environmental microbiology reports

Environmental Microbiology Reports (2013) **5**(1), 117–126 doi:10.1111/j.1758-2229.2012.00394.x

Distribution and diversity of bacterial secretion systems across metagenomic datasets

Matthieu Barret, Frank Egan and Fergal O'Gara*
BIOMERIT Research Centre, Department of Microbiology, University College Cork, Cork, Ireland.

Summary

Bacteria can manipulate their surrounding environment through the secretion of proteins into other living organisms and into the extracellular milieu. In Gram stain negative bacteria this process is mediated by different types of secretion systems from type I through type VI secretion system (T1SS–T6SS). In this study the prevalence of these secretion systems in 312 publicly available microbiomes derived from a wide range of ecosystems was investigated by a gene-centric approach. Our analysis demonstrates that some secretion systems are over-represented in some specific samples. In addition, some T3SS and T6SS phylogenetic clusters were specifically enriched in particular ecological niches, which could indicate specific bacterial adaptation to these environments.

Introduction

Gram stain negative bacteria rely on several secretion systems to influence their environment by translocating protein and DNA into host cells and the extracellular milieu. These secretion systems can range from simple transporters to multi-component complexes and have been classified into six types: from type I through type VI secretion system (T1SS–T6SS) (Filloux, 2011). In addition, T5SS could be further divided into monomeric (T5aSS) or trimeric autotransporters (T5cSS) and classic (T5bSS) or fused two-partner secretion systems (T5dSS) (Filloux, 2011). Although all the secretion systems have been thought to be primarily involved in the virulence of plant and animal pathogenic bacteria, it is now generally accepted that bacterial protein secretion plays a key role in the modulation of different biotic interactions, ranging from symbiosis to pathogenesis (Henderson *et al.*, 2004; Preston, 2007; Juhas *et al.*, 2008; Schwarz *et al.*, 2010a; Barret *et al.*, 2011a). Numerous comparative genomics analyses have indeed highlighted that non-pathogenic bacterial strains (i.e. symbiotic or commensal) also harbour a wide diversity of secretion systems (Pallen *et al.*, 2005; Boyer *et al.*, 2009; Persson *et al.*, 2009; Tseng *et al.*, 2009). Moreover, these secretion systems are not solely encoded in the genomes of bacteria belonging to the *Proteobacteria* phylum. For example, a T3SS expressed under *in vitro* conditions has been identified in the genome of a bacterium belonging to the *Verrucomicrobia* phylum (Sait *et al.*, 2011).

While several studies have investigated the distribution of secretion systems in genome sequences, relatively few reports have investigated the prevalence of these secretion systems in natural environments (Mazurier *et al.*, 2004; Warmink and van Elsas, 2008; Persson *et al.*, 2009). Moreover, these reports have been restricted to specific ecosystems (i.e. aquatic or terrestrial) and/or to specific secretion systems (e.g. T3SSs). The availability of more than 300 metagenomic datasets derived from different ecosystems in the IMG/M database (Markowitz *et al.*, 2012) provided an opportunity to investigate the prevalence and distribution of secretion systems in natural environments. In this study, the prevalence of secretion systems in 312 metagenomic datasets derived from 10 different ecosystem categories was assessed through a process-centric approach, which considers a community from the point of view of its functions rather than its organisms (Tringe *et al.*, 2005; Hugenholtz and Tyson, 2008). With such an approach, secretion systems over-represented in one environmental sample are expected to be selected by the local environment and therefore may exert an important function in that niche.

Results and discussion

Distribution of secretion systems in microbiomes

The prevalence of secretion systems (T1SS to T6SS) in bacterial communities associated with different ecological niches was investigated in a total of 312 metagenomic datasets (Table S1) present in the IMG/M database (Markowitz *et al.*, 2012). In this study, protein-coding genes were assigned to a specific secretion system using cluster of orthologous group (COG) as a functional classification (Tatusov *et al.*, 2003) for the following reasons: (i) the percentage of protein-coding genes assigned to

Received 5 June, 2012; accepted 29 August, 2012. *For correspondence. E-mail f.ogara@ucc.ie; Tel. (+353) (0)21 427 2646; Fax (+353) (0)21 427 5934.

COG identifiers is higher than for Pfam or TIGRfam (Table S1) and (ii) COGs, which group proteins with sequence similarity over the entire length, are more sensitive in detection of overall protein relationships (Mavromatis *et al*., 2009). COG identifiers corresponding to the specific protein family of each secretion system were selected according to literature (Delepelaire, 2004; Henderson *et al*., 2004; Cianciotto, 2005; Cornelis, 2006; Alvarez-Martinez and Christie, 2009; Boyer *et al*., 2009). As numerous proteins involved in the assembly of secretion systems are derived from other membrane-bound systems (i.e. T2SS and type IV pilus, T3SS and flagella), the specificity of each COG was initially assessed by a systematic investigation of their distribution and genetic location in complete genomes (summarized in Table S2). The prevalence of secretion systems in microbiomes was assessed using the number of single COGs for single-component secretion systems (T5aSS, T5bSS and T5cSS) or the average number of COGs per metagenome for multi-component secretion systems (T2SS, T3SS, T4SS and T6SS). Although T1SSs are composed of three proteins (TolC, HlyB and HlyD), only one COG identifier is T1SS-specific (Table S2). For this reason, T1SSs are analysed as a single-component secretion system in this study. The recently described T5dSS (Salacha *et al*., 2010) is not considered in this analysis because of the lack of a specific COG identifier.

According to genome sequences, some secretion systems are more prevalent in *Proteobacteria* (Table S3). Linear regression analyses indeed demonstrated that, in general, the prevalence of secretion systems is better correlated with the number of predicted proteobacterial genes (Fig. 1) rather than the number of predicted bacterial genes present in each metagenomic dataset (Fig. S1). The only secretion system which seems unaffected by the relative abundance of proteobacterial sequences is the T4SS. This is somewhat unsurprising as T4SSs are used for conjugation in many Gram stain positive and Gram stain negative bacteria (Table S3) and are also frequently encoded on self-transmissible plasmids and integrative conjugative elements (Juhas *et al*., 2007; Alvarez-Martinez and Christie, 2009).

