MOBILE GENETIC ELEMENTS: THE AGENTS OF OPEN SOURCE EVOLUTION

Laura S. Frost, Raphael Leplae‡, Anne O. Summers§ and Ariane Toussaint‡*

Abstract | Horizontal genomics is a new field in prokaryotic biology that is focused on the analysis of DNA sequences in prokaryotic chromosomes that seem to have originated from other prokaryotes or eukaryotes. However, it is equally important to understand the agents that effect DNA movement: plasmids, bacteriophages and transposons. Although these agents occur in all prokaryotes, comprehensive genomics of the prokaryotic mobile gene pool or 'mobilome' lags behind other genomics initiatives owing to challenges that are distinct from cellular chromosomal analysis. Recent work shows promise of improved mobile genetic element (MGE) genomics and consequent opportunities to take advantage — and avoid the dangers — of these 'natural genetic engineers'. This review describes MGEs, their properties that are important in horizontal gene transfer, and current opportunities to advance MGE genomics.

*Department of Biological Sciences, Biological Sciences Centre, University of Alberta Edmonton, Alberta T6G 2E9, Canada.
‡Service de Conformation de Macromolécules Biologiques et de Bioinformatique, Université Libre de Bruxelles, Bruxelles, Belgium.
§Department of Microbiology, Biological Sciences Building, University of Georgia, Athens, Georgia 30602-2605, USA.
Correspondence to A.O.S.
e-mail: summers@uga.edu
doi:10.1038/nrmicro1235

Mobile genetic elements (MGEs) are segments of DNA that encode enzymes and other proteins that mediate the movement of DNA within genomes (intracellular mobility) or between bacterial cells (intercellular mobility). Intercellular movement of DNA takes three forms in prokaryotes: TRANSFORMATION, CONJUGATION and TRANSDUCTION (see also the article by C. M. Thomas and K. M. Nielsen in this issue). Transformation was the first mechanism of prokaryotic horizontal gene transfer (HGT) to be discovered. This process involves the transfer of cellular DNA between closely related bacteria and is mediated by chromosomally encoded proteins that are found in some naturally transformable bacteria. By contrast, as shown in FIG. 1, conjugation requires independently replicating genetic elements called conjugative plasmids, or chromosomally INTEGRATED CONJUGATIVE ELEMENTS (ICEs), which include conjugative transposons (CTns)[1,2]. These genetic elements encode proteins that facilitate their own transfer and occasionally the transfer of other cellular DNA from the 'donor' plasmid-carrying cell to a recipient cell that lacks the plasmid or ICE (see also the article by C. M. Thomas and K. M. Nielsen in this issue). Transduction is also a form of DNA transfer that is mediated by independently replicating bacterial viruses called bacteriophages (or phages). At low frequency, bacteriophages can accidentally package segments of host DNA in their capsid and can inject this DNA into a new host, in which it can recombine with the cellular chromosome and be inherited. Intracellular movement of DNA is a property of promiscuously recombining loci that are generically called transposons, which randomly recombine or 'jump' between replicons. As transposons can 'hop' into phages or plasmids, they can also be transferred with them into other cells.

Traces of MGE activity are evident in all prokaryotic genome sequences (see the article by J. P. Gogarten and J. P. Townsend in this issue). MGE transposases and site-specific recombinases catalyse the intracellular movement of MGEs and, with the HOMOLOGOUS RECOMBINATION systems of the host, they enable chromosomal deletions and other rearrangements[3]. Recent efforts to understand the origins and roles of immigrant

chromosomal genes, and the recognition that MGEs have important roles in infectious diseases, antibiotic resistance, bacterial symbioses, and biotransformation of xenobiotics, has kindled interest in the comprehensive genomic analysis of the MGEs. Although these genetic elements are potent agents of change, their contributions to the mode and tempo of bacterial evolution have just begun to be examined[4].

The following sections briefly review the most important genes that define these elements as agents of HGT, selected accessory MGE genes that are involved in medically, agriculturally and environmentally important processes, and the unique challenges of MGE genomics.

Plasmids and other conjugative elements

A plasmid is a collection of functional genetic modules that are organized into a stable, self-replicating entity or 'replicon', which is smaller than the cellular chromosome and which usually does not contain genes required for essential cellular functions. The classic plasmids are covalently closed, circular double-stranded DNA molecules (FIG. 1), but linear double-stranded DNA plasmids have been found in an increasing number of species[5–7]. The general anatomy of a plasmid includes the essential 'backbone' of genes that encode replicative functions and a variable assortment of accessory genes that encode processes that are distinct from those encoded by the bacterial chromosome (see below). Such accessory traits can be accumulated in the cell without altering the gene content of the bacterial chromosome[8,9].

