975/BH0972), as does *B. anthracis*. Curiously, in addition to the *yukE* orthologue, among the GLIMMER-predicted products from the *B. anthracis* chromosome are three proteins that show amino-terminal similarity to the

  PII: S0966-842X(02)02345-4

Fig. 1. Multiple alignment of WXG100 proteins. Sequences obtained from the PSI-BLAST search were retrieved from the NCBI or the ViruloGenome databases. Sequences were aligned using ClustalW/Jalview then purged of redundancy at a cut-off of 50% identity to show a subset of the most diverse sequences. Alignments were then shaded using the BOXSHADE server, with positions highlighted if they showed identity (red) or similarity (blue) in 30% of sequences with ESAT-6. The first column shows the species of origin, the next the sequence name. The numbers show the position in the sequence at which the alignment starts. Entries starting VGE represent temporary arbitrary designations from the ViruloGenome database. All sequences are available in the supplementary Table that is available online at http://archive.bmn.com/supp/tim/esat/esat6.html

WXG100 proteins but which are longer than usual (289, 348 and 420 residues). The 410-residue protein pXO1-98, which is encoded by a gene in the anthrax toxin pathogenicity island on the *B. anthracis* pXO1 plasmid, also shows such similarity [13].

The solventogenic saprophyte *Clostridium acetobutylicum* possesses nine WXG100 proteins, encoded by genes in four clusters, three of which also contain *yukA* homologues. Nölling and colleagues [14] make passing reference to the presence of multiple YukE/YfjA-like proteins in both *C. acetobutylicum* and mycobacteria, suggesting they have independently uncovered the YukE/YfjA/ESAT-6 link.

A single *yukA–yukE*-like gene cluster was detected in all six *Staphylococcus aureus* genome sequences (Fig. 2) and in the two *Listeria* genome sequences. One truncated WXG100 protein sequence was identified from *Streptococcus gordonae*, and one WXG100-like sequence was found by visual inspection of the region around the *Streptococcus equi yukA* homologue. Analyses with the COILS program suggested that most WXG100 proteins possess extensive coiled-coil domains (Fig. 2). None has signal peptides.

YukA-like transporters

In every species in which full-length *esx/yukE*-like genes are found (including *M. tuberculosis*), at least some of them cluster with *yukA* homologues (Fig. 2). This clustering puts the link between the mycobacterial ESAT-6 family and the YukE/YfjA-like proteins beyond doubt. PFAM searches with the YukA-like proteins show that they all possess two or three FtsK/SpoIIIE domains. Single copies of this domain, which contains a putative ATP-binding P-loop motif, are found in the FtsK cell division protein from *E. coli* and in the SpoIIIE sporulation protein from *B. subtilis* (see PFAM entry PF01580), where they are thought to be involved in DNA translocation. However, this domain is a specific example of the more general AAA+ domain [15], common to many ATPases involved in macromolecular secretion, including those from Gram-negative type IV protein secretion systems (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR003593). Thus, it is reasonable to hypothesize that the FtsK/SpoIIIE domains in the YukA-like proteins are involved in generating energy for the secretion of WXG100 proteins.

Discussion and conclusions

Others have already speculated that proteins encoded by the ESAT-6 gene cluster might be components of a secretion system for the ESAT-6-like proteins [7–9]. This study strengthens and extends this hypothesis by finding new distant ESAT-6 homologues and by identifying the most conserved – and therefore probably most important – component of the cognate secretion system. It will now be possible to explore the general features of this system in

organisms that are much more tractable genetically and easier to handle in the laboratory than the mycobacteria, including the model organism, *B. subtilis*.

As the *yukA–yukE*-like gene clusters are found in several non-pathogenic species, neither the WXG100 proteins nor the YukA transporter system can be a specific adaptation to virulence. However, these proteins have been recruited to virulence functions in pathogenic mycobacteria, and probably also in the anthrax bacillus (note the *esx* homologue in the anthrax toxin pathogenicity island [13]), and perhaps even in the staphylococci. In both pathogenic mycobacteria and in *C. acetobutylicum* there have been parallel amplifications in the number of the *yukA–yukE*-like gene clusters.

If the ESAT-6-like WXG100 proteins were the primary targets of the YukA secretion system, it is hard to think of a common effector function useful to so many different bacterial lifestyles. A possible clue to the function of these proteins comes from Gram-negative protein secretion systems, where it is clear that secreted proteins are often components, rather than targets, of the translocation system. The WXG100 proteins are reminiscent of secretion system subunits, particularly in their size and possession of coiled-coil domains (a common feature of such proteins [16]). The involvement of FtsK/SpoIIIE domains in both DNA and protein transport is also reminiscent of Gram-negative type IV secretion systems [17]. One could thus speculate that the ESAT-6-like proteins are components of a surface-located assemblage involved in the secretion of other as-yet-unidentified proteins.