As the relative abundance of proteobacterial genes could vary significantly between distinct samples derived from different ecosystem categories (Table S1 and Fig. S2), we decided to analyse the prevalence of secretion systems in relation to the number of predicted proteobacterial genes present in each metagenome. The relative abundance of each secretion system per sample was estimated by analysis of the standardized residuals of the linear regression. Secretion systems with standardized residuals inferior to –2.576 and superior to 2.576 were defined as under-represented and over-represented respectively (Tables 1 and S4). For instance, the T4SS is particularly enriched in a metagenome sample associated to wastewater treatment plant (Table 1). This is in accordance with the fact that wastewater treatment plants are favourable environments for horizontal gene transfer, because of intense selective pressures imposed by a wide range of xenobiotics (Schlueter *et al*., 2007). Interestingly, T3SSs and T6SSs are often enriched in arthropoda-associated microbiomes. However, while T3SSs are enriched in distinct ambrosia beetles microbiomes (i.e. *Dendrotocus frontalis* or *Xyleborus affinis*), T6SSs are preferentially associated with different fungus-growing ants of the Attini's tribe (i.e. *Atta*, *Apterostigma*, *Cyphomyrmex* or *Trachymyrmex*). Unlike the multi-component secretion systems, the distribution of single-component secretion systems seems to be relatively ubiquitous (Fig. S2), as there is no clear correlation between the abundance of T1SS, T5aSS and T5bSS and a particular ecosystem (Table S4).

Distribution of specific phylogenetic clusters in metagenomes

Distinct phylogenetic clusters have been distinguished within each multi-component secretion system (Peabody *et al*., 2003; Troisfontaines and Cornelis, 2005; Juhas *et al*., 2008; Boyer *et al*., 2009). Investigating the correlation between secretion system prevalence and environment might be more informative if the presence of distinct phylogenetic families of the secretion systems is taken into account, as it may reflect the different host–bacterial interactions these systems mediate. To date, this correlation has only been investigated using genome sequences (Pallen *et al*., 2005; Boyer *et al*., 2009), which are biased towards a phylogenetically limited range of bacterial pathogens such as salmonellae (3.5% of all proteobacterial genomes) or *Escherichia coli* (8.5%). To assess if a potential correlation between phylogenetic cluster and a particular ecological niche exists, individual phylogenetic analysis was performed on each multi-component secretion system. To do so, we first selected a coding DNA sequence (CDS) which was relatively short (the vast majority of the metagenomic datasets are composed of small scaffolds and therefore CDS are frequently truncated) and, more importantly, specific to the secretion system considered. Accuracy of the analysis was first validated on CDS derived from genome sequences. If data were congruent with previous studies, reference genomic sequences from each phylogenetic cluster were retrieved and used in subsequent metagenomic analysis. As the T2SS and T4SS phylogenetic trees obtained with metagenomic sequences were not congruent to those obtained with genomic sequences, the distribution of T2SS and T4SS phylogenetic clusters within different ecosystem categories was not assessed further.

