# A first global analysis of plasmid encoded proteins in the ACLAME database

Leplae, Raphaël; Lima-Mendez, Gipsi; Toussaint, Ariane

*Published in:*
FEMS microbiology reviews

*DOI:*
10.1111/j.1574-6976.2006.00044.x

*Publication date:*
2006

*Document Version*
Publisher's PDF, also known as Version of record

# A first global analysis of plasmid encoded proteins in the ACLAME database

Raphaël Leplae, Gipsi Lima-Mendez & Ariane Toussaint

SCMBB, Université Libre de Bruxelles, Bvd du Triomphe, Bruxelles, Belgium

**Correspondence:** Raphaël Leplae, SCMBB, CP 263, Université Libre de Bruxelles, Bvd du Triomphe, B1050 Bruxelles, Belgium. Tel.: +32 2 6505426; fax: +32 2 6505425; e-mail: raphael@scmbb.ulb.ac.be

Received 17 March 2006; revised 25 July 2006; accepted 30 August 2006.
First published online October 2006.

DOI:10.1111/j.1574-6976.2006.00044.x

Editor: Rafael Giraldo

**Keywords**
plasmid; protein family; global analysis; ACLAME; network.

## Abstract

Many plasmids are mobile genetic elements (MGEs) and, as other members of that group of DNA entities, their genomes display a mosaic and combinatorial structure, making their classification extremely difficult. As other MGEs, plasmids play a major role in horizontal transfer of genetic materials and genome reorganization. Yet, the full impact of such phenomenon on major properties of the host cell, such as pathogenicity, the ability to use new carbon sources or resistance to antibiotics, remains to be fully assessed. More and more complete plasmid genome sequences are available. However, in the absence of standards for storing plasmid sequence data and annotating genes and gene products on sequenced plasmid genomes, the resulting information remains rather limited. Using 503 sequenced plasmids organized in the ACLAME database, we discuss how, by structuring information on the genomes, their host and the proteins they code for, one can gain access to either global or more detailed analysis of the plasmid sequence information, as illustrated by a network representation of the relationships between plasmids.

## Introduction

Plasmids are autonomously replicating, circular or linear DNA entities, residing in some microbial cells in addition to the chromosome(s). Bacterial conjugative plasmids encode conjugation machinery, using it to transfer to recipient cells, and hence are prokaryotic mobile genetic elements (MGE). MGEs form a diverse group, which includes genetic entities/DNA segments, playing a central role in the rapid adaptation of bacteria to, sometimes drastic, environmental changes. They are susceptible to one or more enzymes or enzyme machineries, allowing them to move within genomes (intracellular mobility) and/or between bacterial cells (intercellular mobility). Active MGEs encode their own mobility proteins that can eventually complement related defective elements. Together with bacteriophages, archaeal viruses and some types of genomic islands [i.e. in a broad sense, potentially mobile DNA segments that reside integrated in bacterial chromosomes, see Frost *et al.* (2005) for more details], conjugative plasmids are responsible for the intra and interspecific horizontal transfer of large segments of genetic material and can express crucial functions involved in pathogenicity/virulence, biodegradation or drug resistance. In addition, MGEs express promiscuous site-specific and transpositional recombination systems that can reshuffle the host genome without a need for sequence relatedness [see (Merlin *et al.*, 2000) for an extensive review]. The importance of these properties of MGEs on bacterial evolution and more specifically on the emergence of new pathogens or strains with new catabolic properties is now clearly emphasized by the comparison of the complete genome sequences of closely related strains that often differ mainly by their MGE content (Glaser *et al.*, 2001, Parkhill *et al.*, 2001).

More and more plasmid and other prokaryotic MGE nucleotide sequences are available, whether individually or as part of completely sequenced genomes. Unfortunately, they are generally poorly annotated, and when integrated in bacterial genomes, they often remain unrecognized. In a first step towards a more comprehensive collection and analysis of MGE sequences, we have developed the ACLAME database [for **A** **CLA**ssification of **M**obile genetic **E**lements, (Leplae *et al.*, 2004)], which provides a common framework for the representation of MGEs and their components and tools to facilitate their analysis. A computational procedure allows for the classification in families of the proteins encoded by the MGEs. The families are here defined as ensembles of similar proteins that can share one or more functions. A functional annotation is performed on these families using an ontology open to a community-wide

discussion to ensure a continuous development and refinement of the function definitions. The ACLAME website (http://aclame.ulb.ac.be) presently provides access to the families of proteins encoded by 184 pages and 503 plasmids (version 0.2) downloaded from the National Center for Biotechnology Information (NCBI) (Benson *et al.*, 2005) repository sections devoted to phage and plasmid genomes by the end of January 2003. In the present review, we discuss a few examples illustrating how the information compiled and structured in ACLAME can be used to annotate protein families. These families can then be used to detect or recover sets of proteins that can be linked with specific plasmid features or to study cooccurring proteins on the whole set of 503 plasmids. This allows, for the first time, to visualise the entire evolutionary relationships of the sequenced plasmids across species and genus boundaries of their hosts. We discuss some of the problems we encountered while populating and organising the ACLAME database, and possible solutions that could bring more homogeneity, e.g. in new GenBank plasmid sequence files.

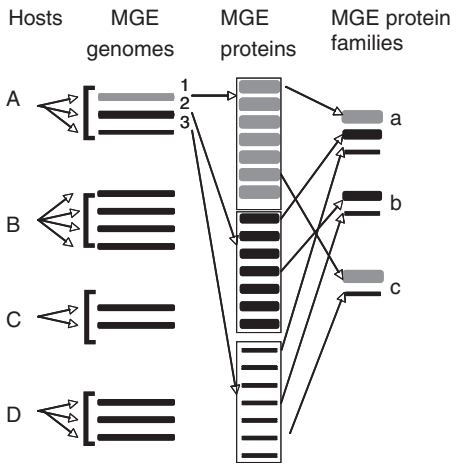
## Overview of the ACLAME database organization

The general scheme for the population and organization of information in the ACLAME database is described on the ACLAME website (http://aclame.ulb.ac.be/Classification/description.html). Plasmids were automatically downloaded in GenBank format from the NCBI website. During the GenBank file parsing step, a consistency check was made on the data and whenever dubious or faulty information was encountered, a report was produced so that a manual check and fix of the data could be performed. Some plasmids sequenced at the time did not appear in our downloaded files. Some of those turned out not to have any annotated coding sequence; unfortunately the plasmid referred to as the 'Birmingham IncP-α plasmid' (NC_001621) (Pansegrau *et al.*, 1994), a compilation of sequences originating from the quasi-identical IncP plasmids R68, RK2 and RP4 (E. Lanka, personal communication), was among them. All three are among the best-characterized plasmids *in vivo* and would thus have been a most useful reference for our functional annotations.

Other plasmid genomes were correctly downloaded, but had only one defined open reading frame (ORF) although their size was well over 1 kb, suggesting that they were incompletely annotated. We also noticed the inappropriate use of the GenBank organism field with some downloaded plasmids (*Phaseolus* or *Homo sapiens* for instance, i.e. the eukaryotic host of the bacterial host of the plasmid). These inappropriately assigned hosts have been corrected in ACLAME when supplementary information was available, either from the sequence file or from the literature. For broad host range (BHR) plasmids that have several hosts in several genera and plasmids that have been isolated from unknown or unculturable bacteria, we so far used the term 'not defined' in the ACLAME host field. Future versions of the database will accommodate the possibility to assign multiple hosts to a given plasmid.

Once the plasmids were loaded in ACLAME, the protein clustering procedure and the search for similar proteins in sequence databases was made as described earlier [(Leplae *et al.*, 2004), see also Fig. 1].

Table 1 summarizes the content of ACLAME version 0.2. Five hundred and three plasmid genomes (15 of which




| Genome 1 | Family | ACLAME function |
|---|---|---|
| Protein 1 | a | Function U |
| Protein 2 | b | Function V |
| Protein 3 | c | Function W |
| Protein 4 | d | Function X |
| Protein 5 | e | Function Y |
| Protein 6 | f | Function Z |

**Fig. 1.** Summary of the content and organization of information in the ACLAME database. Information within ACLAME consists of i) a list of bacterial hosts (shown here by capital letters A–C), which provides access to individual MGE genomes in each host (designated here by numbers 1–3 for host A only); ii) each MGE genome entry is linked to a list of all proteins encoded by that genome (boxed for genomes 1–3) and (not shown) to its host, NCBI Entrez file entry and host taxonomy NCBI entry; each protein is again linked to the corresponding NCBI file, and iii) a list of protein families (here a–c for genomes 1–3). The families were built by comparing the proteins 'all vs. all' using BLAST (Altschul *et al.*, 1997). All the *E*-values (similarity measure) obtained were treated using a modified version of the TRIBE-MCL script (Enright *et al.*, 2002), which groups proteins into 'Clusters' (i.e. families), with two adjustable thresholds, the *E*-value (here set at 0.01) and a parameter that controls the cluster granularity (here set at 1.2). These two values were shown previously to be appropriate to reproduce standard protein domains and IS encoded proteins classifications (Leplae *et al.*, 2004). Each ACLAME Cluster (protein family) was manually assigned with a function based on available experimental or other published information and additional analysis (see text for more details). As a result each genome can be visualized as a succession of proteins, each of which is linked to the family it belongs to, and hence to the function assigned to that family (illustrated for genome 1 in the box).

