Annotation of Plasmid Genes

Christopher M. Thomas[1*], Nicholas R. Thomson[2], Ana M. Cerdeño-Tárraga[2a], Celeste J. Brown[3], Eva M. Top[3] and Laura S. Frost[4]

[1]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK; [2]Sanger Centre, Hinxton, UK, CB10 1SA; [3]Department of Biological Sciences, University of Idaho, Moscow, Idaho, 83844-3051; [4]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada, T6G 2E9

[a]Current address: EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

*Author for correspondence

Abstract

Good annotation of plasmid genomes is essential to maximise the value of the rapidly increasing volume of plasmid sequences. This short review highlights some of the current issues and suggests some ways forward. Where a well-studied related plasmid system exists we recommend that new annotation adheres to the convention already established for that system, so long as it is based on sound principles and solid experimental evidence, even if some of the new genes are more similar to homologues in different systems. Where a well-established model does not exist we provide generic gene names that reflect likely biochemical activity rather than overall purpose particularly, for example, where genes clearly belong to a type IV secretion system but it is not known whether they function in conjugative transfer or virulence. We also recommend that annotators use a whole system naming approach to avoid ending up with an illogical mixture of names from other systems based on the highest scoring match from a BLAST search. In addition, where function has not been experimentally established we recommend using just the locus tag, rather than a function-related gene name, while recording possible functions as notes rather than in a provisional name.

Key words: genome; DNA sequence; gene names; gene function; replication; conjugative transfer

Highlights:

Review of current "issues" with plasmid annotation

Overview of annotation developments for other types of genome elements

Suggestions for good practice in plasmid annotation

Discussion of annotation in relation to replication, stability and conjugation functions

Previously unpublished annotation of IncX1 plasmid complete genome sequence as a model

1. Introduction

Efforts to standardize plasmid naming and gene annotation have not kept up with the deluge of data provided by modern high throughput sequencing and automated annotation. In 1976, Novick et al. published a schema for naming plasmids (pXY1,2,3, etc) and the genes they carry. The convention for naming plasmids was generally followed for many years but gradually eroded as new plasmids were reported with increasing frequency. Researchers moved away from this simple naming system and instead used names that reflected the strain, the cloning procedure, the institution's or investigator's initials, etc. This was further exacerbated by the discovery of plasmids in genome and metagenome sequencing projects where no naming protocol exists. This problem has been outlined within a broader examination of microbial elements by Klimke and his associates at NCBI (2011). They have worked to standardize experimental data entry into GenBank and other databases through a portal named COMBREX (Anton et al., 2013). They have also been working with other interested parties on the nomenclature of viruses (Brister et al., 2010), Insertion Sequences (Siguier et al., 2012) and genomes and metagenomes (Markowitz et al., 2014a,b). Other groups have tried to impose order on plasmid names with varying levels of success (Angiuoli et al., 2008; Martinez-Garcia et al., 2011; Seiler et al., 2014; Wang et al., 2009; Zuo et al., 2007). However, no system has received as widespread approval as the original proposal by Novick et al. (1976). Therefore, we encourage researchers to follow the naming scheme from Novick et al. (1976) whereby a plasmid is designated with a small "p" at the beginning of the name followed by a combination of letters and numbers that are unique to that plasmid. Some recently named plasmids have the "p" at the end, which in older nomenclature designated a protein, or in the middle, which lacks the clarity of the initial "p". However, plasmids that were named prior to 1976 should keep their names (F, RP4, ColIb-P9etc) because much confusion could arise if well studied plasmid paradigms were renamed at this point in time. Also, where a convention has developed within a research community, such as those who work on plasmids of Rhizobium, we encourage new researchers to use that system rather than develop something new (Cevallos et al., 2008). In addition, it is worth making sure that the annotation starts at a similar point and progresses in the same direction round the plasmid as annotations of related plasmids already in the databases unless what has gone before is deemed unsatisfactory with good reasons.

A greater area of concern is the annotation (naming) of genes and their gene products belonging to plasmids and associated elements (conjugative transposons, ICEs). Of particular concern are the "backbone" genes that define plasmid maintenance and spread within bacterial populations.

The chaotic naming of plasmid genes is the result of biases in sequence analysis programs such as BLAST and a lack of familiarity with plasmid-encoded functions. This is compounded by the propagation of these errors in automated annotation programs. The International Nucleotide Sequence Database Collaboration (INSDC) including NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and SwissProt (http://web.expasy.org/groups/swissprot/) has made herculean efforts to manually correct and organize gene products into families. Staff at NCBI are re-annotating genomes using the NCBI annotation pipeline and cataloguing them in the Reference Sequence (RefSeq) database (Pruitt et al., 2009; O'Leary et al., 2016; https://www.ncbi.nlm.nih.gov/refseq/). However, ensuring that the extra level of detail required in the latest annotations is unambiguous and error free will require expert input from specialists in the plasmid community to ensure that the results of this effort are fully accepted and used by the community. The authors wish to review current practices and make suggestions, based on the wisdom of members of the International Society for Plasmid Biology that will be adopted by automated annotation services.

2. A brief history of plasmid annotation

The phases of plasmid annotation reflect the history of bacterial genetic analysis and can be split into four, illustrating the predicament that we are now experiencing.

First, there are the historically important plasmids, such as F, whose genes were named based on the order in which the complementation groups were identified using classical bacterial genetics or gene cloning. Thus, we have genes ordered *traALEKB etc* within the F transfer region. This random naming scheme does not reflect the position of the gene within an operon nor does it suggest the presence of genes within separate operons.