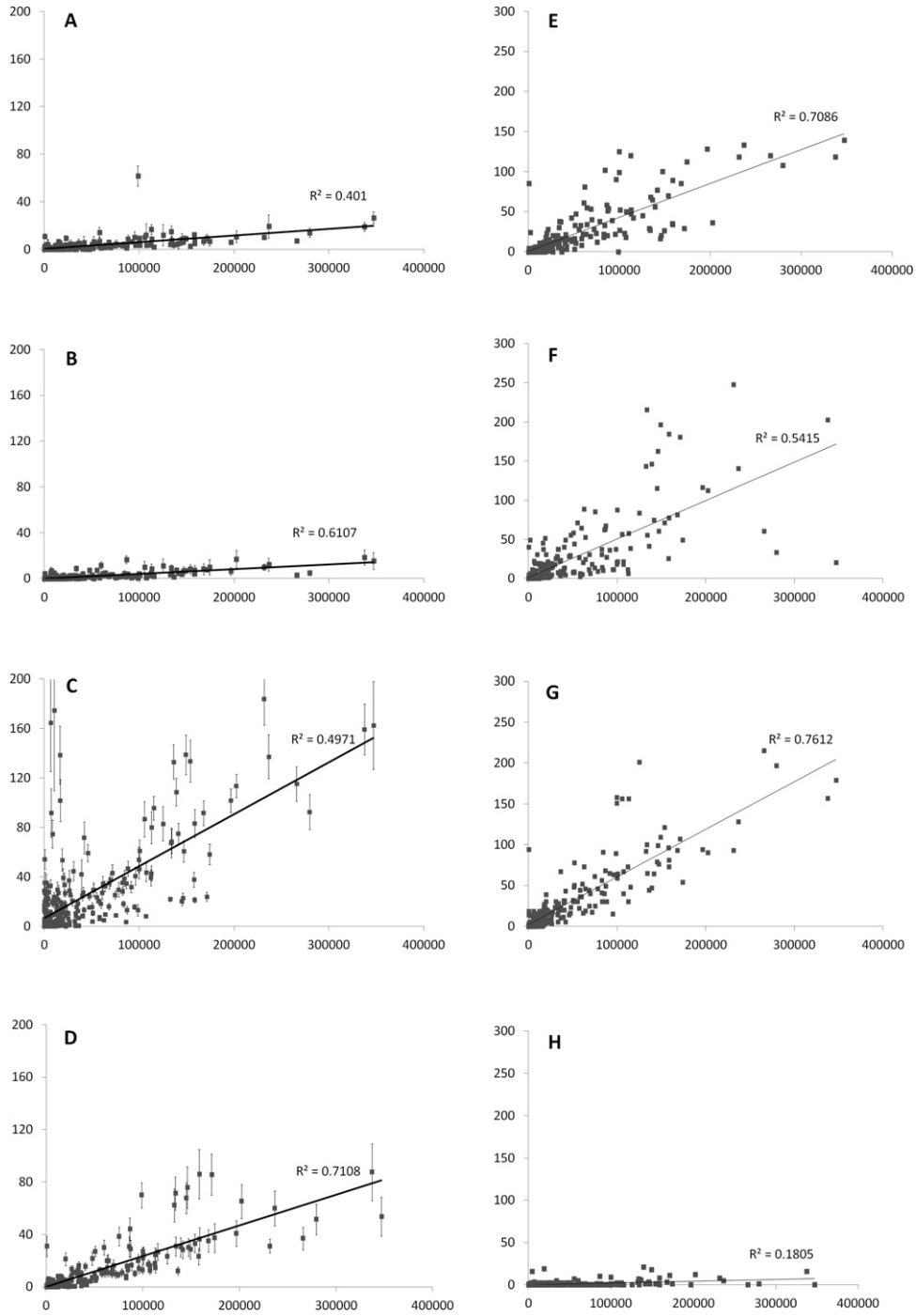

Fig. 1. Distribution of secretion systems in metagenomic datasets. The *x*-axis represents the number of predicted proteobacterial protein-coding genes (those with more than 60% identity to proteins from proteobacterial genomes) present in each metagenomic dataset. The *y*-axis represents the average of COGs per metagenome for multi-component secretions systems (A–D) and the number of COGs for single-component secretion system (E, F). Graphics A, B, C and D represent the distribution of T2SS, T3SS, T4SS and T6SS, while graphics E, F, G and H represent the distribution of T1SS, T5aSS, T5bSS and T5cSS.

T3SS. T3SS is a nanomachine composed of approximately 25 proteins (Cornelis, 2010), which has evolved into seven different families: Ysc, Hrp1, Hrp2, SPI-1, SPI-2, *Rhizobiales* and *Chlamydiales* (Pallen *et al*., 2005; Troisfontaines and Cornelis, 2005). Phylogenetic analysis performed on 544 YscT homologues (COG4791) derived from genome sequences indeed highlighted seven main clades, which correspond to the previously described

Table 1. Metagenome samples enriched or depleted in multi-component secretion systems.

IMG/M ID	Metagenome	Ecosystem category	Frequency	SR
T2SS				
2077657010	Saline water microbial communities from Great Salt Lake	Aquatic	0.041	2.613
2088090006	Sediment microbial communities from Lake Washington	Aquatic	0.010	2.767
2156126005	*Trichodesmium* cyanobacterial communities	Aquatic	0.059	2.800
2189573029	*Bankia setacea* microbiome	Mollusca	0.260	14.755
T3SS				
2088090013	Sediment microbial communities from Lake Washington	Aquatic	0.004	–4.395
2088090005	Sediment microbial communities from Lake Washington	Aquatic	0.003	–3.693
2199352008	Rice-straw enriched compost microbial community	Engineered	0.017	2.749
2032320008	*Dendroctonus ponderosae* fungus gallery microbiome	Arthropoda	0.108	2.826
2088090007	Sediment microbial communities from Lake Washington	Aquatic	0.010	3.174
2100351007	Sediment microbial communities from Lake Washington	Aquatic	0.011	3.374
2030936000	*Amitermes wheeleri* gut microbiome	Arthropoda	0.021	3.557
2032320009	*Dendroctonus ponderosae* microbiome from jack pine	Arthropoda	0.208	4.062
2084038008	*Xyleborus affinis* microbiome	Arthropoda	0.046	4.668
2035918003	*Dendroctonus ponderosae* microbiome from lodgepole pine	Arthropoda	0.132	4.943
2044078007	*Dendroctonus frontalis* microbiome	Arthropoda	0.127	7.188
T4SS				
2100351011	Coastal water and sediment microbial communities from Arctic	Aquatic	0.015	–2.815
2038011000	*Atta colombica* fungus garden	Arthropoda	0.123	2.633
2019105002	Faecal microbiome of *Canis familiaris*	Mammals	0.212	2.673
2030936000	*Amitermes wheeleri* gut microbiome	Arthropoda	0.285	2.716
2048955003	Poplar biomass decaying microbial community	Engineered	0.150	2.759
2035918001	Wastewater treatment plant	Engineered	3.128	2.904
2030936006	*Atta texana* microbial communities, dump top	Arthropoda	0.257	3.133
2084038013	Gut microbiome of *Anoplophora glabripennis*	Arthropoda	0.200	3.434
2084038008	*Xyleborus affinis* microbiome	Arthropoda	0.312	3.525
2019105001	Faecal microbiome of *Canis familiaris*	Mammals	0.204	3.630
2032320007	*Atta texana* microbial communities, dump top	Arthropoda	0.264	4.110
2100351002	Crop microbiome from Hoatzin	Birds	0.206	4.367
2199352012	Rice-straw enriched compost microbial community	Engineered	0.217	5.040
2013843001	Wastewater treatment plant	Engineered	4.030	7.107
T6SS				
2100351000	Biofuel metagenome	Engineered	0.020	–3.351
2088090013	Sediment microbial communities from Lake Washington	Aquatic	0.045	–3.046
2199352012	Rice-straw enriched compost microbial community	Engineered	0.037	–2.786
2044078007	*Dendroctonus frontalis* microbiome	Arthropoda	0.337	2.878
2088090006	Sediment microbial communities from Lake Washington	Aquatic	0.027	3.720
2030936005	*Cyphomyrmex longiscapus* microbiome	Arthropoda	0.327	3.767
2084038018	*Trachymyrmex* microbiome	Arthropoda	0.338	4.067
2065487013	Hindgut termite microbiome	Arthropoda	0.198	4.820
2029527006	*Atta colombica* microbiome, garden bottom	Arthropoda	0.380	5.013
2029527005	*Atta colombica* microbiome, garden top	Arthropoda	0.355	5.494
2189573029	*Bankia setacea* microbiome	Mollusca	0.294	5.642
2029527003	*Apterostigma dentigerum* microbiome	Arthropoda	0.389	5.877

The relative abundance of each secretion system in the different samples was estimated by analyses of the standardized residuals (SRs) of the linear regression. Samples with SRs inferior to –2.576 and superior to 2.576 were defined as under-represented and over-represented respectively. Information relative to SRs of single-component secretion systems can be found in Table S4.