**Table 1.** Overview of the ACLAME database content in version 0.2

| Category | Phage-encoded 'Vir' | Plasmid-encoded 'Plasmid' | Phage and plasmid-encoded 'All' |
|---|---|---|---|
| Total number of proteins in ACLAME | 10 465 | 20 722 | 31 187 |
| Number of proteins in the largest family (Cluster:plasmids:0) | 112 | 319 | 319 |
| Number of hosts | 71 | 241 | 281 |
| Total number of ACALME protein families (Clusters) | 5219 | 6904 | 11 907 |
| Number of families containing three or more proteins | 844 | 1377 | 2199 |
| Total number of proteins in families of three or more proteins | 5488 | 14 051 | 19 772 |
| Percentage of proteins in families of three or more proteins | 52% | 67% | 63% |
| Number of families of three or more plasmid proteins only* | – | – | 1197 (54%) |
| Number of families of three or more phage proteins only* | – | – | 693 (31%) |
| Number of families of three or more phage and plasmid proteins* | – | – | 309 (14%) |
| Number of families containing four or more proteins | 581 | 1009 | 1578 |
| Total number of proteins in families of four or more proteins | 4699 | 12 947 | 17 909 |
| Percentage of proteins in families of four or more proteins | 45% | 62% | 57% |

The term 'Cluster' refers to a family of related proteins resulting from 'clustering' with the MCL algorithm (van Dongen, 2000). Percentages are calculated vs. the numbers of proteins shown in the first line.

*Only valid for the 'All' category. See the text for definitions of the 'Vir', 'Plasmid' and 'All' categories

belong to eukaryotic organisms) contributed 20 722 annotated protein sequences, which have been clustered in 6904 protein families. These proteins and families form the category 'Plasmids'. Version 0.2 also includes a set of 10 465 protein sequences encoded by 184 phage genomes. These and the corresponding families form the category 'Viruses'. Finally the combined set of the 31 187 phage and plasmid protein sequences and the associated families form the category 'All'. The three categories can also be visualized and browsed through the ACLAME web interface. Sixty three percent of the protein set 'All', 67% of the protein set 'Plasmids' and 53% of the protein set 'Viruses' belong to families of three or more members. A comparison of the cluster composition between the three sets shows that 97.2% of the families obtained from the 'Viruses' set and 98.6% of the families obtained from the 'Plasmids' set remain unchanged in the 'All' set. This indicates that most plasmid encoded protein families differ from phage encoded ones and that the clustering procedure is not affected by the heterogeneity of the sequences classified. However, in that set 'All', 309 out of 2199 (i.e. 14%) of the families with three or more members contain both phage and plasmid proteins. These include families of site-specific tyrosine and serine recombinases (e.g. cluster:all:4, 5, 473, 717), replication functions (e.g. cluster:all:305, 792, 197, 195, 88, 69, 45, 43, 39, 14) and regulatory proteins (e.g. cluster:all:11, 57, 75, 475), which are expected to be found on all types of MGEs. Some families have phage tail and fibre proteins (e.g. cluster:all:477–479, and 304, respectively) encoded by the plasmids pMT1 from *Yersinia pestis* and pHCM2 *Salmonella enterica* ssp. *enterica* serovar Typhi and are likely corresponding to phage tail-like bacteriocins, which to our knowledge have not yet been identified experimentally.

## Functional annotation of the protein families

As it can be easily seen on the ACLAME website, the individual protein annotations found in the NCBI sequence files were often of little help in deciding on the nature of the function that could be assigned to the ACLAME family the proteins belong to. The additional information from sequence similarities detected via database searches, both at the level of individual proteins and at the level of the families, was thus very valuable to manually analyze and assign a functional annotation to the protein families. We originally considered two ontologies for this purpose: MultiFun (Serres & Riley, 2000), the standard for bacterial genomes, and the Gene Ontology (GO) (Harris *et al*., 2004), used for several eukaryotic genomes. GO is organized as a hierarchy of biological processes, cellular components and molecular functions and has now integrated most of the MultiFun definitions and many other database resources. Therefore, we presently use only GO as a primary resource for annotations in ACLAME and redirect the users to the GO consortium website (http://www.geneontology.org/) to benefit from all their additional information and the permanent update of the ontology. However, GO (and MultiFun) is not precisely MGE-oriented and many MGE-specific functions are as yet unavailable. Our strategy therefore was to maximize the use of suitable existing GO functions, whether at one or another category (molecular function, biological process or cellular component) of the ontology. In many instances, a more precise description was added (for instance when experimental data support the function of one or more protein in a given family) generating 'child' functions. As a result, several ACLAME functions have the same GO

accession (e.g. all type IV secretion proteins). When no suitable function was available in GO, new terms and descriptions were generated. The complete list of functions used so far is accessible on the ACLAME website (see http://aclame.ulb.ac.be/functions). The list should be considered as 'work in progress' since it is under continuous evolution and some supplementary work will be required before it fully fits the appropriate data structure for the integration in GO. Future ACLAME versions will provide multiple functional assignments to a single protein family, allowing a combination of annotations for a biological process and/or a component and/or a function to a family to conform to the GO ontology.

So far, enzyme functions were deliberately given a generic annotation because e.g. transacetylases or transporters can be responsible for drug resistance (to chloramphenicol for the first, to tetracycline or several drugs for the latter). Once appropriate criteria is defined, the biological process 'antibiotic resistance' or 'multi drug resistance' could be assigned to a family of proteins with the molecular function 'transacetylase' or 'transporters' providing a more comprehensive annotation. Multiple-function assignment will also allow for more accurate annotation of families containing multi-domain proteins with a different function per domain, or families that contain very similar proteins with different functions (as for instance transposases and transposition regulators encoded for some IS, see below).

With the protein families and their functional annotation, individual genomes can be displayed on the ACLAME website as an organized succession of proteins (the same as the corresponding genes), each of which is linked to the family containing that protein and to its function where it was assigned. As a result, the relationship of each individual protein of a given plasmid with any other(s) is readily accessible.

## General features of 503 sequenced plasmids

Three hundred and eight out of the 503 plasmids analyzed are 10 kb or shorter, 153 are between 10 and 100 kb and 42 are over 100 kb. Whether this reflects the size distribution of plasmids in nature is an open question, since this information is not readily accessible. Figure 2 shows the distribution of plasmids per host, calculated as the number of plasmids and the number of nucleotides sequenced per host species. Leaving aside *Borrelia*, which can carry as many as 21 plasmids in a single strain [(Casjens *et al.*, 2000) and references therein], *Staphylococcus* contributed to the largest number of sequenced plasmids. However, agrobacteria and rhizobia plasmids, due to their large size, contributed to the largest number of base pairs and proteins in ACLAME. The distribution shown in Fig. 2 highlights the fragmentary and biased information we have on plasmids and, as recently discussed by Frost *et al.* (2005), there is an imperious need for new large-scale plasmid sequencing projects to fill the gaps.

## Distribution of the ACLAME functional annotations in plasmids

Seven hundred and twenty seven families of five or more plasmid proteins were manually assigned with a functional annotation. Most of the smaller families were so far left unannotated as they mainly contain proteins of unknown function. This is illustrated in Fig. 3, which shows how the number of families, as a function of the cluster size, differs between defined functional categories (here illustrated by the cumulated 'DNA metabolism, replication, recombination and repair' and 'Transport and secretion' functional categories defined as listed in Table S1) and the 'Molecular function unknown' category. Only the latter displays a sharp increase among smaller families. Functional categories were

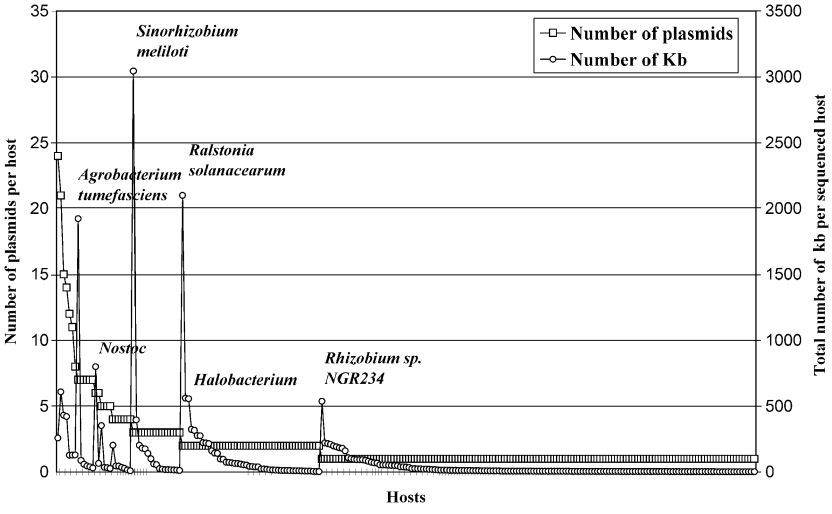



**Fig. 2.** Distribution of the sequenced plasmid genomes among host species. The left-hand Y axis refers to the number of plasmids per host (line with squares, 37 genomes from 'not-defined' hosts, 21 from *Borrelia burgdorferi*, 20 from *Staphylococcus aureus*, 12 from *Escherichia coli* and *Lactococcus lactis*, 11 from *Clostridium glutamicum*, seven from *Clostridium jeikeium*, *Helicobacter pylori*, *Aeromonas salmonicida* and *Bacillus longum*, 6 from *Nostoc* sp. PCC 7120, *Streptococcus epidermidis* and five from *Yersinia pestis, Bacillus subtilis* and *Lactococcus plantarum*), and the right-hand Y axis refers to the total number of base pairs (in kb) sequenced per host (line with circles). Each genome is represented by a point of the X axis.