Several predictions can be made on the basis of this study that can now be addressed in the laboratory:

- The new WXG100 proteins will be secreted into culture supernatants.
- Such secretion will be abolished by non-polar deletions in the cognate YukA transporters (although it is unclear how much redundancy occurs when more than one YukA homologue is present in the cell).
- Protein–protein interactions are likely to occur between the YukA-like proteins and their cognate WXG100 proteins and/or between individual WXG100 monomers (there is experimental evidence of polymerization of ESAT-6 [1]). If coiled-coil domains mediate these interactions, then mutations of key hydrophobic residues will abrogate inter-molecular associations [18,19].
- If the analogies with Gram-negative systems hold up, then one might expect post-translational processing of the WXG100 proteins (c.f. pilin cyclization in type IV systems [17]), and the involvement of less-well-conserved proteins in chaperoning and presenting substrates or as additional structural components of the secretion system.
- Immunogold electron-microscopy studies might reveal that the ESAT-6-like proteins are components of a surface-located secretion apparatus [20]. However, by analogy with Gram-negative secretion, this apparatus might be assembled only under specific conditions (e.g. contact with host cells [20]) and could prove elusive (c.f. the mating channel in Gram-negative conjugation [17]).
- The new ESAT-6 homologues in *S. aureus* and *B. anthracis* are potential virulence factors, and could also prove useful immunodominant antigens.
- The effector proteins of the YukA-family secretion systems have yet to be identified and, by analogy with type III secretion, could be encoded by genes distant from the genes encoding the secretion system.

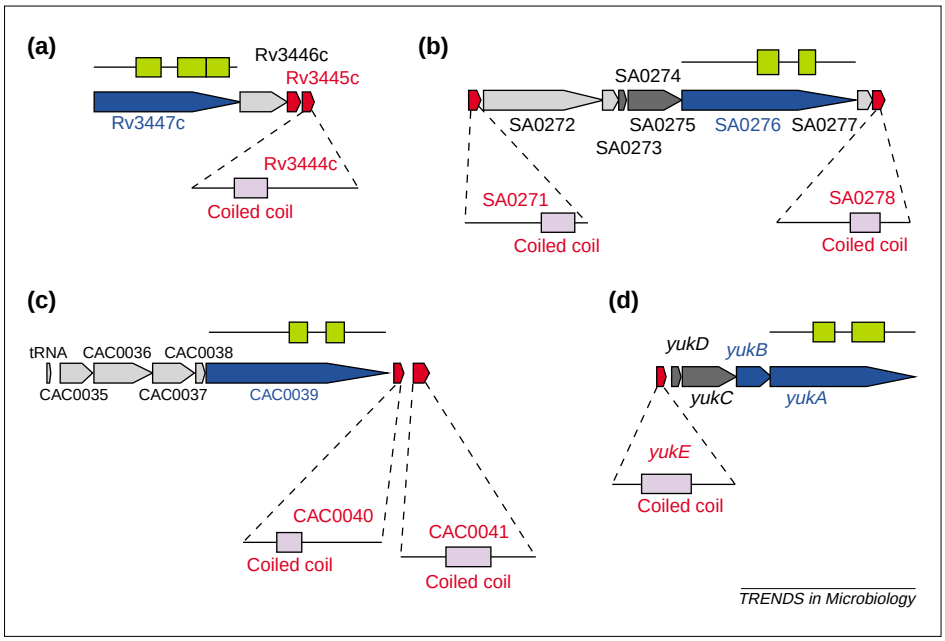

Fig. 2. Selected clusters of *yukA–yukE* homologues in (a) *Mycobacterium tuberculosis* H37Rv ESAT-6 cluster 4; (b) *Staphylococcus aureus* N315; (c) *Clostridium acetobutylicum* cluster 1; and (d) *Bacillus subtilis*. Filled arrows show the orientation and size of genes. *yukA* homologues are shown in blue and *yukE* homologues in red (note that *yukB* shows similarity to the amino terminus of some *yukA* homologues and thus is also shown in blue. Indeed, our preliminary sequencing studies suggest there is a frameshift in the published sequence and YukA and YukB form a single coding sequence). *yukCD*-related genes are shown in dark grey, and unrelated genes are shown in light grey. Regions encoding FtsK/SpoIIIE domains identified from PFAM searches are shown above genes as green boxes. Regions encoding coiled-coil domains in *yukE* homologues are depicted as lilac boxes in the expanded figures. Other YukAB homologues not shown include CAC00408 and CACO3709 from *Clostridium acetobutylicum*, and lino0054 and lmo0061 from *Listeria innocua* and *Listeria monocytogenes*, respectively.