T3SS families (Fig. 2). These YscT homologues are only encoded in genomes of bacteria belonging to the *Proteobacteria*, *Chlamydia* and *Verrucomicrobia* phyla (Table S4).

To investigate the distribution of the T3SS families in different ecosystems, protein-coding genes derived from metagenomic datasets were submitted to phylogenetic analysis. Results obtained were congruent with the phylogenetic analysis previously performed on genome sequences (Fig. 2). Based on our data, the Hrp1 family is the most common T3SS cluster among metagenomic datasets, comprising 20.7% of all YscT homologues (Table 2). This distribution contrasts with genomic data where SPI-1 is the most abundant clade. Interestingly, the relative abundance of each T3SS family varies significantly between the different ecosystems (Table 2). For example, the Hrp1 family seems over-represented in arthropoda (especially in ambrosia beetles fungus garden) and plant-associated microbiomes. This is in accordance with enrichment of Hrp1 positive strains in

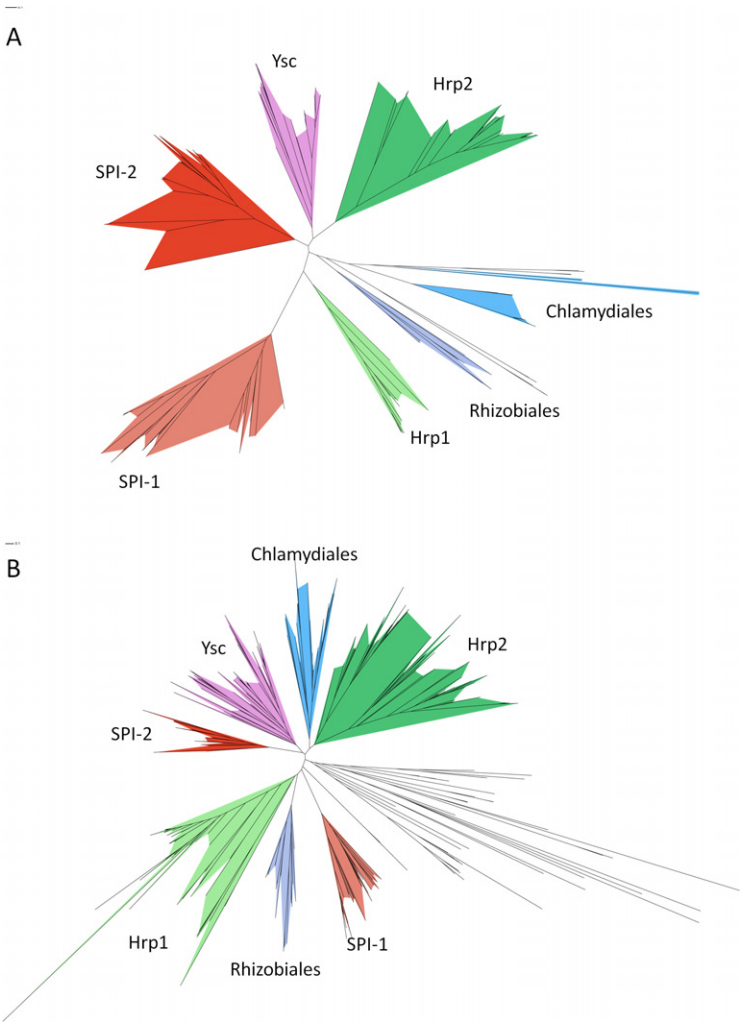

Fig. 2. Phylogenetic distribution of T3SSs in genomes and metagenomes. A distance tree (Maximum-Likelihood) was calculated from 544 and 241 YscT sequences (COG4791) derived from genomes (A) and metagenomic datasets (B) respectively. Only protein sequences longer than 100 amino acids in length were conserved for further analysis (241 out of 369). Protein sequences were aligned using the default parameter of MAFFT (Katoh and Toh, 2008). Maximum-likelihood trees were built with PhyML (Guindon and Gascuel, 2003) using the WAG amino acid substitution model of evolution (Whelan and Goldman, 2001) and four categories of substitution rates. Branch supports were evaluated using the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006). Phylogenetic trees were visualized and exported using the web-based tool Interactive Tree Of Life (iTol) (Letunic and Bork, 2011).

the rhizosphere (Mazurier *et al.*, 2004) and in the mycorrhizosphere of different plants species (Warmink and van Elsas, 2008; Viollet *et al.*, 2011), which suggest that this T3SS could be involved in plant–bacterial interactions but also in bacterial–fungal interactions (Leveau and Preston, 2008). Another interesting result is the high abundance of Chlamydiale family within microbiomes derived from aquatic ecosystems (Table 2). As there is

Table 2. Relative abundance of T3SS families in different ecosystems.