Second, we have plasmids whose current naming system was established after manual DNA sequencing became a more routine part of genetic analysis, and where cistrons were often first identified by DNA sequencing and therefore named in order of their occurrence within operons on the plasmid. Good examples of this are RP4 whose two transfer regions contain genes *traABCDE etc* and *trbABCDEFGHIJK* etc (Pansegrau *et al*., 1994). Perhaps the most influential plasmids in this category are the Ti plasmids (Christie and Gordon, 2014), such as pTiC58 that carries operons involved in tumorigenesis in plants named *virA,B,C,D* etc with the genes in each operon named *virA, virB1-11, virC1-2, virD1-4* etc. The *virB* operon defines the type IV secretion system (T4SS) involved in transfer of the tumorigenic DNA (tDNA) to the target plant cell (a process related to plasmid conjugative transfer) whereas the *virD* operon defines the gene

products involved in DNA processing (VirD2 is the relaxase, VirD4 is the coupling protein). Because the mechanism of tumorigenesis provided such powerful insights into the mechanism of T4SS and DNA transfer, its gene products are now often used to define gene families within the databases even if those genes are not actually involved in a "virulence" phenotype (as discussed below).

The third group is comprised of newly discovered plasmids of interest to researchers whose genes are named after homologues in the database often reflecting the top hits in BLAST. These names often do not match the proposed function. For example, genes encoding T4SS proteins are often named after Ti plasmid *vir* genes, based on homology, rather than a role in true virulence. In some cases, genes within an operon or regulon involved in a single process are named using a variety of gene names (often taken from different systems) based on homology rather than function which can lead to confusion.

The fourth group of plasmids is comprised of the thousands of sequences, either of circularized plasmids or contigs suspected to be of plasmid origin, that have fallen out of metagenome projects. Often, their provenance (e.g., host) is unknown and details about their backbone functions are not provided. These sequences languish in databases but their gene products do provide fodder for homology algorithms such as BLAST. If their gene products have been incorrectly named, they perpetuate and propagate these errors and exert undue influence on future analyses.

3. Issues and possible solutions

Thus, we are left with databases that have multiple names for identical proteins, the same name for often distantly related proteins and proteins that are incorrectly named based on their occurrence within an operon encoding other functions. Ideally we should be able to rectify this, if not for annotated plasmids in the databases, then for future annotations. We have previously addressed this issue (Frost and Thomas, 2014) but the context and our thoughts have moved on since then.

3.1 General issues

The issues and a possible solution are illustrated in Table 1. The issues are highlighted by three historically important plasmids, F, R751 and pSK41, selected from RefSeq (NC_000000) and four others from GenBank chosen to illustrate the problems in annotation. There are two IncP plasmids in Table 1 because there are genes/loci annotated for the IncP Birmingham plasmid

sequence (an amalgamation of RP1/RP4/RK2 sequences) that are not featured explicitly in R751. Only backbone genes are presented and even these have not been presented in their entirety for brevity's sake. The most completely annotated reference plasmids are F and R751 that are paradigms for F- and P-type backbone functions. So the issues as we see them are as follows. First, with the exception of the single-stranded binding protein Ssb (although it is TraM in pSK41, Table 1), the names for various homologs vary considerably and in some cases the gene/gene product names are very plasmid-specific - for example *trw* for transfer of IncW plasmids (see Table 1; R7K). Second, there are gaps within the gene clusters which underline the difficulty of recognizing homologs even though there is a reasonable expectation of them being present. An example is the partitioning system in R7K (IncW), if there is one. There is also the difficulty of getting automated processes to call cis-acting sequences such as the origin of vegetative or conjugative transfer (*oriV*, *oriT*) in pNDM-1_Dok01 or pRA3, or the partitioning centromere in most systems with the exception of F (*sopC*). Proteins such as propilin and the entry exclusion protein are difficult to predict because of their low sequence identity with homologues such as in R7K (IncW). Third, some gene functions have been identified but the locus tag (see below) remains as the name of the gene (pNDM-1_Dok01_N0219 for the soluble transglycosylase Slt, or pRA3.23 for VirB7/TivB7) and some genes are named after homologs found using BLAST such as VirB2-11 in the IncU conjugative plasmid pRA3.

NCBI has made a welcome effort to clarify the annotation of genome sequences, including plasmids, by re-annotating them using the NCBI annotation pipeline based on their criteria for acceptable annotation (Angiuoli et al., 2008; O'Leary et al., 2016). They have also tackled the problem of redundant protein sequences by assigning an NP tag to each protein whose non-redundant RefSeq protein record is then assigned a WP tag. Thus in RefSeq NC001735 for the R751 plasmid, TrfA1 (locus tag R751p25), the replication protein, is given the protein id NP_044236 that links to WP_010890124 "which represents a single, non-redundant, protein sequence which may be annotated on many different RefSeq genomes from the same, or different, species." This is summarized at https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/.

What can we learn from this? First, there is a repository of reasonably thoroughly annotated plasmids at RefSeq (identified by the locus line, e.g., NC_000000) that can serve as paradigms for plasmid sequences. The core backbone genes and proteins that we identified from records in GenBank for seven plasmids are shown in Table 1. Table 1 also lists the principal functions (column 1), the suggested names for the genes and their products within these groups (/gene;

/product; column 2) and their proposed function in the /note (comment) in column 3. Second, each gene and genetic locus (that is something other than a protein coding sequence or CDS) on a plasmid is given a locus tag (the range of locus tag numbers is shown after the accession number in columns 4-10) that represents a unique designation for that gene within that plasmid sequence. In the face of many confusing names for gene products with the same overall function, the locus tag leads the investigator to the given name (for example TrfA1), its putative function (eg initiation of replication or activation of *oriV*) and its membership in a family (in this case of the family of TrfA proteins designated pfam07042).

