

1 **Annotation of Plasmid Genes**

2 Christopher M. Thomas^{1*}, Nicholas R. Thomson², Ana M. Cerdeño-Tárraga^{2a}, Celeste J. Brown³, Eva
3 M. Top³ and Laura S. Frost⁴

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

9
10 ^aCurrent address: EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD,
11 UK.

12 *Author for correspondence

13 14 **Abstract**

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

28
29 **Key words:** genome; DNA sequence; gene names; gene function; replication; conjugative transfer

30 31 **Highlights:**

32 Review of current “issues” with plasmid annotation
33 Overview of annotation developments for other types of genome elements
34 Suggestions for good practice in plasmid annotation
35 Discussion of annotation in relation to replication, stability and conjugation functions
36 Previously unpublished annotation of IncX1 plasmid complete genome sequence as a model

37

38 **1. Introduction**

39 Efforts to standardize plasmid naming and gene annotation have not kept up with the deluge of
40 data provided by modern high throughput sequencing and automated annotation. In 1976, Novick
41 et al. published a schema for naming plasmids (pXY1,2,3, etc) and the genes they carry. The
42 convention for naming plasmids was generally followed for many years but gradually eroded as
43 new plasmids were reported with increasing frequency. Researchers moved away from this
44 simple naming system and instead used names that reflected the strain, the cloning procedure, the
45 institution's or investigator's initials, etc. This was further exacerbated by the discovery of
46 plasmids in genome and metagenome sequencing projects where no naming protocol exists. This
47 problem has been outlined within a broader examination of microbial elements by Klimke and his
48 associates at NCBI (2011). They have worked to standardize experimental data entry into
49 GenBank and other databases through a portal named COMBEX (Anton et al., 2013). They
50 have also been working with other interested parties on the nomenclature of viruses (Brister et al.,
51 2010), Insertion Sequences (Siguier et al., 2012) and genomes and metagenomes (Markowitz et
52 al., 2014a,b). Other groups have tried to impose order on plasmid names with varying levels of
53 success (Angiuoli et al., 2008; Martinez-Garcia et al., 2011; Seiler et al., 2014; Wang et al., 2009;
54 Zuo et al., 2007). However, no system has received as widespread approval as the original
55 proposal by Novick et al. (1976). Therefore, we encourage researchers to follow the naming
56 scheme from Novick et al. (1976) whereby a plasmid is designated with a small "p" at the
57 beginning of the name followed by a combination of letters and numbers that are unique to that
58 plasmid. Some recently named plasmids have the "p" at the end, which in older nomenclature
59 designated a protein, or in the middle, which lacks the clarity of the initial "p". However,
60 plasmids that were named prior to 1976 should keep their names (F, RP4, ColIb-P9etc) because
61 much confusion could arise if well studied plasmid paradigms were renamed at this point in time.
62 Also, where a convention has developed within a research community, such as those who work
63 on plasmids of *Rhizobium*, we encourage new researchers to use that system rather than develop
64 something new (Cevallos et al., 2008). In addition, it is worth making sure that the annotation
65 starts at a similar point and progresses in the same direction round the plasmid as annotations of
66 related plasmids already in the databases unless what has gone before is deemed unsatisfactory
67 with good reasons.

68

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

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

86

87 **2. A brief history of plasmid annotation**

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

90

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

96

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

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

112

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

120

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

128

129 **3. Issues and possible solutions**

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

136

137 **3.1 General issues**

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

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

160

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

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

171

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

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

185

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

198

199 Similarly, many plasmid gene products are too well known to change their names at this late date.
200 For instance, new IncP plasmids should maintain the IncP-specific gene names such as *trf* (*trans-*
201 *acting* replication function), *kor* (*kil* over-ride for genes that turned out to encode DNA binding
202 proteins that repress transcription) and *kla/klc/kle* (Kil locus A, C and E) that are in common
203 usage (see BN000925 and U67194). It also makes sense that new plasmids that are closely related
204 to these paradigms should be annotated using the same set of gene names, for instance *tra* and *trb*
205 for the transfer genes of IncP plasmids. On the other hand, it might not be appropriate to use
206 KorA as the name for a close homolog of KorA unless it is known to regulate a *kil* gene (a gene
207 that is unclonable due to a bacteriostatic or bacteriocidal effect if unregulated). Thus we
208 recommend that gene names reflect gene function as much as possible and we definitely
209 recommend against plasmids being annotated using a mixture of names based on the top BLAST
210 hits.