	Chlamydiale	Hrp1	Hrp2	Rhizobiale	SPI-1	SPI-2	Ysc	n.d.
Genome (544)	8.3 (45)	6.3 (34)	20.4 (111)	2.9 (16)	28.3 (154)	21.5 (117)	11.4 (62)	0.9 (5)
Metagenome (241)	15.8 (38)	20.7 (50)	16.2 (39)	9.5 (23)	8.7 (21)	3.3 (8)	12.0 (29)	13.7 (33)
Aquatic (67)	43.3 (29)	11.9 (8)	14.9 (10)	1.5 (1)	1.5 (1)	0	7.5 (5)	19.4 (13)
Arthropoda (98)	2.0 (2)	28.6 (28)	11.2 (11)	13.3 (13)	19.4 (19)	7.1 (7)	11.2 (11)	7.1 (7)
Engineered (21)	4.8 (1)	0	23.8 (5)	19.0 (4)	0	4.8 (1)	38.1 (8)	9.5 (2)
Mammals (2)	0	0	0	0	0	0	100 (2)	0
Mollusca (1)	0	0	0	0	0	0	0	100 (1)
Plants (14)	0	35.7 (5)	28.6 (4)	21.4 (3)	0	0	7.1 (1)	7.1 (1)
Terrestrial (38)	15.8 (6)	23.7 (9)	23.7 (9)	5.3 (2)	2.6 (1)	0	5.3 (2)	23.7 (9)

The percentage of each T3SS family within genomes and metagenomes is presented. The number of sequences is indicated in parentheses. The distribution of secretion systems in bacterial communities was investigated in 10 ecosystem categories using the classification system proposed by Ivanova and colleagues (2010). For sake of simplicity all the ecosystem categories belonging to engineered ecosystems (bioremediation, biotransformation, solid waste and wastewater) were grouped together.
n.d.: not determined.

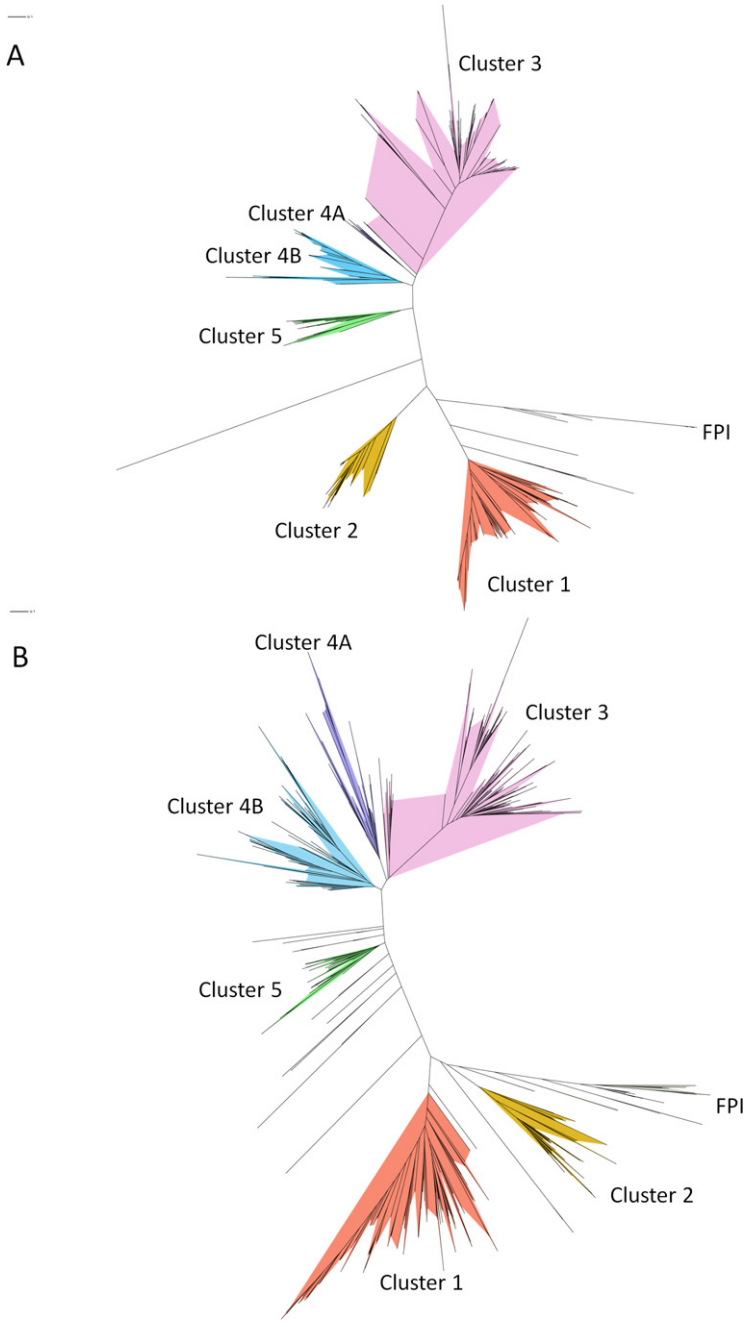

Fig. 3. Phylogenetic distribution of T6SSs in genomes and metagenomes. A distance tree (Maximum-Likelihood) was calculated from 1127 and 894 IglA homologues (COG3516) derived from genomes (A) and metagenomic sequences (B) respectively. Only protein sequences longer than 75 amino acids in length were conserved for further analysis (894 out of 1740).

no evidence for lateral gene transfer of chlamydiale T3SSs (Collingro *et al.*, 2011), the relative abundance of these T3SS protein-coding genes in aquatic environments is probably explained by a high prevalence of *Chlamydia*.

A total of 13.7% of the YscT homologues (Fig. 2B) are not grouped with any phylogenetic cluster and could therefore represent alternative T3SS phylogenetic clusters or simply be involved in the assembly of another protein complex (i.e. bacterial flagellum). To verify that the YscT homologues derived from metagenomes were indeed T3SS-related proteins, the genomic location of 85 YscT proteins encoded on contigs greater than 1 kb in length (Fig. S3) was investigated. Twenty-one contigs encode only the *yscT* gene, 59 contigs encode other T3SS-related genes, while five represent flagellar biosynthetic loci. YscT from these five contigs are all related to the unidentified T3SS phylogenetic cluster (Fig. 2B), which therefore is likely to represent flagellar-derived proteins and not YscT homologues.