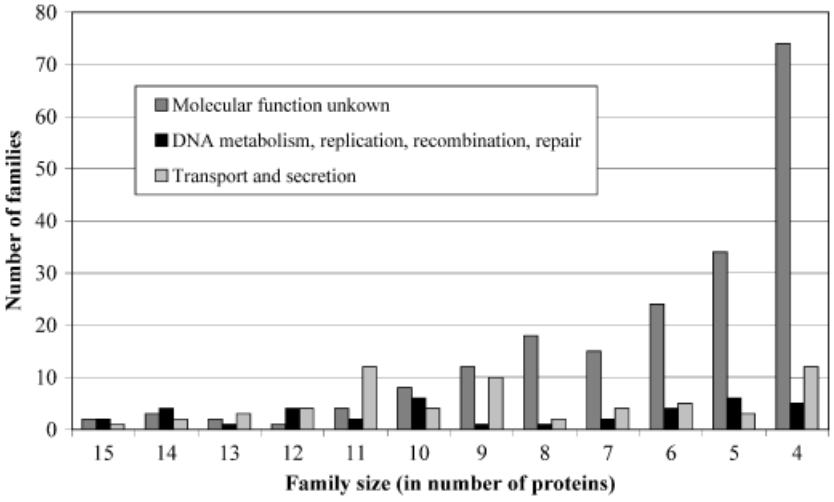



**Fig. 3.** Size distribution of protein families with 'Unknown molecular function', 'DNA metabolism, replication, recombination and repair' and 'Transport and secretion'. See Table S1 for a detailed description of the functional categories. The number of protein families (cluster:plasmids) containing from four to 15 members were cumulated for the three functional categories 'Molecular function unknown', 'DNA metabolism, replication, recombination and repair' and 'Transport and secretion' (see Table S1 for detailed description).

used to calculate the distribution of the most frequent functions shown in Tables 2A and B. This is the first general snapshot of plasmid-encoded functions, although some caution is required due to the fragmented and biased data the analysis relies on. One of the biases shown in Fig. 2 appears in more detail here. Indeed, more than 37% of the proteins (7850 out of 20 722) in the set analyzed are encoded by only 11 very large plasmids, all hosted by plant associated bacteria: three Ti plasmids from *Agrobacterium tumefaciens*, pRi1724 from *Agrobacterium rhizogenes*, the symbiotic plasmids pSymA and B from *Sinorhizobium meliloti*, pMLa and pMLb from *Mesorhizobium loti*, p42d from *Rhizobium elti*, pNGR234a from *Rhizobium* sp. NGR234 and the *Ralstonia solanacearum* pGMI1000MP megaplasmid. Among those 11 plasmids, four particularly large ones contribute to a very important fraction of the proteins in the most populated families (see Table 2C).

**Table 2A.** Main functional categories found on plasmids in ACLAME

| Functional categories | Number of families | Number of proteins |
|---|---|---|
| Unknown function | 212 | 1826 |
| Metabolic activity | 132 | 1704 |
| Transporters | 107 | 2241 |
| DNA replication recombination and repair | 76 | 2762 |
| Regulation | 45 | 1316 |
| Partition | 8 | 355 |

Functional categories were assembled from ACLAME functions as listed in Table S1, some of which are detailed in Table 2B. The number of families is the cumulated number of cluster:plasmids appended with the corresponding functions. The number of proteins is the cumulated number of proteins in those families.

## Families of plasmid specific proteins

Essential backbone plasmid genes encode proteins involved in replication initiation, copy number control, partitioning and stability or maintenance. One of the largest protein families, cluster:plasmids:2 with 213 proteins, contains Walker-box ATPases, some of which have been shown to control plasmid partitioning [ParA proteins, for review see (Gerdes *et al*., 2000)]. Ebersbach & Gerdes (2005) recently reviewed partitioning (*par*) loci, which ensure ordered plasmid segregation prior to cell division. They come in two types. The first *parCMR* type encode ParM actin-like ATPases. These form dynamic filaments that segregate plasmids paired at mid-cell to daughter cells. Like microtubules, ParM filaments exhibit dynamic instability whose regulation is an important component of the DNA segregation process. ParM proteins were found in cluster:plasmids:260 (11 proteins) and a ParR partner regulator could be readily identified for 10 of them in two families, cluster:plasmids:445 and cluster:plasmids:1312 with, respectively, seven and three proteins. The second *parABS/sopABC* type encode a ParA deviant Walker-type ATPase (found here in cluster:plasmids:2) related to MinD. ParA proteins form highly dynamic, oscillating filaments that are required for the subcellular movement and positioning of plasmids. The original annotations of the Walker box ParA ATPases were particularly uninformative, due in part to the traditional use of different names for homologous proteins encoded by different, rather well characterized, plasmids (SopA protein in plasmid F and RepA in Ti). The family is quite heterogeneous, with member proteins having sequence lengths spanning from 100 to 750 aa. Considering that genuine ParA proteins are 200–400 residues long (Gerdes *et al*., 2000), this family illustrates well the difficulty to cope with multidomain proteins. Several plasmids contributed with more than one protein to cluster:plasmids:2, including

**Table 2B.** Distribution among some main functional categories

| Function | | | | | |
|---|---|---|---|---|---|
| DNA recombination* | | Transporters | | Regulators† | |
| Subfamilies | Number of proteins | Subfamilies | Number of proteins | Subfamilies | Number of proteins |
| Tyr recombinase | 179 | ABC transporters | 849 | LysR-like | 172 |
| Ser recombinase | 175 | Type II secretion | 78 | MarR-like | 20 |
| DNA transposition | 1384 | Type III secretion | 138 | CRP/FNR-like | 16 |
| | | Type IV secretion | 576 | Other | 1108 |
| | | Other | 600 | | |

*only site specific recombinases and transposases are taken into account. Proteins involved in homologous recombination are not considered.
†LysR-like, Mar-like and CRP/FNR-like stand respectively for transcriptional regulators with sequence similarity to LysR, controls a regulator of lysine biosynthesis in *Escherichia coli* (Schell, 1993), MarR, a multiple antibiotic resistance regulator widespread among enteric bacteria (Cohen *et al.*, 1993) and the CRP/FNR global regulators (Korner *et al.*, 2003). 'Other' stands for all families with the same function but not within the subfamilies listed in the column.

**Table 2C.** Contribution of a few large plasmids to some of the most populated protein families in ACLAME

| | Function | | | | | |
|---|---|---|---|---|---|---|
| | Host<br>Plasmid name | SM<br>pSymA | SM<br>pSymB | RS<br>pGMI1000 | AT<br>AT* | %Total |
| | Number of proteins | (1264) | (1569) | (1674) | (547/542) | (27)** |
| LysR-like regulators | 172 | 43 | 22 | 31 | 17/17 | 75 |
| ABC transporters | 849 | 142 | 237 | 35 | 70/74 | 66 |
| Sensor-regulators | 199 | 26 | 30 | 54 | 9/7 | 63 |
| GntR-like regulators | 102 | 17 | 23 | 14 | 5/5 | 63 |
| MDR permeases | 108 | 7 | 9 | 32 | 1/1 | 46 |
| Transposition | 1384 | 61 | 24 | 65 | 13/11 | 14 |
| Site specific recombination | 355 | 10 | 1 | 8 | 2/2 | 6 |

The numbers in the second column show the cumulated number of proteins in cluster:plasmids families appended with the function in the first column. Numbers in parentheses are the total number of proteins encoded by the plasmid in the first line.
*two very similar *Agrobaterium* AT plasmid sequences are present in ACLAME version 0.2. They are here considered side by side.
%Total is the percentage over the total number of proteins encoded by the plasmids listed in this table.
**Percentage of the cumulated number of proteins encoded by those same plasmids over the cumulated number of proteins in cluster:plasmids families in the second column.
Hosts of the plasmids: SM, *Sinorhizobium meliloti*; RS, *Ralstonia solanacearum*; AT, *Agrobaterium tumefasciens* for AT.

characterized proteins such as an arsenite-induced protein from plasmid R64 and a nitrogenase component from plasmid p42d. The latter proteins are likely not involved in partitioning, but because they share the ATPase domain with genuine ParA proteins they were aggregated by the clustering procedure. The ParA/SopA proteins are encoded from *parABS/sopABC* operons where the A and B proteins of the operons act on the *parS/sopC* site [see (Gerdes *et al.*, 2000) for review]. ParB/SopB partner protein families in ACLAME were readily identified (cluster:plasmids:20, 86, 375 and 411, representing a total of 113 proteins). From these ParB/SopB families, 50 out of the 80 proteins in cluster:plasmids:20 and all of the 24 and eight proteins in cluster:plasmids:86 and 375 had at least one partner in cluster:plasmids:2. Based on this evidence, the cluster:plasmids:2 was annotated with the 'Partitioning-ParA' function. Dividing the family into groups of more closely related proteins combined with the analysis of genes adjacent to *parA* without an identified *parB* partner (some of which may not have been annotated as coding sequences in the original GenBank files) would help to validate this functional assignment but is beyond the scope of this more global analysis.