Plasmids must replicate, control their copy number, and ensure their inheritance at each cell-division by a process known as partitioning. It is impossible for plasmids with the same replication mechanism to co-exist in the same cell, a phenomenon termed 'incompatibility' (Inc). The Inc trait provided the basis for the initial classification of some plasmids that is still in use today. Incompatibility groups have been defined for plasmids of the enterobacteriaceae (26 groups), the

TRANSFORMATION
Gene transfer that is mediated by the uptake of free DNA.

CONJUGATION
Gene transfer that is mediated by certain plasmids or ICEs with relevant transfer genes. Cell–cell contact is required for conjugation, unlike transduction or transformation.

TRANSDUCTION
Gene transfer that is mediated by certain types of bacteriophage.

INTEGRATIVE CONJUGATIVE ELEMENTS
(ICEs). Together with conjugative transposons (CTns) and genomic islands, these are chromosomally located gene clusters that encode phage-linked integrases and conjugation proteins as well as other genes associated with an observable phenotype such as virulence or symbiosis. ICEs and CTns are gene clusters that can be transferred between cells, whereas genomic islands have not been shown to transfer. Although these gene clusters have some phage-like genes, they do not lyse the cell or form extracellular particles.

HOMOLOGOUS RECOMBINATION
DNA recombination that requires extensive sequence similarity in the involved DNA segments. It is usually effected by chromosomally encoded genes, but some phages also have orthologues of such chromosomal genes.