Table 3. Relative abundance of T6SS families in different ecosystems.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4A	Cluster 4B	Cluster 5	*Francisella*	n.d.
Genome (1127)	27.3 (307)	14.7 (166)	35.4 (398)	2.1 (24)	11.6 (131)	5.3 (60)	3.5 (40)	0.1 (1)
Metagenome (894)	16.7 (149)	14.7 (131)	32.6 (291)	5.6 (50)	13.0 (116)	9.2 (82)	6.5 (58)	1.9 (17)
Aquatic (193)	14.0 (27)	5.2 (10)	36.3 (70)	9.8 (19)	11.9 (23)	7.3 (14)	11.4 (22)	4.1 (8)
Arthropoda (347)	16.4 (57)	31.4 (109)	36.9 (128)	3.2 (11)	9.8 (34)	0.9 (3)	0.3 (1)	1.2 (4)
Engineered (58)	17.2 (10)	1.7 (1)	34.5 (20)	8.6 (5)	19.0 (11)	19.0 (11)	0	0
Mammals (4)	75 (3)	25 (1)	0	0	0	0	0	0
Microbial (1)	0	0	100 (1)	0	0	0	0	0
Mollusca (70)	10.0 (7)	0	0	0	0	48.6 (34)	41.4 (29)	0.0
Plants (44)	29.5 (13)	13.6 (6)	31.8 (14)	11.4 (5)	9.1 (4)	2.3 (1)	2.3 (1)	0.0
Terrestrial (177)	18.1 (32)	2.3 (4)	32.8 (58)	5.6 (10)	24.9 (44)	10.7 (19)	2.8 (5)	2.8 (5)

The percentage of each T6SS family within genomes and metagenomes is presented. The number of sequences is indicated in parentheses. The distribution of secretion systems in bacterial communities was investigated in 10 ecosystem categories using the classification system proposed by Ivanova and colleagues (2010). For sake of simplicity all the ecosystem categories belonging to engineered ecosystem (bioremediation, biotransformation, solid waste and wastewater) were grouped together.
n.d.: not determined.

T6SS. T6SSs are macromolecular machineries composed of approximately 15 conserved proteins, which are involved in virulence towards different eukaryotes and/or in bacterial killing (Schwarz *et al*., 2010a). T6SSs have been initially divided in five phylogenetic clusters (Boyer *et al*., 2009), but a recent study has further divided cluster 4 into 4A and 4B (Barret *et al*., 2011b). In addition, a seventh family, called FPI (for *Francisella* pathogenicity island), has also been identified (de Bruin *et al*., 2007), but although some protein-coding genes of the FPI show some similarities with T6SS components, its genetic organization suggests that FPI does not encode a 'real' T6SS (Boyer *et al*., 2009).

The protein IglA (COG3516) is a highly conserved T6SS component that has been successfully used in a previous phylogenetic analysis (Schwarz *et al*., 2010b). Our analysis performed on 1127 IglA homologues encoded in numerous genome sequences clearly identified seven clusters corresponding to T6SS-1, T6SS-2, T6SS-3, T6SS-4A, T6SS-4B, T6SS-5 and FPI (Fig. 3). Surprisingly, a single protein sequence encoded in the genome of the proteobacteria *Crenothrix polyspora* did not group with any T6SS clusters. Taxonomic analysis of the IglA sequences revealed that T6SSs are widely encoded in the genomes of *Proteobacteria*, but are also present in the genomes of some bacterial strains related to *Acidobacteria*, *Gemmatimonadetes*, *Nitrospirae*, *Planctomycetes* and *Verrucomicrobia* (Table S3).

The genomic context of 318 *iglA* hits from metagenomes was analysed in contigs greater than 1 kb in length using the gene neighbourhood tool of IMG/M. A total of 295 contigs (92.8%) encode an additional T6SS-related gene. Some of the remaining 23 contigs may represent T6SS loci, as 18 have sequence on only one side of the *iglA* gene, two contain only *iglA* and three encode hypothetical protein either side of *iglA*. The topology of the phylogenetic tree constructed from 894 IglA homologues derived from metagenomic datasets is in accordance with the genomic tree (Fig. 3). Moreover, the same phylogenetic clusters are present in each analysis. The T6SS-3 cluster (which includes HSI-I, SPI-6 and T6SS-3) is the most prevalent group in genome and metagenomic datasets, accounting for approximately 30% of all T6SSs (Table 3). This cluster has been shown to be involved in bacterial killing in a number of *Proteobacteria* including *Pseudomonas aeruginosa* (Hood *et al*., 2010; Russell *et al*., 2011) and *Serratia marcescens* (Murdoch *et al*., 2011). Moreover, recent evidence suggests that T6SSs belonging to cluster 3 and 4B are able to deliver antibacterial effectors (Russell *et al*., 2012). In contrast, clusters 1 and 2 are particularly prevalent in mammals and arthropoda-associated metagenomes respectively. This is an interesting observation as these secretion systems have mainly been characterized in the context of bacterial–eukaryote interactions (Lesic *et al*., 2009; Schwarz *et al*., 2010b; Suarez *et al*., 2010; Zhou *et al*., 2012). Finally, the cluster 5 previously linked with the aquatic ecological niche (Boyer *et al*., 2009) seems instead highly prevalent in mollusca-associated microbiomes, more precisely in an endosymbiotic bacterial consortium of wood-boring marine bivalves.

Conclusion

In this study the distribution of secretion systems in natural bacterial communities associated with distinct environmental niches was investigated using a gene-centric approach. This analysis has shown that some secretion systems such as T3SS and T6SS are enriched in environmental samples associated with different arthropoda. Moreover, some specific T3SS (e.g. Hrp1) and T6SS phylogenetic clusters (e.g. cluster 5) seem to be over-represented in some particular environmental

niches. Such correlations suggest that the interrogation of metagenomic datasets is a promising approach for developing new hypotheses about the basis for important interactions in different environments. The rapid increase in metagenome sequence datasets will allow more comprehensive studies of the distribution of these secretion systems.