211

212 Below is a discussion of these main concepts.

213

214 **3.2 Replication functions**

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

234

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

246

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

259

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

268

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

278

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

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

286

287 **3.3 Partitioning functions**

288 Partitioning refers to the distribution of newly replicated plasmids into daughter cells after cell
289 division. In general, it is a feature of large, low copy number plasmids that cannot rely on random
290 distribution through a “safety in numbers” mechanism. Three main types of partitioning systems
291 have been described in plasmids: I, II and III with I subdivided into Ia and Ib. In addition to an
292 NTPase, there is a centromere sequence and a centromere-binding protein CBP (Schumacher,
293 2012) with the NTPase and CBP defining the groups I, Par A,-B; II, ParR,-M; and III, TubZ,-R.
294 The most difficult partitioning proteins to predict are the type Ib CBPs that vary in structure
295 considerably – the putative *cbp* gene of R6K being an example (Supplementary Data Table S1,
296 CDS R6K0033). In general, the NTPases of group Ia and the CBP of Ib, II, and III autoregulate
297 *par* expression. Thus, DNA-binding proteins originally identified as repressors were later shown
298 to be CBPs involved in partitioning. An example of this is KorB from the IncP plasmids, which is
299 a Ia CBP. Unfortunately, CBPs in annotated sequences are often described as repressors and their
300 role in partitioning is overlooked. Again, this requires that the context of the gene within a region
301 be examined manually since computer algorithms are currently unable to connect position to
302 function. For instance, since plasmid partitioning regions contain three characteristic sequences,
303 if one is identified, the other two should be nearby.

304

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

312

313 **3.4 Conjugation functions**

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

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

324

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

330

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

343

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

351 Examples include the IncA/C plasmid pNDM-1_Dok01 (Table 1) and the IncHI1 plasmid R27

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

354

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

375

376 Table 1 illustrates various attempts to come to terms with naming T4SS genes and their products.
377 The IncA/C plasmid pNDM-1_Dok01 has a circular P-type pilin subunit named TraA, which is
378 also the name for the historically important linear F-type pilin subunit. It is processed by the
379 peptidase/cyclase TrhF, a name derived from TrhF from the IncHI1 plasmid R27 involved in the
380 maturation of the circular TrhA pilin. The name TrhF is, in turn, derived from the TraF
381 peptidase/cyclase of IncP plasmids (Table 1) which was first referred to as a peptidase in the *traF*
382 /function= " peptidase / maturation of TrbC pilin protein" of IncP plasmid pKJK5(AM261282). The
383 T4SS gene products in the IncU plasmid pRA3 are named after their closest homologues, VirB2-
384 11, which suggests these proteins having a role in virulence (Table 1). These names could easily
385 be changed to the TivB1-11 nomenclature. A further refinement would be to designate whether
386 the propilin is F- or P-like by using TivF1 and TivB2 (since the Ti plasmid VirB system is P-like)

387 respectively and TivF2,-F3, etc for the other essential gene products in F-type T4SS (Table 1,
388 column 2). Núñez (1998) foresaw the problems in T4SS nomenclature and suggested PilX1-11
389 for the T4SS of the IncX plasmid R6K. However, to avoid confusion we suggest that the TivB1-
390 11 nomenclature be adopted, as illustrated in Supplementary Data Table S1. With the realization
391 that Gram-positive and archaeal conjugative systems also use a modified T4SS, albeit with no
392 visible pili, and in cases where no incompatibility group is known, we suggest using TivB1-11 for
393 the appropriate homologues as the default nomenclature (see Supplementary Data Table S1).

394

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

405

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

413

414 Other “transfer” genes and proteins actually reduce transfer efficiency. These include proteins
415 that block mating pair formation (**S**urface **e**xclusion or Sfx), block DNA entry (**E**ntry **e**xclusion or
416 Eex) and reduce transfer gene expression (**F**ertility **i**nhibition or Fin). We encourage investigators
417 to not refer to these genes as *tra* genes.

418

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

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

424

425 **5. Conclusions**

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

433

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

441

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

451

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

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

458

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

464

465 **6. Acknowledgements**

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

471

472 **7. References.**

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

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

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

490 (2013) The COMBREX project: design, methodology, and initial results. PLoS Biol. 11:
491 e1001638. doi: 10.1371/journal.pbio.1001638.

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

496 Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller
497 Aarestrup F, H. (2014) *In Silico* detection and typing of plasmids using PlasmidFinder
498 and Plasmid Multilocus Sequence Typing. Antimicrob Agents Chemother. 58: 3895–
499 3903.

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

502 Christie PJ, Gordon JE (2014). The Agrobacterium Ti Plasmids. Microbiol Spectr. 2. doi:
503 10.1128/microbiolspec.PLAS-0010-2013.