While this is reassuring, the alphanumeric designation for sequences, genes and gene products is not as intuitively gratifying as seeing old-fashioned names for genes and their protein products. At this late stage, we strongly suggest that researchers initially use locus tags to identify coding sequences in sequential fashion and only use names for those genes/genetic loci that are known or strongly suspected to be involved in functions such as replication, partitioning, stability and conjugation, the main backbone functions of plasmids. It is also important to use gene names in a logical and transparent way. Thus in enterococcal plasmids there are a set of conjugative transfer genes uniquely named *prg* (for example in pCF10 transfer region, AY855841), for **p**heromone **r**esponsive **g**ene on the basis of the first phenotype by which they were identified whereas now it is common to equate *prg* with transfer genes rather than their regulatory mode. We suggest that it is acceptable to propagate such names for genes and their products as long as they are unambiguous.

Similarly, many plasmid gene products are too well known to change their names at this late date. For instance, new IncP plasmids should maintain the IncP-specific gene names such as *trf* (*trans*-acting replication function*), kor* (*kil* over-ride for genes that turned out to encode DNA binding proteins that repress transcription*)* and *kla/klc/kle* (Kil locus A, C and E) that are in common usage (see BN000925and U67194). It also makes sense that new plasmids that are closely related to these paradigms should be annotated using the same set of gene names, for instance *tra* and *trb* for the transfer genes of IncP plasmids. On the other hand, it might not be appropriate to use KorA as the name for a close homolog of KorA unless it is known to regulate a *kil* gene (a gene that is unclonable due to a bacteriostatic or bacteriocidal effect if unregulated). Thus we recommend that gene names reflect gene function as much as possible and we definitely recommend against plasmids being annotated using a mixture of names based on the top BLAST hits.

Below is a discussion of these main concepts.

3.2 Replication functions

Replication is an absolute requirement for plasmid survival. Therefore all plasmids should have an origin of vegetative replication, *ori* (or *oriV* to distinguish it from the conjugative transfer replication origin, *oriT*) and most plasmids should have a *rep* gene. Some plasmids have multiple *rep* genes as in the IncF (Table 1) and IncHI1 plasmid groups because there are multiple replicons whereas others like IncQ plasmids have a single replicon that is more complex and requires multiple *rep* genes – see below). Some plasmids, for example ColE1, do not encode a *rep* protein because their replicon consists of an *ori* that is activated by an RNA transcript produced by RNA polymerase. For many plasmids, the *rep* gene is easily identified by BLASTX if it is related to an already characterised plasmid group. The *oriV* is an A+T-rich region that can often contain or is adjacent to multiple small repeats called iterons in forward or reverse orientations to one another. These iterons can be the basis for a phenotype called incompatibility (*inc*) whereby closely related plasmids ie, ones with the same iteron sequences, are unable to be stably inherited in the same cell-line. Alternatively incompatibility can result from the tight control over replication exerted by regulatory RNA molecules as in the IncFII replicons. In the Enterobacteriaceae, plasmids can be classified by comparison to known replicons using computer algorithms described by Carattoli et al. (2014). This assigns a new plasmid to a sequence group that corresponds to a putative incompatibility group (Inc). However, it needs to be stressed that while this is a useful classification system in its own right, incompatibility assays are still required to demonstrate the phenotype of incompatibility (Thomas, 2014).

Names for gene products should represent protein function such as replication (e.g., Rep) rather than a phenotype such as incompatibility (Inc) or copy number control (Cop). Thus acceptable names for replication proteins are Rep, RepA, RepB etc. If the incompatibility group (Inc) is known, this information can be given in the /product line in the annotation. Thus the gene product RepB of plasmid F (NC_002483; locus tag D616_p97094 or old locus tag Fpla035) is noted as the RepFIB replication protein in the /product line. The repeat regions that define the iterons are described separately in the /note lines as RepFIB repeat sequences. The *ori* sequences can be *ori*, *oriV* (the vegetative replication origin as noted above), *oriS* (a secondary origin identified when the primary *oriV* was deleted), or *ori-1*, *ori-2* as has been used for F in the past. Similarly, in plasmids that replicate via rolling circle (RC) replication, the /product line could indicate Rep(RC) as in the Gram-positive plasmid pSK41 (NC_005024).

An important exception is TrfA (already referred to above, its name being derived from ***t****rans*-acting **r**eplication **f**unction when its role was not clear), the replication protein of plasmids belonging to *E. coli* plasmid incompatibility group IncP (Pseudomonas plasmid incompatibility group IncP-1) (Pansegrau *et al*., 1994). Because of its historical significance, new replication proteins related to TrfA should also be named TrfA. However, it should be noted that a BLAST search with such a protein will identify many homologues that are called Transcriptional Regulator rather than Replication Initiation Protein, illustrating the way in which misinformation about the true function of a protein can be propagated. Homology in this class of proteins is usually based on the type of DNA binding domain within the protein, a useful first step that overlooks its true function. The interested investigator needs to manually identify the hallmarks of a replication region (*rep, ori* and possibly nearby *par* genes) before assigning the name Rep and the proposed function of replication initiator protein.

Another exception is the *rep* gene of IncX1 plasmid R6K which is called *pir* (**p**rotein for the **i**nitiation of **r**eplication) and encodes a protein called Pi (the Greek letter π) (Stalker *et al*., 1982). Although there have been a number of publications covering IncX plasmids in recent years and a number of complete plasmid sequences of much more recently isolated IncX plasmids, we use this occasion to deposit the R6K sequence in EMBL (accession number LT827129) and report the complete annotation of the R6K genome following the principles proposed in this short paper (Supplementary Data Table S1). This is significant because its replication system involves multiple origins as well as a terminator (*ter*) (Sista *et al*., 1991).