Acknowledgements

This research was supported in part by grants awarded to FOG by the Science Foundation of Ireland (07IN.1/B948, 08/RFP/GEN1295, 08/RFP/GEN1319, SFI09/RFP/BMT2350); the Department of Agriculture, Fisheries and Food (RSF grants 06-321 and 0-377; FIRM grants 06RDC459, 06RDC506 and 08RDC629); the European Commission (MTKD-CT-2006-042062, Marie Curie TOK-:TRAMWAYS, EU256596, MicroB3-287589-OCEAN2012, MACUMBA-CP-TP 311975; PharmaSea-CP-TP 312184); IRCSET (05/EDIV/FP107/INTERPAM, EMBARK), the Marine Institute Beaufort award (C&CRA 2007/082), the Environmental Protection Agency (EPA 2006-PhD-S-21, EPA 2008-PhD-S-2) and the HRB (RP/2006/271, RP/2007/290, HRA/2009/146). The authors would like to acknowledge the Boole Centre for Research in Informatics at University College Cork for providing access to computational facilities.

References

Alvarez-Martinez, C.E., and Christie, P.J. (2009) Biological diversity of prokaryotic type IV secretion systems. *Microbiol Mol Biol Rev* **73:** 775–808.

Anisimova, M., and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* **55:** 539–552.

Barret, M., Morrissey, J., and O'Gara, F. (2011a) Functional genomics analysis of plant growth-promoting rhizobacterial traits involved in rhizosphere competence. *Biol Fertil Soils* **47:** 1–15.

Barret, M., Egan, F., Fargier, E., Morrissey, J.P., and O'Gara, F. (2011b) Genomic analysis of the type VI secretion systems in *Pseudomonas* spp.: novel clusters and putative effectors uncovered. *Microbiology* **157:** 1726–1739.

Boyer, F., Fichant, G., Berthod, J., Vandenbrouck, Y., and Attree, I. (2009) Dissecting the bacterial type VI secretion system by a genome wide *in silico* analysis: what can be learned from available microbial genomic resources? *BMC Genomics* **10:** 104.

de Bruin, O.M., Ludu, J.S., and Nano, F.E. (2007) The *Francisella* pathogenicity island protein IglA localizes to the bacterial cytoplasm and is needed for intracellular growth. *BMC Microbiol* **7:** 1.

Cianciotto, P.M. (2005) Type II secretion: a protein secretion system for all seasons. *Trends Microbiol* **13:** 581–588.

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R.C., *et al*. (2011) Unity in variety – the pangenome of the *Chlamydiae*. *Mol Biol Evol* **28:** 3253–3270.

Cornelis, G.R. (2006) The type III secretion injectisome. *Nat Rev Microbiol* **4:** 811–825.

Cornelis, G.R. (2010) The type III secretion injectisome, a complex nanomachine for intracellular 'toxin' delivery. *Biol Chem* **391:** 745–751.

Delepelaire, P. (2004) Type I secretion in gram-negative bacteria. *Biochim Biophys Acta* **1694:** 149–161.

Filloux, A. (2011) Protein secretion systems in *Pseudomonas aeruginosa*: an essay on diversity, evolution and function. *Front Microbiol* **2:** 155.

Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C., and Ala'Aldeen, D. (2004) Type V protein secretion pathway: the autotransporter story. *Microbiol Mol Biol Rev* **68:** 692–744.

Hood, R.D., Singh, P., Hsu, F.S., Guvener, T., Carl, M.A., Trinidad, R.R.S., *et al*. (2010) A type VI secretion system of *Pseudomonas aeruginosa* targets a toxin to bacteria. *Cell Host Microbe* **7:** 25–37.

Hugenholtz, P., and Tyson, G.W. (2008) Microbiology – metagenomics. *Nature* **455:** 481–483.

Ivanova, N., Tringe, S.G., Liolios, K., Liu, W.-T., Morrison, N., Hugenholtz, P., and Kyrpides, N.C. (2010) A call for standardized classification of metagenome projects. *Environ Microbiol* **12:** 1803–1805.

Juhas, M., Crook, D.W., Dimopoulou, I.D., Lunter, G., Harding, R.M., Ferguson, D.J.P., and Hood, D.W. (2007) Novel type IV secretion system involved in propagation of genomic islands. *J Bacteriol* **189:** 761–771.

Juhas, M., Crook, D.W., and Hood, D.W. (2008) Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. *Cell Microbiol* **10:** 2377–2386.

Katoh, K., and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9:** 286–298.

Lesic, B., Starkey, M., He, J., Hazan, R., and Rahme, L.G. (2009) Quorum sensing differentially regulates *Pseudomonas aeruginosa* type VI secretion locus I and homologous loci II and III, which are required for pathogenesis. *Microbiology* **155:** 2845–2855.

Letunic, I., and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39:** W475–W478.

Leveau, J.H.J., and Preston, G.M. (2008) Bacterial mycophagy: definition and diagnosis of a unique bacterial-fungal interaction. *New Phytol* **177:** 859–876.

Markowitz, V.M., Chen, I.M.A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., *et al*. (2012) IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40:** D123–D129.

Mavromatis, K., Chu, K., Ivanova, N., Hooper, S.D., Markowitz, V.M., and Kyrpides, N.C. (2009) Gene context analysis in the integrated microbial genomes (IMG) data management system. *PLoS ONE* **4:** e7979.

Mazurier, S., Lemunier, M., Siblot, S., Mougel, C., and Lemanceau, P. (2004) Distribution and diversity of type III secretion system-like genes in saprophytic and phytopathogenic fluorescent pseudomonads. *FEMS Microbiol Ecol* **49:** 455–467.