Replication initiation functions, often called RepA (but RepB in Agrobacteria and Rhizobia), were spread among at least 19 families for a total of 416 proteins. Among those, some appeared corresponding to previously proposed classes (http://www.essex.ac.uk/bs/staff/osborn/DPR_home.htm, see Table 3). Eight matching families contained Rep proteins encoded by rolling circle replicating plasmids. The largest among those, cluster:plasmids:11, mapped well on a group of relaxases encoded by conjugative plasmids [reviewed in (Francia *et al.*, 2004)]. These enzymes are part

**Table 3.** Replication initiation families

| ACLAME families | DPR RepA families | Inc group* |
|---|---|---|
| cluster:plasmids:11 | RCR XIII (conjugation relaxases) | |
| Cluster:plasmids:25 | RCR III | |
| Cluster:plasmids:132 | RCR V | |
| Cluster:plasmids:141 | RCR II | |
| Cluster:plasmids:204 | RCR I | |
| Cluster:plasmids:205 | RCR IV | |
| Cluster:plasmids:621 | RCR IX (A) | |
| Cluster:plasmids:528 | Theta Group A | IncB, L/M, K, I |
| Cluster:plasmids:114 | Theta Group A | RepFIC, FII, FIII, Z |

*the data on incompatibility come from the Database of Plasmid Replicons (DPR) website (http://www.essex.ac.uk/bs/staff/osborn/DPR_home.htm). RCR, rolling circle replication; Theta, theta replication. Proteins from the families in the DPR were aligned with the proteins in ACLAME using the ACLAME BLAST search engine, Evalue threshold of 0.01. The sets of proteins in ACLAME and the DPR overlap only partially and hence the two sets of families overlap also partially. No correspondence could be found between ACLAME proteins and DPR RCR families VI–VIII, IX (B), X–XII and XIV–XVII. Theta Group A is the only theta group accessible in DPR.

of the relaxosome complex and have a DNA binding and nicking activity at the origin of the DNA transfer site. They are shared between conjugative plasmids from Gram-negative and Gram-positive bacteria, while the rest of the conjugative transfer machinery is quite distinct in most known instances. Conjugative plasmids residing in Gram-negative hosts encode a set of proteins that assemble in a mating pair formation (Mpf) apparatus. This apparatus is related to the type IV secretion system [see (Zechner *et al.*, 2000) for reviews] and forms the channel between two bacteria for plasmid transfers. The Mpf and the relaxosome complexes are coupled via a coupling protein of the TraG-like ATPase family. This protein is called VirD4 in Ti and related plasmids and TrwB in the R388 plasmid. The last one was not completely sequenced at the time of our download and was therefore not part of our analysis. To retrieve more information on conjugation systems from ACLAME, we checked the plasmids that had a relaxase protein in cluster:-plasmids:11 for the presence of a TraG/VirD4 ATPase and other components of the Mpf apparatus. This identified 28 plasmids (listed in supplementary material as Table S2), all of which carried at least one stretch of genes coding for additional proteins belonging to, or related to, a type IV secretion system. These included a second ATPase typical of type IV and Type II secretion systems (classified in cluster:-plasmids:33) as well as several proteins assigned with the type IV/Mpf protein function (cluster:plasmids:34, 157, 192, 194, 212–214, 265, 629, 773, 784). The two well-characterized and distantly related Mpf complexes, those of the paradigm plasmids Ti/RP4 (Cascales & Christie, 2004) and F [see (Frost *et al.*, 2005) and references therein] are represented in these clusters although unevenly (only cluster:plasmids:157 and 265 contain F-like proteins). This search did obviously not reveal all the components expected in the F-like conjugative plasmids. Further analysis is needed to identify the families of missing components. They could be small unannotated families, or families annotated as 'unknown function' because the information available from the NCBI sequence files and similarity searches was not convincing enough to decide on a functional assignment. Cluster:plasmids:192 to 194 and cluster:plasmids:212 to 214 contain Mpf proteins from Ti-related plasmids as well as proteins encoded by plasmids R751, pB4, pB10, R64, pADP-1 or pUO1 (see supplementary Table S2 for their respective hosts). The specific role in Mpf, and the relationship with other Mpfs, has been discussed in details for the proteins encoded by plasmid pB4 (Schluter *et al.*, 2003) and pB10 (Tauch *et al.*, 2003). Plasmids pMRC01 from *Lactococcus lactis* and pSK41 and pLW043 from *Staphylococcus aureus* (i.e. Gram-positive bacteria) encode a relaxase and a TraG-like coupling protein that clustered in the same families discussed above. This is consistent with the earlier reports of similarities between the components of these transfer machineries and those of plasmids from Gram-negative hosts (Firth *et al.*, 1993, Hickey *et al.*, 2001). The details of the complex relationships existing between conjugation machineries are difficult to tackle by manual analysis. As illustrated, bioinformatics tools are available or being developed that, for instance, allow scoring the cooccurrence of pairs or other combination of conserved genes/proteins.

## Main families of non plasmid specific proteins

Most of the largest plasmid protein families could not be considered *a priori* as plasmid-specific. These include site-specific or transpositional recombinases. Most transposases probably belong to IS sequences. Among the 19 IS families defined so far [(Chandler & Mahillon, 2003) and http://www-is.biotoul.fr], some include IS sequences that encode three polypeptides. One of them is the transposase, which is produced by a programmed translational frameshift, which fuses most of the two other IS-encoded proteins involved in transposition regulation. The clustering procedure obviously groups these proteins sharing a common portion of the sequence. Some of the families annotated with the function 'DNA recombination, transposase activity' (e.g. function : 498 defined as 'DNA recombination, transposase activity, IS*3* family') do presently include such transposition regulators. This ambiguity will be resolved with the possibility to assign several functions to one family.

One hundred and thirty five out of the 503 plasmids (26.8%) featured proteins annotated with a transposition function (ACLAME function:352, 471, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513 and

514). Amongst those plasmids, 55% had more than 10% of their genes encoding such proteins. The most populated one was the *Shigella* plasmid pWR501 with, as previously reported, 150 out of its 293 genes (51.2%) encoding transposase related proteins (Venkatesan *et al*., 2001).

Many tyrosine and serine site-specific recombinases constituted another set among the largest families. Eleven Tyr recombinases, two of which were originally annotated as defective, bore the typical motifs of integron integrases (D. Mazel, personal communication). They are encoded by plasmids R46, R100, R721, R751, pJR2, pTET3, pHCM1, pCG4, pCTX-M3, pB10 and p1658/97 (see Supplementary Table S3 for a more exhaustive description). All these Tyr recombinases were linked with at least one resistance gene and, on 10 plasmids, the adjacent genes encoded several proteins typically encoded by gene cassettes. In most cases, these integron candidates were flanked by transposase genes (the IS*6* family found in cluster:plasmids:14, Tn*7* family found in cluster:plasmids:175 or Tn*3* family found in cluster:plasmids:32). This is a well-known feature in several cases. Plasmid R100 has an integron inserted in its Tn*21* transposon, which encodes a transposase from the Tn*3* family (Liebert *et al*., 1999). Plasmid R751 has the integron in the Tn*5090* transposon from the Tn*7* family (Radstrom *et al*., 1994). Plasmid pHCM1 (Wain & Kidgell, 2004) has it between a Tn*3* and an IS*6* transposase and in pCG4 from the Gram-positive *Corynebacterium glutamicum* the integron was suggested to be associated with transposases (Nesvera *et al*., 1998). However, the association of integrons with other mobile sequences does not seem to be the rule since in pB10 the integron is located between the *tra* and *trb* regions (Schluter *et al*., 2003). No published information seems to be available on the other integron candidates we found. A systematic analysis of gene neighbours should shed some light on the overall tendency of integrons to be associated with other MGEs.