Some plasmids have multiple replication genes, the best studied being the IncQ plasmids which encode a helicase and a primase in addition to a "normal" origin binding protein (Meyer, 2009). The genes encoding these proteins were named *repA*, *repB* and *repC* before biochemical characterisation revealed RepA as the helicase, RepB as the primase and RepC as the iteron-binding *oriV*-activator. Therefore in this system RepA is not equivalent to RepA in many other systems. In addition, in the IncQ system there is a very close relationship between replication and mobilisation functions: RepB is produced by an internal translational signal within the *mobA* open reading frame (orf). In cases of such complexity, the new orf should be named using its locus tag until the system is adequately characterised, providing a neutral solution to the problem.

A number of other proteins such as the single-stranded DNA binding protein Ssb, encoded by *ssb*, as well as genes involved in stability (*stb*) are often found in large plasmids. It is not always clear what basic plasmid process these are associated with but in the case of *ssb* we know that it encodes

an accessory protein in replication, either vegetative or conjugative, and is therefore classed as a replication gene. During annotation, gene products that have high sequence identity to well-described accessory proteins such as these can be named with some confidence. Others should be left as locus tag designations and their putative function stated in the /product line.

3.3 Partitioning functions

Partitioning refers to the distribution of newly replicated plasmids into daughter cells after cell division. In general, it is a feature of large, low copy number plasmids that cannot rely on random distribution through a "safety in numbers" mechanism. Three main types of partitioning systems have been described in plasmids: I, II and III with I subdivided into Ia and Ib. In addition to an NTPase, there is a centromere sequence and a centromere-binding protein CBP (Schumacher, 2012) with the NTPase and CBP defining the groups I, Par A,-B; II, ParR,-M; and III, TubZ,-R. The most difficult partitioning proteins to predict are the type Ib CBPs that vary in structure considerably – the putative *cbp* gene of R6K being an example (Supplementary Data Table S1, CDS R6K0033). In general, the NTPases of group Ia and the CBP of Ib, II, and III autoregulate *par* expression. Thus, DNA-binding proteins originally identified as repressors were later shown to be CBPs involved in partitioning. An example of this is KorB from the IncP plasmids, which is a Ia CBP. Unfortunately, CBPs in annotated sequences are often described as repressors and their role in partitioning is overlooked. Again, this requires that the context of the gene within a region be examined manually since computer algorithms are currently unable to connect position to function. For instance, since plasmid partitioning regions contain three characteristic sequences, if one is identified, the other two should be nearby.

In terms of annotation, we recommend using the nomenclature for *par* systems already in existence, namely ParA,-B, ParR,-M and TubZ,-R and historically important names such as SopABC in F and IncC (ParA) KorB (ParB) in IncP plasmids. The *par* group and identification as belonging to a protein family (pfam) should be mentioned in the /function and /note lines during annotation. If the CBP coding sequence is not immediately apparent, the gene should be referred to by its locus tag and putative function mentioned elsewhere as shown in Supplementary Data Table S1.

3.4 Conjugation functions

This is probably the thorniest function or set of functions to annotate because of the variation in conjugative mechanisms and the often low sequence identity among members of a particular pfam group. The key protein in conjugation is an AAA+ ATPase of the pfam VirD4, called the

coupling protein or T4CP, a distant relative of the chromosome segregation protein FtsK and the sporulation protein SpoIIIE (Moncalian *et al*., 1999). T4CPs enable the transport of DNA through a pore formed during cell division, sporulation and conjugation. In some Gram-positive and archaeal plasmids, conjugation only requires this protein, named Tra, and a few inessential accessory genes for plasmid spread (*spd*) etc, for the transfer of double-stranded DNA. A more complete discussion of the requirements for conjugation and the role of the T4CP are discussed in Smillie et al., (2010).

In more complex systems, an endonuclease or relaxase (also nickase) cleaves the plasmid in a site-specific, single-stranded manner to initiate transfer of a single-strand of DNA covalently bound at its 5' end by the relaxase. Together with accessory proteins that direct the relaxase to the cleavage site *oriT* or *nic* and coordinate interactions with the T4CP, they form the relaxosome or Dtr (**D**NA **tr**ansfer) complex (Smillie et al., 2010; Guglielmini et al., 2012).

The bridge between the donor and recipient cells is the result of the activity of type IV secretion systems (T4SS) that can vary substantially in complexity and protein identity. These proteins are involved in **m**ating **p**air **f**ormation or Mpf. In Gram-negatives, an extracellular filament, the pilus, is assembled by the T4SS and is involved in identifying competent recipient cells. Originally pili were found to be of two broad two types – long, thin and flexible (F-like) and short and rigid (P-like) named after the F and P plasmids with which they were first associated. Currently, eight different T4SS systems, including the less studied I-like systems, have been identified as discussed by Guglielmini et al. (2014) with more surely to come. All Gram-negative and Gram-positive ssDNA transfer systems contain an ATPase of the VirB4 family that is responsible for protein secretion (Guglielmini et al., 2014). A second Mpf ATPase, VirB11, is found in a large subset of these systems whereas Mpf$_F$ systems lack a VirB11 homologue but instead have additional proteins involved in mating pair stabilization (Mps) and pilus assembly and retraction.

Other key proteins in Gram-negative T4SS are the VirB7,-9,-10 complex (Fronzes et al., 2009), the VirB6,-8 complex that completes the mating bridge and the more obscure VirB2,-B3,-B5 proteins involved in pilus assembly. The pilus protein itself can be represented by F-like pilin (Costa et al., 2016), a linear, acetylated polypeptide (TraX is the acetylase in F) and by P-pilin, an unusual circular polypeptide that requires a peptidase/cyclase protein (TraF in IncP plasmids) for maturation (Table 1). As sequences accumulate in the databases, it is apparent that both F- and P-like T4SS can assemble P-like pili whereas F-like pili are assembled by F-like T4SS alone. Examples include the IncA/C plasmid pNDM-1_Dok01 (Table 1) and the IncHI1 plasmid R27

that encodes TrhF, which completes the processing and cyclization of the TrhA protein within an otherwise classic F T4SS (Rooker et al., 1999).