In conclusion, this study has forged a compelling link between a crucial problem in mycobacteriology and the biology of other, more tractable Gram-positive bacteria, including both pathogens and a model organism, and highlights the potential of sequence analysis in hypothesis generation in post-genomic age.

Supplementary material

Results of PSI-BLAST searches with ESAT-6 and SA0271, together with a table of WXG100 proteins, can be found online at http://archive.bmn.com/supp/tim/esat/esat6.html

Acknowledgements

I thank Ajit Lalvani for stimulating my interest in ESAT-6. I thank the BBSRC for funding ViruloGenome and Alex Lam and Nick Loman for help setting it up. I gratefully acknowledge staff at The Wellcome Trust Sanger Institute,

The John Innes Centre, The Institute for Genomic Research, the University of Oklahoma and Genome Therapeutics for making incomplete genome sequence data publicly available.

References

1 Sorensen, A.L. *et al.* (1995) Purification and characterization of a low-molecular-mass T-cell antigen secreted by *Mycobacterium tuberculosis*. *Infect. Immun.* 63, 1710–1717

2 Gordon, S.V. *et al.* (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol. Microbiol.* 32, 643–655

3 Wards, B.J. *et al.* (2000) An *esat6* knockout mutant of *Mycobacterium bovis* produced by homologous recombination will contribute to the development of a live tuberculosis vaccine. *Tubercle Lung Dis.* 80, 185–189

4 Mollenkopf, H.J. *et al.* (2001) Protective efficacy against tuberculosis of ESAT-6 secreted by a live *Salmonella typhimurium* vaccine carrier strain and expressed by naked DNA. *Vaccine* 19, 4028–4035

5 Lalvani, A. *et al.* (2001) Rapid detection of *Mycobacterium tuberculosis* infection by enumeration of antigen-specific T cells. *Am. J. Respir. Crit. Care Med.* 163, 824–828

6 Berthet, F.X. *et al.* (1998) A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). *Microbiology* 144, 3195–3203

7 Gey Van Pittius, N.C. *et al.* (2001) The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. *Genome Biol.* 2, RESEARCH0044

8 Cole, S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544

9 Tekaia, F. *et al.* (1999) Analysis of the proteome of *Mycobacterium tuberculosis in silico*. *Tubercle Lung Dis.* 79, 329–342

10 Rindi, L. *et al.* (1999) Search for genes potentially involved in *Mycobacterium tuberculosis* virulence by mRNA differential display. *Biochem. Biophys. Res. Commun.* 258, 94–101

11 Skjot, R.L. *et al.* (2000) Comparative evaluation of low-molecular-mass proteins from *Mycobacterium tuberculosis* identifies members of the ESAT-6 family as immunodominant T-cell antigens. *Infect. Immun.* 68, 214–220

12 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

13 Okinaka, R.T. *et al.* (1999) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J. Bacteriol.* 181, 6509–6515

14 Nolling, J. *et al.* (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol.* 183, 4823–4838

15 Neuwald, A.F. *et al.* (1999) AAA+: a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* 9, 27–43

16 Pallen, M.J. *et al.* (1997) Coiled-coil domains in proteins secreted by type III secretion systems. *Mol. Microbiol.* 25, 423–425

17 Christie, P.J. (2001) Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Mol. Microbiol.* 40, 294–305

18 Delahay, R.M. *et al.* (1999) The coiled-coil domain of EspA is essential for the assembly of the type III secretion translocon on the surface of enteropathogenic *Escherichia coli*. *J. Biol. Chem.* 274, 35969–35974

19 Daniell, S.J. *et al.* (2001) Coiled-coil domain of enteropathogenic *Escherichia coli* type III secreted protein EspD is involved in EspA filament-mediated cell attachment and hemolysis. *Infect. Immun.* 69, 4055–4064

20 Knutton, S. *et al.* (1998) A novel EspA-associated surface organelle of enteropathogenic *Escherichia coli* involved in protein translocation into epithelial cells. *EMBO J.* 17, 2166–2176

Mark J. Pallen
Divn of Immunity and Infection,
Birmingham University Medical School,
Birmingham, UK B15 2TT.
e-mail: m.pallen@bham.ac.uk