Transporter proteins were also among the most frequently represented functional categories. A large number of ABC transporters (Table 2B and C) were almost exclusively located on the large soil bacteria plasmids. A very limited number of transporters appeared to belong to Type II and III (Table 2B and 3) secretion systems, including the paradigms *Yersinia, Shigella* and *Ralstonia* Type III systems '[(see for instance (Cornelis & Van Gijsegem, 2000, Tran Van Nhieu *et al*., 2000)]. Those systems can be reconstituted from the ACLAME families (see details in Table 4).

**Table 4.** Families of proteins belonging to the type II and III secretion systems

| ClusterID | Number of proteins | Protein name | Function |
|---|---|---|---|
| TTSS (Type III) | | | |
| Cluster:plasmids:123 | 19 | YscC/HrcC | Secretin |
| Cluster:plasmids:240 | 12 | YscT | Transmembrane protein |
| Cluster:plasmids:241 | 12 | YscR | Transmembrane protein |
| Cluster:plasmids:242 | 12 | YscJ | Lipoprotein |
| Cluster:plasmids:243 | 12 | YscN | ATPase |
| Cluster:plasmids:244 | 12 | YscV | Transmembrane protein |
| Cluster:plasmids:245 | 12 | YscU | Transmembrane protein |
| Cluster:plasmids:273 | 11 | YscS | Transmembrane protein |
| Cluster:plasmids:314 | 10 | YscF* | Type III secretion |
| Cluster:plasmids:315 | 10 | YscQ* | Type III secretion |
| Cluster:plasmids:416 | 8 | YscL/HrpF† | Transmembrane protein |
| T2S (Type II) | | | |
| Cluster:plasmids:33 | 59 | T2SE | ATPase |
| Cluster:plasmids:123 | 19 | T2SD | Secretin |
| Cluster:plasmids:234 | 12 | T2SK‡ | Major pilin |
| Cluster:plasmids:297 | 10 | T2SO | Peptidase |
| Cluster:plasmids:397 | 8 | T2SF | Integral membrane protein |

Protein names and Functions were taken from the review articles cited in the text and these functions were entered in the ACLAME list of functions. The second column shows the number of proteins present in the family in the first column.

*protein not present in *Shigella* plasmids

†protein present only on plasmids harboured by strains pathogenic for animals

‡minor pilins encoded by pGMI1000MP belong to that same family.

The 12 plasmids with Type III systems are: pWR501 from *Shigella flexneri*, pCP301 from *Shigella flexneri 2a*, pYVe227, pYVa127/90 and pYVe8081 from *Yersinia enterocolitica*, 1 pCD1 plasmid from *Yersinia pestis CO92* and 2 from *Yersinia pestis KIM*, p42d from *Rhizobium etli,* pNGR234a from *Rhizobium sp*. NGR234 and pGMI1000MP from *Ralstonia solanacearum*, which in addition to the HPR system carries a second set of similar genes annotated as being involved in flagellar biosynthesis (Salanoubat *et al*., 2002).

Type II systems appeared complete only on pO157 from *Escherichi coli O157-H7* (Schmidt *et al*., 1997) and pGMI1000MP.

## Borrelia plasmids

Twenty three plasmids in the set analyzed (13 linear and 10 circular) originate from *Borrelia burgdorferi* strains. As reported earlier (Casjens *et al.*, 2000), these plasmids are related and redundant but share very limited relationship with plasmids from other species. Hence, they form a good group to assess the reliability of the clustering in ACLAME. A manual comparison of the ACLAME families that contain most of those plasmid-encoded proteins with the Paralogous Families (PF) defined in the Borrelia sequencing project (Fraser *et al.*, 1997, Casjens *et al.*, 2000) shows a good correspondence (Table S4 supplementary materials). All plasmids but lp9 and cp5 encode a protein in the PF32 family, which corresponds to the very large ParA family (cluster:plasmids:2) discussed earlier. However, no recognizable ParB homologue could be seen on any of the plasmids. All but lp36, lp38 and cp26 contributed to cluster:plasmids:40 annotated with the function 'DNA replication initiation' which corresponds to PF57, previously shown to be essential for plasmid replication where it was tested (Eggers *et al.*, 2002, Stewart *et al.*, 2003). The family cluster:plasmids:40 also includes proteins from plasmids hosted by *Aquifex aeolicus*, *Clostridium acetobutylicum* and *Clostridium tetani*. As earlier observed by Casjens *et al.* (2000) a few Borrelia plasmid proteins are found in large ACLAME protein families with a well-defined function. The most notable ones are the ABC transporter-like proteins encoded by cp26, lp38 and lp54, a possible adenine deaminase on lp26 that is also present on lp28-3, or a DNA helicase on lp28-2. There are a few cases of discrepancies between our and the PF protein families. Most of them will not be discussed further because of the absence of experimental evidence for the function of any of the proteins concerned. However one protein, BBA74 encoded by lp54, deserves some comment. It was originally annotated as a porin in a family of two proteins (PF171). In the ACLAME families BBA74 stands alone in cluster:plasmid:3007 and interestingly clusters with proteins related to phage tail proteins in cluster:all:17 in the classification of the combined set of proteins from phages and plasmids (set 'All'). BBA31, another lp54 protein, belongs to the PF145 family of conserved hypothetical proteins, which includes orthologs from plasmids cp32-1 to 9 and lp56. Our combined clustering of plasmid- and phage-encoded proteins puts these proteins within a phage terminase large subunit family (cluster:all:262) with terminases encoded by phage Phi ETA from *Staphylococcus aureus* and phages O1205 and Sfi11 from *Streptococcus thermophilus*. This functional assignment is consistent with the recent finding that some cp32 plasmids are prophages (Eggers *et al.*, 2000). Plasmid cp26 codes for the only protelomerase ResT in *Borrelia burgdorferi* B31 [BBB03, (Kobryn & Chaconas, 2002)]. The enzyme processes all the hairpin ends on linear chromosomes and plasmids in the strain. In the ACLAME set 'All' BBB03 clusters with protelomerases encoded by phage N15 from *Escherichia coli* (Deneke *et al.*, 2000), phage PY54 from *Yersinia enterolitica* (Hertwig *et al.*, 2003), phage PhiKO2 from *Klebsiella oxytoca* (Casjens *et al.*, 2004) and one protein encoded by phage VHML from *Vibrio harveyi* (Oakey & Owens, 2000). The latter may thus be expected to form a linear prophage, as it is the case for the three other phages.

## ORFan genes in plasmids

Amongst the plasmid proteins in ACLAME, some belong to the so-called ORFan category, i.e. are without significant similarity to any known sequence (Fischer & Eisenberg, 1999). The origin of these ORFan genes remains an open question. For some, a function has been demonstrated experimentally (Siew *et al.*, 2004) but some ORFans either could result from an overprediction by ORF prediction methods, or could be 'dead' genes (Amiri *et al.*, 2003) or, on the contrary, rapidly evolving genes (Domazet-Loso & Tautz, 2003). One proposed cause of the existence of ORFans has been related to the low number of sequenced genomes or, in other words, a gapped sequence space. Therefore, it was expected that with the increasing number of sequenced genomes, the number of ORFans would decrease.

We counted the number of ORFans in plasmids and measured the effect of sequence database growth on ORFan frequency. Three versions of the NCBI nonredundant protein sequence database (NRDB) were used, defined by the date of the download from the FTP server (ftp://ftp.ncbi.nih.gov): 18 July 2002, 16 October 2003 and 11 September 2004, here called version 2002, 2003 and 2004 with, respectively, 1034241, 1539396 and 2138225 protein sequences, i.e. a difference of more than one million sequences between 2002 and 2004. Only plasmid ORFs available in 2002 (19 942 sequences) were used to search for similarities with PSI-BLAST 4 iterations and an Evalue threshold of 0.001 in each NRDB version. The number of hits obtained for each protein sequence against each NRDB version was extracted. The percentage of protein occurrence with one to 12 and more hits is shown in Fig. 4. The number of singleton ORFans (proteins with one hit, i.e. self-matches) decreased from 16.5% in 2002 to 11.2% in 2004, while the percentage of proteins with 12 or more hits increased from 63.1% to 70.7%. However, if the proteins from plasmids sequenced in 2003 and 2004 were added in their respective dataset (data not shown), the singleton ORFans, in 2004, represented 13.1% and the proteins with 12 or more hits represented 68.4%. These results suggest that the increase of sequences has an effect on the number of singleton ORFans (decreased by 5.3% between 2002 and 2004). However, adding new