What is a beleaguered annotator to do with all this variation in mechanism, sequence and synteny of genes responsible for conjugation? In general, we recommend simplicity with the limitation that genes are not just named after the best known member of their family but are given a name that reflects their biochemistry where that is clear. For example, genes should not be named *vir* unless there is evidence that they contribute to virulence. They may belong to Vir pfams as denoted in the /note or /product lines but their name should be more reflective of their structural or enzymic nature. We recommend that T4SS proteins, when encountered, be named TivB1-11 (Tiv stands for **T**ype **IV**; Table I; Supplementary Data Table S1), which keeps the B1-11 designations of the VirB proteins (but see below). The R6K sequence also raises an interesting question about annotating genes that are fusions of two adjacent orfs in a well studied system. In our sequence of R6K and a number of other IncX plasmids (such as pNGX2-Qnr51, pYD786 and pEGB1) already in the databases a gene that is clearly a fusion of *virB3* and *virB4* is called variously *pilX4*, *pilX3_4* or *pilX3-4*. We recommend that this gene is called *tivB3-4* to indicate its hybrid nature. As for VirD4, using the name TivD4 is unsatisfactory because the coupling protein is not required for Type IV protein secretion. The Tiv nomenclature should be reserved for the proteins that form the trans-envelope complex required for secretion. We suggest that the term Rlx and Cpl be used as an appropriate name for relaxase and coupling protein genes, respectively. Other existing names for the relaxase such as Nic and Nes (Table I) or TaxA,-B,-C (R6K see Supplementary Data Table S1; Núnez et al., 1997) should be discouraged in future annotation projects.

Table 1 illustrates various attempts to come to terms with naming T4SS genes and their products. The IncA/C plasmid pNDM-1_Dok01 has a circular P-type pilin subunit named TraA, which is also the name for the historically important linear F-type pilin subunit. It is processed by the peptidase/cyclase TrhF, a name derived from TrhF from the IncHI1 plasmid R27 involved in the maturation of the circular TrhA pilin. The name TrhF is, in turn, derived from the TraF peptidase/cyclase of IncP plasmids (Table 1) which was first referred to as a peptidase in the *traF* /function= " peptidase / maturation of TrbC pilin protein" of IncP plasmid pKJK5(AM261282). The T4SS gene products in the IncU plasmid pRA3 are named after their closest homologues, VirB2-11, which suggests these proteins having a role in virulence (Table 1). These names could easily be changed to the TivB1-11 nomenclature. A further refinement would be to designate whether the propilin is F- or P-like by using TivF1 and TivB2 (since the Ti plasmid VirB system is P-like)

respectively and TivF2,-F3, etc for the other essential gene products in F-type T4SS (Table 1, column 2). Núnez (1998) foresaw the problems in T4SS nomenclature and suggested PilX1-11 for the T4SS of the IncX plasmid R6K. However, to avoid confusion we suggest that the TivB1-11 nomenclature be adopted, as illustrated in Supplementary Data Table S1. With the realization that Gram-positive and archaeal conjugative systems also use a modified T4SS, albeit with no visible pili, and in cases where no incompatibility group is known, we suggest using TivB1-11for the appropriate homologues as the default nomenclature (see Supplementary Data Table S1).

In Gram-positive bacteria, beside the relaxase (Rlx), T4CP (Cpl) and VirB4 (TivB4) homologues, the soluble lytic transglycosylase (Slt), usually non-essential in Gram-negative bacteria (Koraimann, 2003), acquires increased importance and is key in identifying a conjugative system (Abajy et al., 2007; Goessweiner-Mohr et al., 2013). We suggest these enzymes be named *slt* rather than VirB1 to reflect their function. Guglielmini et al. (2014) make the important point that the presence of a VirB4 family member signals the possible presence of a T4SS especially when accompanied by TivB4, Cpl and Rlx homologues. All of the selected plasmids have a coupling protein, a relaxase and a T4SS NTPase of the VirB4/CagE superfamily (Table 1). The presence of an *slt* gene in most of these plasmids in Table 1 also confirms the presence of a putative T4SS that must span the cell wall.

The presence of an F-like TraN (*tivF6*), a **m**ating **p**air **s**tabilization protein (Mps), is characteristic of F-like T4SS conjugative systems and is usually the easiest of the F T4SS gene cluster (Table 1) to pick out because of its large size and high cysteine content (Lawley et al., 2003). When manually annotating plasmids, finding one or more of these proteins should trigger a further search for other components of the conjugative Dtr and Mpf/T4SS as mentioned above. We recommend TivF1, TivF2 etc (Table 1, column 2) to designate these proteins, which are essential for transfer and are specific to F-type T4SS (Lawley et al., 2003).

Other "transfer" genes and proteins actually reduce transfer efficiency. These include proteins that block mating pair formation (**S**urface **ex**clusion or Sfx), block DNA entry (**E**ntry **ex**clusion or Eex) and reduce transfer gene expression (**F**ertility **in**hibition or Fin). We encourage investigators to not refer to these genes as *tra* genes.

DNA binding proteins, a subject that extends well beyond the scope of this review, are often encoded by plasmids and can be involved in replication, partitioning, relaxosome formation or control of transcription. Unless their function is known, they should be left as locus tags, Orfs

(open reading frames) or Upfs (Unknown protein function) and their similarities to known DNA binding proteins and their putative functions noted on separate lines of the annotation.

5. Conclusions

What we have tried to do in this short review is to prompt the reader to think about the problems associated with plasmid annotation and some ways of minimising these problems for the future. We are not saying that all plasmids need to be re-annotated or even that all new plasmids need to be annotated in exactly the same way. But we feel it is important that people think more critically about the annotation process and base it on a better understanding of plasmids and the evidence needed to establish the function of a gene in the replication, maintenance and transfer of that plasmid.