504 Costa TR, Ilangovan A, Ukleja M, Redzej A, Santini JM, Smith TK, Egelman EH, Waksman G.
505 (2016) Structure of the Bacterial Sex F Pilus Reveals an Assembly of a Stoichiometric
506 Protein-Phospholipid Complex. Cell 166: 1436-1444. e10. doi:
507 10.1016/j.cell.2016.08.025.

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

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

514 Goessweiner-Mohr N, Arends K, Keller W and Grohmann E (2013) Conjugative type IV
515 secretion systems in Gram-positive bacteria. Plasmid 70: 289–302. doi:
516 10.1016/j.plasmid.2013.09.005

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

519 Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, Rocha EP (2014) Key
520 components of the eight classes of type IV secretion systems involved in bacterial
521 conjugation or protein secretion. Nucleic Acids Res. 42: 5715-27. doi:
522 10.1093/nar/gku194.

523 Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrahi 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, Tennesen 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, Cabezón 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 Núñez B. (1998) Ph.D. Thesis. Departamento de Biología Molecular, Universidad de Cantabria,
551 Facultad de Medicina, Cardena Herrera Oria sn, Santander, 39011, SPAIN

552 Núñez 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, Ciufu 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,

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

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

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

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

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

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

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

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

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

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

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

590 Wang Z, Jin L, Yuan Z, Wegrzyn G, Wegrzyn A (2009) Classification of plasmid vectors using
591 replication origin, selection marker and promoter as criteria. *Plasmid.* 61: 47-51. doi:
592 10.1016/j.plasmid.2008.09.003.

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

596 Websites:

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

598 International Nucleotide Sequence Database Collaboration (INSDC):
599 <http://www.insdc.org/>

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

601 NCBI Reference Sequence Database:
602 <http://www.ncbi.nlm.nih.gov/refseq/>
603 <https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>

604 Swiss-Prot Group:
605 <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 ₁ , 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	<i>rep</i> ; Rep	replication initiator protein; or helicase; or primase; or regulator	RepA	RepB/FIB, RepE	<i>trfA</i> ; TrfA1, TrfA2	TrfA1, TrfA2	RepB	RepA	Rep, Rep(AC)
	<i>oriV</i>	origin of vegetative replication		<i>ori-1/oriV</i> , <i>ori-2/oriS</i>	<i>oriV</i>				
	<i>ssb</i> ; Ssb	single-stranded DNA binding protein	Ssb	Ssb	Ssb	Ssb		Ssb	TraM
Partitioning	<i>parA</i> ; ParA	partitioning protein	ParA	SopA	IncC; IncC1, IncC2	IncC1, -C2	IncC		ParM
	<i>parB</i> ; <i>cbp</i> ; ParB	centromere binding protein	ParB	SopB	KorB	KorB	KorB		ParR
	<i>parC</i> ; <i>parS</i> ; <i>cen</i>	centromere		<i>sopC</i>					
Conjugative DNA transfer (Dtr)	<i>rlx</i> ; Rlx	relaxase	Tral	Tral	Tral	Tral	Nic	TrwC	Nes
	<i>nic</i>	nick site, origin of conjugative replication		<i>oriT</i>	<i>oriT</i>	<i>nic</i>			<i>oriT</i>
	<i>dtr</i> ; Dtr	relaxosome auxiliary proteins		TraY, -M	TraH, -J, -K	TraH, -J, -K		TrwA	
	<i>pri</i> ; Pri	DNA primase			TraC	TraC	TraC3, -C4		
	<i>cpl</i> ; Cpl	coupling protein	TraD	TraD	TraG	TraG	VirD4	TrwB	TraK
Exclusion	<i>sfx</i> ; Sfx	surface exclusion protein		TraT					
	<i>eex</i> ; Eex	entry exclusion protein		TraS	TrbK	TrbK	Eex		
Type IV secretion system (TivB)	<i>slt</i> (<i>virB1</i>) ³	Soluble transglycosylase	pNDM-1_Dok01_N0219	GeneX	TrbN	TrbN			
	<i>tivB2</i> (<i>virB2</i>)	P-type propilin	TraA		TrbC	TrbC	VirB2		
	<i>tivB3</i> (<i>virB3</i>)	pilus assembly	TraL	TraL	TrbD	TrbD	VirB3		
	<i>tivB4</i> (<i>virB4</i>)	T4SS ATPase	TraC	TraC	TrbE	TrbE	VirB4	TrwK	TraE
	<i>tivB5</i> (<i>virB5</i>)	pilus assembly	TraE	TraE	TrbF	TrbF	VirB5	TrwJ	

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