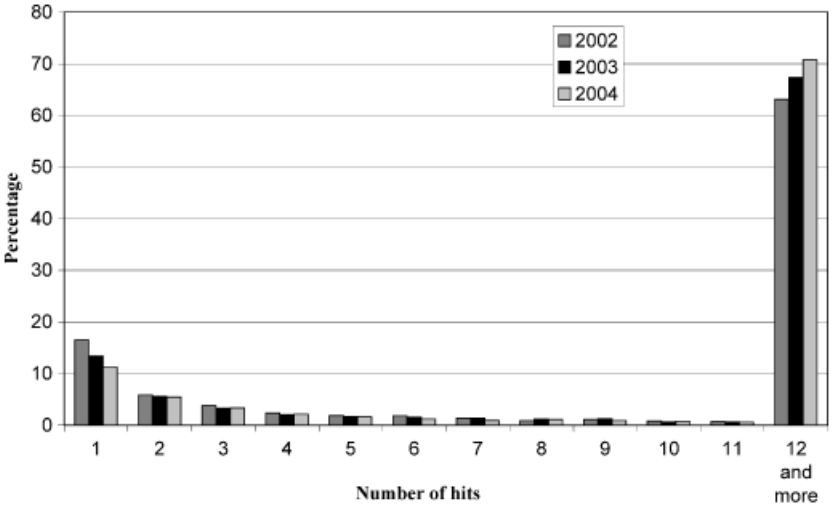


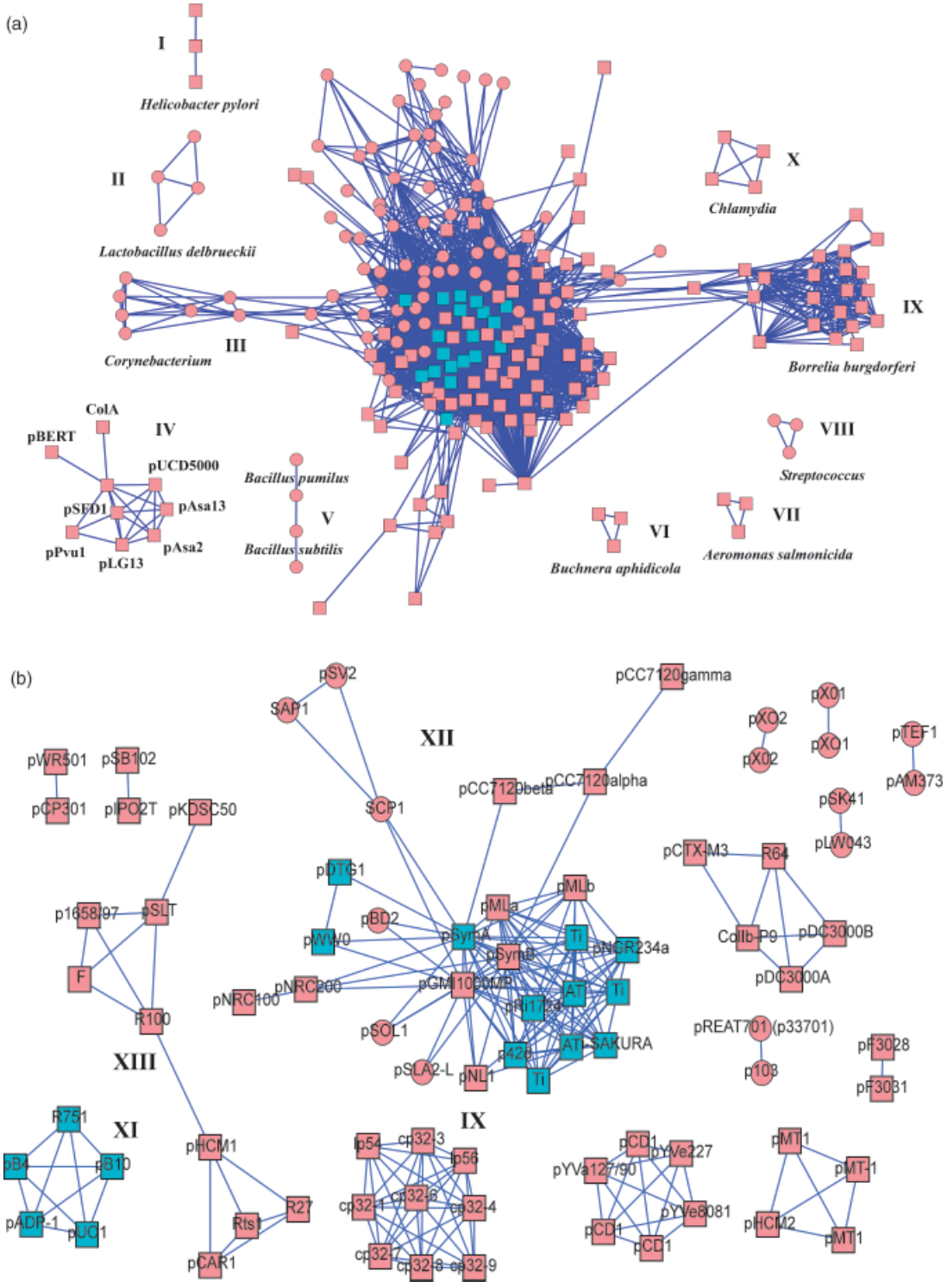

**Fig. 4.** Distribution of the proteins based on their number of hits in NRDB 2002, 2003 and 2004. The graph summarizes the database search results in NRDB-NCBI from 2002 (dark grey), 2003 (black) and 2004 (light grey). Each bar in the graph represents the percentage of plasmid proteins that produced the number of hits given on the X axis. Proteins with one hit are singleton ORFans since matching only with themselves. The percentage of singleton ORFans in plasmids is around 13% as observed in bacterial genomes. For more details see text.

plasmid proteins, and hence more ORFans, lowered the difference between 2002 and 2004 to only 3.4%. This is in agreement with the suggestion by Siew & Fischer (2003) that although the number of ORFans slowly decreases, they will most probably never disappear. In addition, the proportion of singleton ORFans in plasmids (13%) was the same as in bacterial genomes [13% from the ORFanage database (Siew *et al*., 2004)].

## The ACLAME plasmid world

The ability to encode MGE genomes as profiles of ACLAME families opens new perspectives in plasmid (evolutionary) relationships studies. We are for instance developing computational tools that allow for the search of pairs or stretches of genes/proteins that tend to remain adjacent on genomes and which could constitute functional modules. The genomes encoded with protein families can be used to build networks, which allow for a visualisation of the reticulate relationships between all or a set of MGEs, generating what we call a proteomic graph. The method, which was developed for the analysis of the phages in ACLAME, will be described elsewhere. It can be applied to the set of plasmids discussed here. Figure 5 shows two plasmid proteomic graphs drawn using the Cytoscape software [http://www.cytoscape.org/, (Shannon *et al*., 2003)]. The nodes correspond to the plasmid genomes. Since genomes in ACLAME are linked to their host, plasmids hosted by Gram-positive and Gram-negative bacteria can be differentiated (here by circles and squares, respectively). The edges that connect the plasmids are drawn based on the number of protein families shared by the connected nodes. A threshold over the minimum number of shared families can be set. This generates pictures where groups of plasmids progressively separate from the bulk when the threshold imposed on the number of families shared increases. In Fig. 5a, the threshold value was set at four shared families. In addition, the system allows discarding any set of protein families from the analysis. In Fig. 5, families corresponding to transposases and site-specific recombinases were not taken into account. This option was chosen because these proteins are abundant and likely belong to IS, transposons or integrons inserted in plasmids and hence they generate connections whose significance for plasmid-plasmid relationship may be marginal. Graphs generated with those families included produce an overall similar picture although the groups visible in Fig. 5 separate from the bulk at a higher cut-off value (because more protein families are shared), leading to the loss of other interesting relationships (data not shown). The Borrelia plasmids (group IX in Fig. 5a and b) form a group of interconnected plasmids, which nevertheless keep a number of connections with the major group. This is also the case for a smaller group of plasmids from *Corynebacteria* (group III in Fig. 5a). More details on the composition of other individualized groups of various sizes can be found in the legend for Fig. 5. Plasmids that code for the three hallmark proteins of the conjugative machinery, a relaxase in cluster:-plamsids:11, TraG and VirB10, are coloured in green. Only plasmids from Gram-negative bacteria are found with the three hallmark proteins and localise inside the major group of plasmids. As the threshold on the number of shared families increases, groups of more related plasmids progressively separate from the bulk. Figure 5b illustrates what happens with the set of plasmids having relaxase, TraG and VirB10 proteins. At a threshold of 27 shared families, group XI (pB4, pB10, R751 and pU01) individualizes. It remains unchanged at a threshold of 30 (data not shown). Detailed sequence comparison within this group of very closely IncP-1β plasmids has been published (Schluter *et al*., 2003). The remaining plasmids that bear these three features remain

within a much larger group (group XII in Fig. 5b), which still combines genomes from Gram-positive and Gram-negative hosts. These remain so when the numerous ABC transporters present on the agrobateria Ti and rhizobia pSym plasmids, which belong to the group, are removed from the graph (data not shown). Increasing the threshold to 30 (data not shown) individualizes the plasmids on the edges of group XII. SAP1, SCP2, pSV2 form a subgroup as well as the cyanobacteria pCC1720α-γ. However, two large plasmids from *Streptomyces rochei* and *Clostridium*