One solution is for each plasmid to have a unique name and for its gene names to consist of a unique subset of these letters plus sequential numbering around the plasmid i.e. the locus tag. The (putative) gene function can be indicated as a qualifier which can be edited as more is learnt. Such annotation can be supplemented with gene names that have more "meaning" so that a functional plasmid map can be easily interpreted based on well understood gene names. We would support this so long as the gene names chosen are not misleading with reference to function and do not propagate errors.

Backbone genes on newly discovered novel plasmids, ICEs (Integrative Conjugative Elements) and even contigs that are likely to be novel plasmids should be named using common terms such as *rep, ori, par, stb, rlx, nic, cpl, tiv, slt, pep, eex, sfx, ssb, fin.* These would reflect their biochemistry and avoid assumptions about function. Also to be avoided is naming genes of unknown function based on their inclusion in operons of predicted function. Thus genes within *rep, par* or *tra* operons/regions, for instance, which have no known homologues, should remain as *orf*s or be referred to by their locus tags until there is experimental proof for their function. Examples include DNA binding proteins and hard-to-predict proteins involved in surface or entry exclusion that are often present within operons for T4SS gene products.

Supplementary Data Table S1 illustrates these principles applied to the complete genome of IncX plasmid R6K. We have used existing nomenclature derived from previous studies of subsections of the plasmid where appropriate (Núnez et al., 1997; Núnez, 1998) but have also applied the principles proposed in this review for features such as the putative partitioning functions and the

T4SS associated with conjugative transfer. We hope this will prompt discussion within the community about this important topic.

In summary, annotation guided by historical paradigms is acceptable if the new plasmid sequence represents a close family member but for other plasmids, a consistent set of names based on established functions is recommended. With time, these names should populate databases and appear as the top hits in BLAST searches etc. Hopefully this will help reduce the ambiguity generated by current algorithms and extend our understanding of plasmid evolution.

6. Acknowledgements

The sequencing and annotation of R6K was supported by core Sanger Institute funding from the Wellcome Trust. The authors are grateful for constructive discussions on this topic with numerous people at NCBI, EBI, the Sanger Institute and at the biennial International Conferences on Plasmid Biology. We would particularly like to mention Bill Klimke who raised the issue at Plasmid Biology 2010 in Bariloche, Argentina.

7. References.

Abajy MY, Kopeć J, Schiwon K, Burzynski M, Döring M, Bohn C, Grohmann E. (2007) A type IV-secretion-like system is required for conjugative DNA transport of broad-host-range plasmid pIP501 in gram-positive bacteria. J Bacteriol. 189: 2487-96.

Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O. (2008**)** Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS. 12: 137-41. doi:10.1089/omi.2008.0017.

Anton BP, Chang YC, Brown P, Choi HP, Faller LL, Guleria J, Hu Z, Klitgord N, Levy-Moonshine A, Maksad A, Mazumdar V, McGettrick M, Osmani L, Pokrzywa R, Rachlin J, Swaminathan R, Allen B, Housman G, Monahan C, Rochussen K, Tao K, Bhagwat AS, Brenner SE, Columbus L, de Crécy-Lagard V, Ferguson D, Fomenkov A, Gadda G, Morgan RD, Osterman AL, Rodionov DA, Rodionova IA, Rudd KE, Söll D, Spain J, Xu SY, Bateman A, Blumenthal RM, Bollinger JM, Chang WS, Ferrer M, Friedberg I, Galperin MY, Gobeill J, Haft D, Hunt J, Karp P, Klimke W, Krebs C, Macelis D, Madupu R, Martin MJ, Miller JH, O'Donovan C, Palsson B, Ruch P, Setterdahl A, Sutton G, Tate J, Yakunin A, Tchigvintsev D, Plata G, Hu J, Greiner R, Horn D, Sjölander K, Salzberg SL, Vitkup D, Letovsky S, Segrè D, DeLisi C, Roberts RJ, Steffen M, Kasif S.

(2013) The COMBREX project: design, methodology, and initial results. PLoS Biol. 11: e1001638. doi: 10.1371/journal.pbio.1001638.

Brister JR, Bao Y, Kuiken C Lefkowitz EJ, Le Mercier P, Leplae R, Madupu R, Scheuermann RH, Schobel S, Seto D, Shrivastava S, Sterk P, Zeng QD, Klimke W, Tatusova T (2010) Towards Viral Genome Annotation Standards, Report from the 2010 NCBI Annotation Workshop Viruses 2: 2258-2268

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, H. (2014) *In Silico* detection and typing of plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrob Agents Chemother. 58: 3895–3903.

Cevallos MA, Cervantes-Rivera R, Gutiérrez-Ríos RM (2008) The *repABC* plasmid family. Plasmid 60, 19–37.

Christie PJ, Gordon JE (2014). The Agrobacterium Ti Plasmids. Microbiol Spectr. 2. doi: 10.1128/microbiolspec.PLAS-0010-2013.

Costa TR, Ilangovan A, Ukleja M, Redzej A, Santini JM, Smith TK, Egelman EH, Waksman G. (2016) Structure of the Bacterial Sex F Pilus Reveals an Assembly of a Stoichiometric Protein-Phospholipid Complex. Cell 166: 1436-1444. e10. doi: 10.1016/j.cell.2016.08.025.

Fronzes R, Christie PJ and Waksman G (2009). The structural biology of type IV secretion systems. Nature Rev., **7**: 703–714.

Frost, L.S. and Thomas, C.M. (2014) Naming and annotation of plasmids. Part of the "Plasmid Genomes" section of Molecular Life Sciences: An Encyclopedic Reference, edited by Ellis Bell, Judith S Bond, Judith P Klinman, Bettie Sue Siler Masters and Robert D Wells, www.SpringerReference.com.