Murdoch, S.L., Trunk, K., English, G., Fritsch, M.J., Pourkarimi, E., and Coulthurst, S.J. (2011) The opportunistic pathogen *Serratia marcescens* utilizes type VI secretion to target bacterial competitors. *J Bacteriol* **193:** 6057–6069.

Pallen, M.J., Beatson, S.A., and Bailey, C.M. (2005) Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS Microbiol Rev* **29:** 201–229.

Peabody, C.R., Chung, Y.J., Yen, M.R., Vidal-Ingigliardi, D., Pugsley, A.P., and Saier, M.H. (2003) Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149:** 3051–3072.

Persson, O.P., Pinhassi, J., Riemann, L., Marklund, B.I., Rhen, M., Normark, S., *et al*. (2009) High abundance of virulence gene homologues in marine bacteria. *Environ Microbiol* **11:** 1348–1357.

Preston, G.M. (2007) Metropolitan microbes: type III secretion in multihost symbionts. *Cell Host Microbe* **2:** 291–294.

Russell, A.B., Hood, R.D., Bui, N.K., LeRoux, M., Vollmer, W., and Mougous, J.D. (2011) Type VI secretion delivers bacteriolytic effectors to target cells. *Nature* **475:** 343–347.

Russell, A.B., Singh, P., Brittnacher, M., Bui, N.K., Hood, R.D., Carl, M.A., *et al*. (2012) A widespread bacterial type VI secretion effector superfamily identified using a heuristic approach. *Cell Host Microbe* **11:** 538–549.

Sait, M., Kamneva, O.K., Fay, D.S., Kirienko, N.V., Polek, J., Shirasu-Hiza, M.M., and Ward, N.L. (2011) Genomic and experimental evidence suggests that *Verrucomicrobium spinosum* interacts with eukaryotes. *Front Microbiol* **2:** 211.

Salacha, R., Kovacic, F., Brochier-Armanet, C., Wilhelm, S., Tommassen, J., Filloux, A., *et al*. (2010) The *Pseudomonas aeruginosa* patatin-like protein PlpD is the archetype of a novel type V secretion system. *Environ Microbiol* **12:** 1498–1512.

Schlueter, A., Szczepanowski, R., Puehler, A., and Top, E.M. (2007) Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance gene pool. *FEMS Microbiol Rev* **31:** 449–477.

Schwarz, S., Hood, R.D., and Mougous, J.D. (2010a) What is type VI secretion doing in all those bugs? *Trends Microbiol* **18:** 531–537.

Schwarz, S., West, T.E., Boyer, F., Chiang, W.C., Carl, M.A., Hood, R.D., *et al*. (2010b) *Burkholderia* Type VI secretion systems have distinct roles in eukaryotic and bacterial cell interactions. *PLoS Pathog* **6:** e1001068.

Suarez, G., Sierra, J.C., Erova, T.E., Sha, J., Horneman, A.J., and Chopra, A.K. (2010) A type VI secretion system effector protein, VgrG1, from *Aeromonas hydrophila* that induces host cell toxicity by ADP ribosylation of actin. *J Bacteriol* **192:** 155–168.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., *et al*. (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Troisfontaines, P., and Cornelis, G.R. (2005) Type III secretion: more systems than you think. *Physiology* **20:** 326–339.

Tseng, T.-T., Tyler, B.M., and Setubal, J.C. (2009) Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol* **9:** S2.

Viollet, A., Corberand, T., Mougel, C., Robin, A., Lemanceau, P., and Mazurier, S. (2011) Fluorescent pseudomonads harboring type III secretion genes are enriched in the mycorrhizosphere of *Medicago truncatula*. *FEMS Microbiol Ecol* **75:** 457–467.

Warmink, J.A., and van Elsas, J.D. (2008) Selection of bacterial populations in the mycosphere of *Laccaria proxima*: is type III secretion involved? *ISME J* **2:** 887–900.

Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18:** 691–699.

Zhou, Y., Tao, J., Yu, H., Ni, J., Zeng, L., Teng, Q., *et al*. (2012) Hcp family proteins secreted via the type VI secretion system coordinately regulate *Escherichia coli* K1 interaction with human brain microvascular endothelial cells. *Infect Immun* **80:** 1243–1251.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Phylum distribution of protein-coding genes. The taxonomic analysis of each metagenomic dataset used in this study was assessed through the phylogenetic distribution tool of the IMG/M database, which displays the phylum distribution of protein-coding genes in each metagenome based on their best match using BLASTp. Proteins which display less than 60% identity are excluded from this analysis.

Fig. S2. Distribution of secretion systems in metagenomic datasets. The *x* axis represents the number of predicted bacterial protein-coding genes present in each metagenomic dataset. The *y* axis represents the average of COGs per metagenome for multi-component secretion systems (A–D) and the number of COGs for single-component secretion system (E–F). Graphics A, B, C and D represent the distribution of T2SS, T3SS, T4SS and T6SS while graphics E, F, G and H represent the distribution of T1SS, T5aSS, T5bSS and T5cSS.

Fig. S3. Length of contigs carrying YscT (A) and IglA (B) protein coding sequences.

Table S1. Metagenomic datasets used in this study.

Table S2. COGs used in this study. COG identifiers corresponding to specific protein families of each secretion systems were selected according to literature (Delepelaire, 2004; Henderson *et al*., 2004; Cianciotto, 2005; Cornelis, 2006; Alvarez-Martinez and Christie, 2009; Boyer *et al*., 2009). The number of COGs used for estimate the distribution each secretion system family is indicated in parentheses. The total number of COGs in all metagenomic datasets examined is also indicated.

Table S3. Taxonomic distribution of secretion systems in genome sequences. The average number of secretion systems per genome was calculated in each bacterial phylum (every bacterial class for the *Proteobacteria*) by using all the specific COG identifiers selected in Table S2. The asterisk indicates that multiple COG identifiers have been selected for the estimation of T5aSS.

Table S4. Distribution of secretion systems within each metagenomic dataset.