*acetobutylicum*, pSLA2-L and pSOL1, remain in the group. Inspection of their genome content shows that the first encodes a relaxase but neither encodes any type IV secretion protein. They both, however, encode numerous enzymes (hydrolases, oxidoreductases, lysases) that they share with the other plasmids in the group. This is the same for the plasmids pNRC100, pNRC200 from *Halobcaterium*, and for plasmids from Gram-negative *Alphaproteobacteria* (e.g. *Ralstonia solanacearum* pGMI1000MP), which are not conjugative. Group XIII in Fig. 5b also includes conjugative plasmids as the paradigm IncF plasmids R100 and F and IncH plasmids R27 and pHCM1. However, the proteins that form these conjugation systems are not similar to those from the conjugative plasmids in group XII. At threshold 30, group XIII splits into two groups, which contain, respectively, the five and four plasmids forming the upper and lower part of group XIII (data not shown). The representation of the plasmid relationships as a graph, besides providing a direct view of the current knowledge we have on the plasmid population, allows for easily generating smaller groups of more related plasmids, which can then be analyzed by more traditional alignment methods. Further exploration of the network allows determining clusters of coevolving genes characteristic for a particular type of function in a particular group of plasmids and possibly for a particular type of host. The proteomic network has the advantage over a phylogenetic tree that it does not rely on the existence of a common ancestor. It is better suited to reflect the modularity of plasmids and to infer the transmission of functional modules through the plasmid population.

## Conclusions

This review highlights some of the possibilities in deriving useful biological information out of the sequenced plasmids on a local and global scale. The plasmid analysis discussed above provides a first assessment of statements appearing more and more often in articles on plasmids or grant applications aiming at sequencing new plasmid genomes, on the scantiness of plasmid genomic data. The analysis was conducted using a combination of manual and computational approaches that would have been hardly possible without the structured information compiled in the ACLAME database. The abundance of IS sequences in plasmids is not a surprising conclusion, but that of site-specific recombinases had, to our knowledge, never been emphasized. Their exact function (likely the resolution of plasmid dimmers, in some instances at least) needs to be further investigated. Plasmids sequenced so far contain few transposons, integrons, secretion systems (including T4SS) or phage-like genes (phage tail-like bacteriocins) and it will be interesting to see how this tendency evolves while more sequences become available. This is also true for the abundance of ABC transporters in the sole plant associated plasmids.

There are clearly more tools needed to explore plasmid data deeper and more comprehensively. For example, computational tools can be developed to identify groups of genes/proteins with conserved functions and organization, to define functional modules relating either to plasmid core functions (replication, conjugation, partition etc.) or plasmid carried 'baggage' genes (catabolic functions, toxins, host interaction functions etc.) or to detect MGEs inserted in plasmids, as IS, transposons or integrons. More elaborated studies on the evolution of defined functional modules, the shuffling of modules between plasmids or of individual genes within modules would then become possible, shedding new light on the overall evolutionary behaviour of plasmids.

However, such work will remain hampered by problems linked with the current plasmid data. The first obvious one, as illustrated in this review, is the bias in sequenced plasmids that will be solved only with more sequencing projects focusing on underrepresented bacterial hosts. Ongoing plasmid sequencing projects (e.g. by the Department of Energy in the US) should fill in some of these gaps. Other problems relate to the format of the sequence files. These should be solved at the level of the GenBank files since the ACLAME resource relies on these files to provide tools for further analysis dedicated to MGE and is not a pipeline for primary sequence annotation. The GenBank file format is obviously not fully appropriate for plasmids and rules should be set on how to fill for instance the 'organism' field when the host is not known as for plasmids directly isolated

**Fig. 5.** The plasmid proteomic graph. The nodes represent the plasmids. (a) Edges have been drawn when plasmids are sharing more than three families. Two hundred and sixty seven plasmids had less than three shared families and were not drawn on this graph. When the number of shared families is set at one, 495 plasmids are connected (data not shown). Circle and square nodes correspond to plasmids hosted by Gram-positive and Gram-negative bacteria respectively. Green nodes are plasmids that encode at least three conjugation hallmark proteins, a relaxase (cluster:plasmids:11 with 115 proteins), a TraG-like coupling protein (cluster:plasmids:29, 67, 142, 185, 571 and 761 for a total of 138 proteins) and a VirB10 Mpf component (cluster:plasmids:58 with 36 proteins). Most groups contain plamids hosted by strains from the same species as indicated. Besides ColA, ColE1 and pLG13 from *Escherichia coli*, group X contains related plasmids hosted by other enterobacteria (pUCD500 from *Pantoea citrea*, pBERT and pSFD1 from Salmonellae, pPVu1 from *Proteus vulgaris*) and pAsal plasmids from *Aeromonas salicida* (also a *Gammaproteobacteria*). (b) Edges were drawn when the number of shared families was 27 or more. Colour and shape code is as in A. For a detailed description of groups IX, XII and XIII, see the text.

from environmental samples [see for instance (Smalla *et al.*, 2000)]. Writing the plasmid name in the 'organism' field should be avoided since any item under this field is automatically assigned by the NCBI with a taxonomy number, which for a plasmid does not make sense. Whether the host of the plasmid would have its place there is a matter of debate. This and other standards to be applied while filling in plasmid sequence files with software such as NCBI Sequin (http://www.ncbi.nlm.nih.gov/Sequin/QuickGuide/sequin.htm#DefinitionLine) and EBI WEBIN (http://www.ebi.ac.uk/embl/Submission/webin.html) should be discussed within the concerned scientific community. This is also the case for the design of a common ontology, to be used in plasmid genes and gene products annotations, which will avoid 'free style' text annotations generally useless for computational analysis. Ontologies such as GO are lagging behind for plasmid functions and a robust and consensual list of plasmid functions needs to be built and added in GO. The ACLAME list of functions, which heavily relies on the classification of protein into families, is a first attempt in that direction. However, here again input from experts in the field is essential to build up a correct list of definitions for plasmid related biological processes, components and molecular functions to fit the GO (or any other) ontology. Within ACLAME, a set of reference plasmids from Gram-negative and Gram-positive bacteria for which proteins have been carefully reannotated by experts would add value to the annotation of protein families that contain a member originating from one of those reference genomes. A robust functional annotation would for instance allow filtering networks as those in Fig. 5 on a selection of functions (rather than on protein families) and hence view their distribution among plasmids and/or among the hosts carrying the plasmids. Any contribution from our readers to complete the list of functions, add information on plasmid properties not readily available in ACLAME or generate a reliable plasmid ontology would be more than welcome. The content of all protein families can be downloaded from the ACLAME website and any additional section of the database content can be made available upon request through the ACLAME website.

## Acknowledgements

We are very much indebted to D. Mazel, G. Schroeder, F. van Gijsegem, M. Chandler and P. Siguier for their help in the annotation of some clusters and J. van Helden and S. Wodak for fruitful discussions. J. Gouëllo actively contributed to the analysis of ORFan proteins. This work was supported in part by the Fonds de la Recherche Scientifique Médicale and Fonds de la Recherche Fondamentale Collective, by the European Space Agency, contract ESTEC 16370/02/NL/CK and the Université Libre de Bruxelles within a collaboration with M. Mergeay and P. de Boever at the Laboratory for Microbiology, SCK-CEN, Mol Belgium.

## References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Amiri H, Davids W & Andersson SG (2003) Birth and death of orphan genes in Rickettsia. *Mol Biol Evol* **20**: 1575–1587.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J & Wheeler DL (2005) GenBank. *Nucleic Acids Res* **33**: D34–38.

Cascales E & Christie PJ (2004) Agrobacterium VirB10, an ATP energy sensor required for type IV secretion. *Proc Natl Acad Sci USA* **101**: 17228–17233.

Casjens S, Palmer N, van Vugt R *et al.* (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol Microbiol* **35**: 490–516.

Casjens SR, Gilcrease EB, Huang WM, Bunny KL, Pedulla ML, Ford ME, Houtz JM, Hatfull GF & Hendrix RW (2004) The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. *J Bacteriol* **186**: 1818–1832.

Chandler M & Mahillon J (2003) Insertion Sequences Revisited. Mobile DNA II. (Craig N, Gellert RCM & Lambowitz AM, eds), pp. 305–366. ASM Press, Washington DC.

Cohen SP, Yan W & Levy SB (1993) A multidrug resistance regulatory chromosomal locus is widespread among enteric bacteria. *J Infect Dis* **168**: 484–488.

Cornelis GR & Van Gijsegem F (2000) Assembly and function of type III secretory systems. *Annu Rev Microbiol* **54**: 735–774.

Deneke J, Ziegelin G, Lurz R & Lanka E (2000) The protelomerase of temperate *Escherichia coli* phage N15 has cleaving-joining activity. *Proc Natl Acad Sci USA* **97**: 7721–7726.

Domazet-Loso T & Tautz D (2003) An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res* **13**: 2213–2219.

Ebersbach G & Gerdes K (2005) Plasmid segregation mechanisms. *Annu Rev Genet* **39**: 453–479.