Goessweiner-Mohr N, Arends K, Keller W and Grohmann E (2013) Conjugative type IV secretion systems in Gram-positive bacteria. Plasmid 70: 289–302. doi: 10.1016/j.plasmid.2013.09.005

Guglielmini J, de la Cruz F, Rocha EPC (2012) Evolution of Conjugation and Type IV Secretion Systems. Mol. Biol. Evol. 30: 315–331 doi:10.1093/molbev/mss221

Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, Rocha EP (2014) Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 42: 5715-27. doi: 10.1093/nar/gku194.

523 Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD,
524 Tatusova T (2011) Solving the Problem: Genome Annotation Standards before the Data
525 Deluge. Stand Genomic Sci 5: 168-193.
526 Koraimann G. (2003) Lytic transglycosylases in macromolecular transport systems of Gram-
527 negative bacteria. Cell Mol Life Sci. 60: 2371-88.
528 Lawley TD, Klimke WA, Gubbins MJ, Frost LS (2003) F factor conjugation is a true type IV
529 secretion system. FEMS Microbiol Lett. 224: 1-15.
530 Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani
531 I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A,
532 Ivanova NN, Kyrpides NC. (2014a) IMG/M 4 version of the integrated metagenome
533 comparative analysis system. Nucleic Acids Res. 42 (Database issue):D568-73. doi:
534 10.1093/nar/gkt919.
535 Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke
536 T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NN,
537 Kyrpides NC (2014b) IMG 4 version of the integrated microbial genomes comparative
538 analysis system. Nucleic Acids Res. 42 (Database issue):D560-7. doi:
539 10.1093/nar/gkt963.
540 Martínez-García E, Calles B, Arévalo-Rodríguez M, de Lorenzo V (2011) pBAM1: an all-
541 synthetic genetic tool for analysis and construction of complex bacterial phenotypes.
542 BMC Microbiol. 11:38. doi: 10.1186/1471-2180-11-38.
543 Meyer R (2009) Replication and conjugative mobilization of broad host-range IncQ plasmids.
544 Plasmid 62: 57-70.
545 Moncalian G, Cabezon E, Alkorta I, Valle M, Valpuesta, JM, Goni, FM, de la Cruz F (1999)
546 Characterization of ATP and DNA binding activities of TrwB, the coupling protein
547 essential in plasmid R388 conjugation. J Biol Chem 274:36117-24
548 Novick RP, Clowes RC, Cohen SN, Curtiss R, Datta N, Falkow S (1976) Uniform nomenclature
549 for bacterial plasmids: a proposal. Bacteriol Rev 40:168-189.
550 Núnez B. (1998) Ph.D. Thesis.Departamento de Biologia Molecular, Universidad de Cantabria,
551 Facultad de Medicina, Cardena Herrera Oria sn, Santander, 39011, SPAIN
552 Núnez B, Avila P, de la Cruz F (1997) Genes involved in conjugative DNA processing of
553 plasmid R6K. Mol Microbiol 24: 1157-1168.
554 O'Leary NA, Wright MW, Brister RJ, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B,
555 Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V,
556 Vyacheslav Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta
557 T, Haft D, Hatcher E, Hlavina W, Joardar VS,Kodali VK, Li W, Maglott D,Masterson P,

McGarvey KM, Murphy, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44, Database issue D733–D745 doi: 10.1093/nar/gkv1189

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

Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 37 (Database issue):D32-6. doi:10.1093/nar/gkn721.

Rooker MM, Sherburne C, Lawley TD, Taylor DE (1999) Characterization of the Tra2 region of the IncHI1 plasmid R27. Plasmid 41: 226-39.

Schumacher M (2012) Bacterial plasmid partition machinery: a minimalist approach to survival. Curr Opin Struct Biol. 22: 72–79. doi:10.1016/j.sbi.2011.11.001.

Seiler CY, Park JG, Sharma A, Hunter P, Surapaneni P, Sedillo C, Field J, Algar R, Price A, Steel J, Throop A, Fiacco M, LaBaer J. (2014) DNASU plasmid and PSI: Biology-Materials repositories: resources to accelerate biological research. Nucleic Acids Res. 42 (Database issue):D1253-60. doi: 10.1093/nar/gkt1060.

Siguier P, Varani A, Perochon J, Chandler M (2012) Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments. Methods Mol Biol 859: 91-103.

Sista PR, Hutchinson CA, Bastia D (1991) DNA-Protein interaction at the replication termini of plasmid R6K. Genes & Development 5: 74-82.

Smillie C, M. Garcillán-Barcia P, M. Francia V, Rocha EPC, de la Cruz F (2010) Mobility of Plasmids. Microbiol Mol Biol Rev. 74: 434–452.

Stalker DM, Kolter R, Helinski DR (1982) Plasmid R6K DNA-Replication .1. Complete Nucleotide sequence of an autonomously replicating segment. J Mol Biol 161: 33-43

Thomas CM (2014) Plasmid incompatibility. Part of the "Plasmid Genomes" section of Molecular Life Sciences: An Encyclopedic Reference, edited by Ellis Bell, Judith S Bond, Judith P Klinman, Bettie Sue Siler Masters and Robert D Wells, www.SpringerReference.com.

Wang Z, Jin L, Yuan Z, Wegrzyn G, Wegrzyn A (2009) Classification of plasmid vectors using replication origin, selection marker and promoter as criteria. Plasmid. 61: 47-51. doi: 10.1016/j.plasmid.2008.09.003.

Zuo D, Mohr SE, Hu Y, Taycher E, Rolfs A, Kramer J, Williamson J, LaBaer J (2007) PlasmID: a centralized repository for plasmid clone information and distribution. Nucleic Acids Res. 35(Database issue):D680-684. 12

Websites:

FTP Directory of Genomes/Plasmids: https://www.ncbi.nlm.nih.gov/genome/browse/

International Nucleotide Sequence Database Collaboration (INSDC):

http://www.insdc.org/

NCBI GenBank home page: https://www.ncbi.nlm.nih.gov/genbank/

NCBI Reference Sequence Database:

http://www.ncbi.nlm.nih.gov/refseq/

https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

Swiss-Prot Group:

http://web.expasy.org/groups/swissprot/

Table 1. Plasmid core functions: generic names plus names of paralogs in examples of different well-studied plasmids.

	Suggested names	Known or putative functions	Examples of different well-studied plasmids						
Inc group, plasmid name			IncA/C, pNDM-1_Dok01	IncF, F	IncPα, RP4	IncPβ, R751	IncU, pRA3	IncW, R7K	IncpSK1, pSK41[1]
Accession no.; locus tag			AP012208; Ndm1Dok1_n0001-0224	NC_002483; D616_p97001-107 or Fpla001-108	BN_000925; NA[2]	NC_001735; R751p01-69	DQ401103; pRA3.01-3.50	AM901564; R7K_001-043	NC_005024; pSK41_p01-p46
/function=	/gene=; /product=	/note (comment)=	/*gene*= or /product=						
Replication	*rep;* Rep	replication initiator protein; or helicase; or primase; or regulator	RepA	RepB/FIB, RepE	*trfA*; TrfA1, TrfA2	TrfA1, TrfA2	RepB	RepA	Rep, Rep(AC)
	oriV	origin of vegetative replication		*ori-1/oriV, ori-2/oriS*	*oriV*				
	ssb; Ssb	single-stranded DNA binding protein	Ssb	Ssb	Ssb	Ssb		Ssb	TraM
Partitioning	*parA;* ParA	partitioning protein	ParA	SopA	IncC; IncC1, IncC2	IncC1, -C2	IncC		ParM
	parB; cbp; ParB	centromere binding protein	ParB	SopB	KorB	KorB	KorB		ParR
	parC; parS; cen	centromere		*sopC*					
Conjugative DNA transfer (Dtr)	*rlx;* Rlx	relaxase	TraI	TraI	TraI	TraI	Nic	TrwC	Nes
	nic	nick site, origin of conjugative replication		*oriT*	*oriT*	*nic*			*oriT*
	dtr; Dtr	relaxosome auxiliary proteins		TraY, -M	TraH, -J, -K	TraH, -J, -K		TrwA	
	pri; Pri	DNA primase			TraC	TraC	TraC3, -C4		
	cpl; Cpl	coupling protein	TraD	TraD	TraG	TraG	VirD4	TrwB	TraK
Exclusion	*sfx;* Sfx	surface exclusion protein		TraT					
	eex; Eex	entry exclusion protein		TraS	TrbK	TrbK	Eex		
Type IV secretion system (TivB)	*slt (virB1)*[3]	Soluble transglycosylase	pNDM-1_Dok01_N0219	GeneX	TrbN	TrbN			
	tivB2 (virB2)	P-type propilin	TraA		TrbC	TrbC	VirB2		
	tivB3 (virB3)	pilus assembly	TraL	TraL	TrbD	TrbD	VirB3		
	tivB4 (virB4)	T4SS ATPase	TraC	TraC	TrbE	TrbE	VirB4	TrwK	TraE
	tivB5 (virB5)	pilus assembly	TraE	TraE	TrbF	TrbF	VirB5	TrwJ	

	tivB6 (virB6)	T4SS protein			TrbL	TrbL	VirB6	TrwI	
	tivB7 (virB7)	T4SS protein	TraV	TraV	TrbH	TrbH	pRA3.23	TrwH	
	tivB8 (virB8)	T4SS protein			TrbJ	TrbJ	VirB8	TrwG	
	tivB9 (virB9)	T4SS protein	TraK	TraK	TrbG	TrbG	VirB9	TrwF	
	tivB10 (virB10)	T4SS protein	TraB	TraB	TrbI	TrbI	VirB10	TrwE	
	tivB11 (virB11)	T4SS protein			TrbB	TrbB	VirB11	TrwD	
	Pep; Pep	P-type propilin processing, cyclization	TrhF		TraF	TraF			
Mating pair formation proteins (Mpf)	*mpfPL-O*	P-type mating pair formation proteins			TrbL,-M,-N	TrbL, -M, -N		TrwL,-N	
F type IV secretion proteins (TivF)	*tivF1 (traA)*	F-type propilin		TraA					
	nac; Nac	F-type pilin acetylase		TraX	TrbP	TrbP			
	tivF2 (*traF*)	F-type T4SS protein	TraF	TraF					
	tivF3(*traG*)	F-type T4SS protein, Mating pair stabilization	TraG	TraG					
	tivF4(*traH*)	F-type T4SS protein	TraH	TraH					
	tivF5 (*trbI*)	F-type T4SS protein		TrbI					
	tivF6 (*traN*)	F-type T4SS protein, Mating pair stabilization	TraN	TraN					
	tivF7 (*traU*)	F-type T4SS protein	TraU	TraU					
	tivF8 (*traW*)	F-type T4SS protein	TraW	TraW					
	tivF9 (*trbC*)	F-type T4SS protein		TrbC					
	dsbC (*trbB*)	DsbC homolog		TrbB					

[1] Several transfer proteins (traA,-B, -C, -D, -F, -G, -H) are not listed because there is no detectable homology to other proteins listed in the Table. IncpSK1 is an incompatibility group in *Staphylococcus aureus.*

[2] Not available.

[3] The VirB1-11 and F-type T4SS homologues from the Ti and F plasmids respectively are given in brackets. The protein name is omitted.