Eggers CH, Casjens S, Hayes SF, Garon CF, Damman CJ, Oliver DB & Samuels DS (2000) Bacteriophages of *Spirochetes*. *J Mol Microbiol Biotechnol* **2**: 365–373.

Eggers CH, Caimano MJ, Clawson ML, Miller WG, Samuels DS & Radolf JD (2002) Identification of loci critical for replication and compatibility of a *Borrelia burgdorferi* cp32 plasmid and use of a cp32-based shuttle vector for the expression of fluorescent reporters in the lyme disease spirochaete. *Mol Microbiol* **43**: 281–295.

Enright AJ, Van Dongen S & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.

Firth N, Ridgway KP, Byrne ME, Fink PD, Johnson L, Paulsen IT & Skurray RA (1993) Analysis of a transfer region from the staphylococcal conjugative plasmid pSK41. *Gene* **136**: 13–25.

Fischer D & Eisenberg D (1999) Finding families for genomic ORFans. *Bioinformatics* **15**: 759–762.

Francia MV, Varsaki A, Garcillan-Barcia MP, Latorre A, Drainas C & de la Cruz F (2004) A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol Rev* **28**: 79–100.

Fraser CM, Casjens S, Huang WM *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**: 580–586.

Frost LS, Leplae R, Summers AO & Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**: 722–732.

Gerdes K, Moller-Jensen J & Bugge Jensen R (2000) Plasmid and chromosome partitioning: surprises from phylogeny. *Mol Microbiol* **37**: 455–466.

Glaser P, Frangeul L, Buchrieser C *et al.* (2001) Comparative genomics of *Listeria* species. *Science* **294**: 849–852.

Harris MA, Clark J, Ireland A *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258–261.

Hertwig S, Klein I, Lurz R, Lanka E & Appel B (2003) PY54, a linear plasmid prophage of *Yersinia enterocolitica* with covalently closed ends. *Mol Microbiol* **48**: 989–1003.

Hickey RM, Twomey DP, Ross RP & Hill C (2001) Exploitation of plasmid pMRC01 to direct transfer of mobilizable plasmids into commercial lactococcal starter strains. *Appl Environ Microbiol* **67**: 2853–2858.

Kobryn K & Chaconas G (2002) ResT, a telomere resolvase encoded by the Lyme disease spirochete. *Mol Cell* **9**: 195–201.

Korner H, Sofia HJ & Zumft WG (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev* **27**: 559–592.

Leplae R, Hebrant A, Wodak SJ & Toussaint A (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32**Database issue: D45–49.

Liebert CA, Hall RM & Summers AO (1999) Transposon Tn21, flagship of the floating genome. *Microbiol Mol Biol Rev* **63**: 507–522.

Merlin C, Mahillon J, Nesvera J & Toussaint A (2000) Gene Recruiters and Transporters: the Modular Structure of Bacterial Mobile Elements. The Horizontal Gene Pool. (Thomas CM, ed), pp. 363–402. Harwood Academic, Amsterdam.

Nesvera J, Hochmannova J & Patek M (1998) An integron of class 1 is present on the plasmid pCG4 from gram-positive bacterium *Corynebacterium glutamicum*. *FEMS Microbiol Lett* **169**: 391–395.

Oakey HJ & Owens L (2000) A new bacteriophage, VHML, isolated from a toxin-producing strain of *Vibrio harveyi* in tropical Australia. *J Appl Microbiol* **89**: 702–709.

Pansegrau W, Lanka E, Barth PT, Figurski DH, Guiney DG, Haas D, Helinski DR, Schwab H, Stanisich VA & Thomas CM (1994) Complete nucleotide sequence of Birmingham IncP alpha plasmids. Compilation and comparative analysis. *J Mol Biol* **239**: 623–663.

Parkhill J, Dougan G, James KD *et al.* (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**: 848–852.

Radstrom P, Skold O, Swedberg G, Flensburg J, Roy PH & Sundstrom L (1994) Transposon Tn5090 of plasmid R751, which carries an integron, is related to Tn7, Mu, and the retroelements. *J Bacteriol* **176**: 3257–3268.

Salanoubat M, Genin S, Artiguenave F *et al.* (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415**: 497–502.

Schell MA (1993) Molecular biology of the LysR family of transcriptional regulators. *Annu Rev Microbiol* **47**: 597–626.

Schluter A, Heuer H, Szczepanowski R, Forney LJ, Thomas CM, Puhler A & Top EM (2003) The 64 508 bp IncP-1beta antibiotic multiresistance plasmid pB10 isolated from a wastewater treatment plant provides evidence for recombination between members of different branches of the IncP-1beta group. *Microbiology* **149**: 3139–3153.

Schmidt H, Henkel B & Karch H (1997) A gene cluster closely related to type II secretion pathway operons of gram-negative bacteria is located on the large plasmid of enterohemorrhagic *Escherichia coli* O157 strains. *FEMS Microbiol Lett* **148**: 265–272.

Serres MH & Riley M (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb Comp Genomics* **5**: 205–222.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B & Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Siew N & Fischer D (2003) Twenty thousand ORFan microbial protein families for the biologist? *Structure (Camb)* **11**: 7–9.

Siew N, Azaria Y & Fischer D (2004) The ORFanage: an ORFan database. *Nucleic Acids Res* **32**Database issue: D281–283.

Smalla K, Heuer H, Gotz A, Niemeyer D, Krogerrecklenfort E & Tietze E (2000) Exogenous isolation of antibiotic resistance plasmids from piggery manure slurries reveals a high prevalence and diversity of IncQ-like plasmids. *Appl Environ Microbiol* **66**: 4854–4862.

Stewart PE, Chaconas G & Rosa P (2003) Conservation of plasmid maintenance functions between linear and circular plasmids in *Borrelia burgdorferi*. *J Bacteriol* **185**: 3202–3209.

Tauch A, Schluter A, Bischoff N, Goesmann A, Meyer F & Puhler A (2003) The 79,370-bp conjugative plasmid pB4 consists of an IncP-1beta backbone loaded with a chromate resistance transposon, the strA-strB streptomycin resistance gene pair, the oxacillinase gene bla(NPS-1), and a tripartite antibiotic efflux system of the resistance-nodulation-division family. *Mol Genet Genomics* **268**: 570–584.

Tran Van Nhieu G, Bourdet-Sicard R, Dumenil G, Blocker A & Sansonetti PJ (2000) Bacterial signals and cell responses during Shigella entry into epithelial cells. *Cell Microbiol* **2**: 187–193.

van Dongen S (2000) Graph clustering by flow simulation. Thesis, Centre for Mathematics and Computer Science, Amsterdam.

Venkatesan MM, Goldberg MB, Rose DJ, Grotbeck EJ, Burland V & Blattner FR (2001) Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. *Infect Immun* **69**: 3271–3285.

Wain J & Kidgell C (2004) The emergence of multidrug resistance to antimicrobial agents for the treatment of typhoid fever. *Trans R Soc Trop Med Hyg* **98**: 423–430.

Zechner EL, de la Cruz F, Eisenbrandt R, Grahn AM, Koraimann G, Lanka E, Muth G, Pansegrau W, Thomas CM & Wilkins BMaZ M (2000) Conjugative DNA transfer processes. The Horizontal Gene Pool. (Thomas CM, eds), pp. 87–174. Harwood Academic Publishers, Amsterdam.

## Supplementary material

The following material is available for this article online at:

**Table S1.** Each functional category name is underlined in the table. The category is followed by its list of ACLAME functions with the ACLAME function ID in the first column, the generic name in the second column and a more specific name in the third column when defined. The same table is available at http://aclame.ulb.ac.be/FEMS_Table_1S.html where the ACLAME function ID is hyperlinked to the clusters or proteins annotated with such function.

**Table S2.** From left to right: UB: Unidentified bacterium; (a): ACLAME function ID; function:374 is unidirectional conjugation, relaxase activity; function:387 is type IV protein secretion system: TraG/VirD4 coupling protein; function:495 is type IV protein secretion system: VirB10, Mpf-coupling protein bridging; function:386, type IV protein secretion system: VirB11, type II/IV secretion NTPase. The numbers in the columns refer to cluster:plasmid:ID, which can be seen in detail on the ACLAME website.

**Table S3.** From left to right: ACLAME cluster:plasmids ID; ACLAME function ID and description (the corresponding resistance phenotype is shown in parenthesis); ACLAME protein ID; Plasmid name; host of that plasmid. More details on the plasmid genomes, the proteins and the clusters can be easily seen on the ACLAME website.

**Table S4.** Column A: the number on the left refers to the Borrelia Project Paralogons family identifier. The number to the right refers to the ACLAME cluster:plasmids identifier. The numbers in parenthesis are the number of Borrelia plasmid proteins and the total number of proteins in the ACLAME cluster respectively. B: Gene/protein names in the Borrelia genomes; C and D: plasmid genomes that encode the protein in column B, in PF and ACLAME respectively.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1574-6968.2006.00044.x (This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.