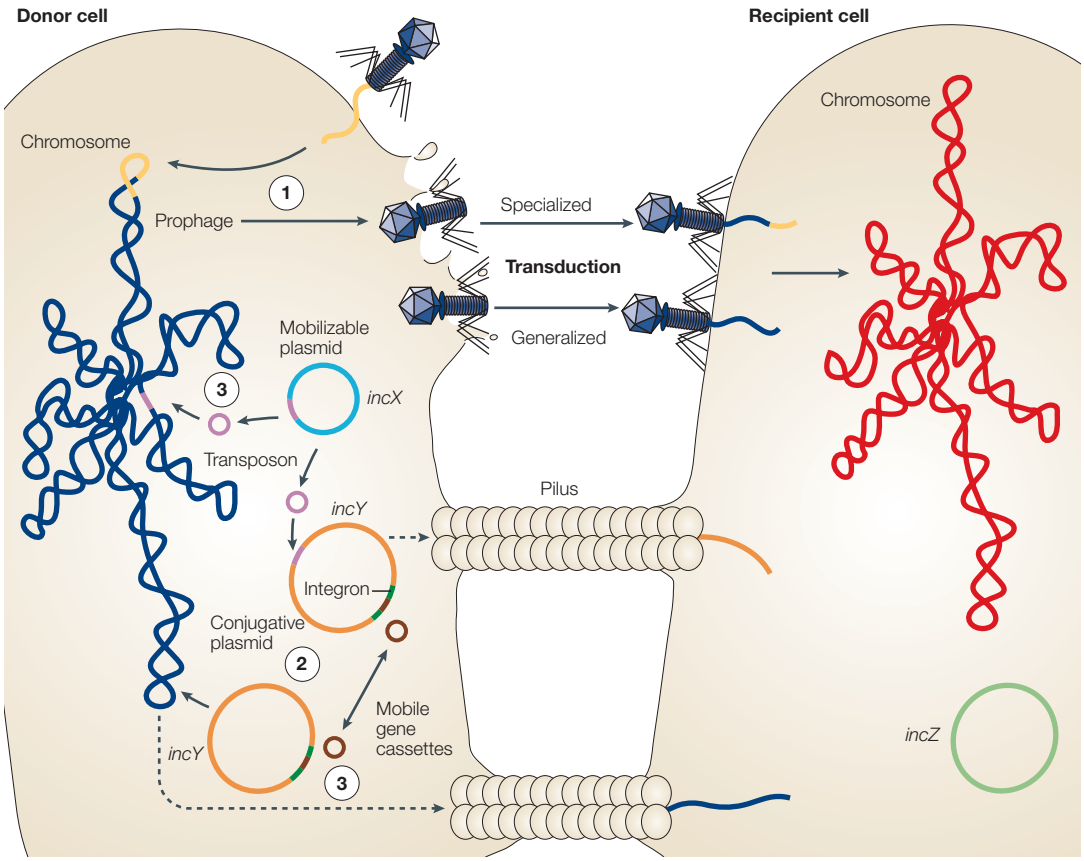

Figure 1 | **Transfer of DNA between bacterial cells.** Transduction (1). The DNA genome (yellow) of a temperate phage inserts into the chromosome (dark blue) as a prophage; it later replicates, occasionally packaging host DNA alone (generalized transduction) or with its own DNA (specialized transduction), lyses the cell, and infects a naive recipient cell in which the novel DNA recombines into the recipient host cell chromosome (red). Conjugation (2). Large, low copy number conjugative plasmids (orange) and integrated conjugative elements (ICEs; not shown) use a protein structure (known as a pilus) to establish a connection with the recipient cell and to transfer themselves to the recipient cell. Alternatively, a copy of a small, multicopy plasmid or defective genomic island or a copy of the entire bacterial chromosome can be transferred to a naive cell, in which these genetic elements either insert into the chromosome or replicate independently if compatible with the resident plasmids (light green). Conjugative transposons and plasmids of Gram-positive bacteria (not shown) do not use pili. Transposition (3). Transposons (pink) integrate into new sites on the chromosome or plasmids by non-homologous recombination. Integrons (dark green) use similar mechanisms to exchange single gene cassettes (brown). Details of these and other MGEs can be found in REFS 119,120.

pseudomonads (14 groups), and for the Gram-positive staphylococci (~18 groups). The number of incompatibility groups of large plasmids in these bacteria seems to be reaching a plateau, so there might be a finite number of successful replication mechanisms in a given bacterial group[10] and, therefore, replicon genes might continue to be useful in classifying plasmids in these bacteria (see the Database of Plasmid Replicons in Online Links box). Plasmids of other bacteria and the Archaea have not been classified because of the difficulty of measuring plasmid competition[11] and of designing appropriate molecular probes in the absence of adequate sequence data[12]. Plasmid complexity increases with size and the so-called 'megaplasmids' can be the size of small chromosomes and can contain several co-integrated compatible replicons. Natural bacterial isolates often contain small, cryptic plasmids that comprise only replication genes and a few genes of unknown function. Such small plasmids can often be transferred to another cell by a larger conjugative plasmid or ICE, a process known as MOBILIZATION[13].

Unlike DNA transfer by transduction or transformation, which occurs as side effects of phage propagation or of nutrient uptake[14], respectively, conjugation has evolved to move the plasmid itself efficiently into other cells. The conjugative or transfer (*tra*) genes establish a stable mating pair and trigger DNA transport from the donor to the recipient cell through a specialized transfer pore. Other genes ensure the survival of DNA in the possibly hostile environment of the new host. The main steps in conjugation are: first, mating-pair formation (Mpf); second, a signalling event that transfer can occur; and third, the transfer of DNA (Dtr). Preparation of the DNA prior to transfer is similar among conjugative systems but the mechanism of mating-pair formation varies considerably.

The hallmark transfer gene encodes the 'coupling protein', which synchronizes mating-pair formation with DNA transfer and is thought to 'pump' the DNA into the recipient cell[15]. Coupling proteins have different names in different systems, but all belong to the TraG-like family of ATPases (named for the RP4 plasmid orthologue)[16] and they associate with the cytoplasmic membrane. They are a subgroup of the larger FtsK/SpoIIIE family of ATPases that are involved in double-stranded DNA transfer during cell division, at forespore formation, and during conjugation by high-GC Gram-positive bacteria and possibly by the Archaea[15,17–23]. A crystal structure has been determined for the TraG orthologue of the IncW plasmid R388, known as TrwB[24].

Most conjugative systems include a relaxase that nicks DNA to give a single-stranded substrate that is suitable for transfer. This nicking occurs in a strand- and site-specific manner at a *nic* site, allowing self-transmissible and mobilizable systems to be classified by their relaxase and *nic* sequences[25]. Early studies indicated the genetic linkage of certain conjugation systems with specific Inc groups[10] but whether this is indeed the case remains an open question. There are several types of conjugative mechanisms, with the greater distinctions being between those of Gram-positive and Gram-negative bacteria. Given their similarities, there might be a limited number of significantly different systems, although the paucity of sequence data makes that idea tentative.

The hair-like surface appendage, known as a pilus, is another hallmark of conjugative systems in Gram-negative bacteria. The anatomy of conjugative pili resembles that of filamentous phages, which underscores the relationship between MGEs[17]. Pilus assembly is a function of a type IV secretion system (T4SS; see REFS 22,26) (FIG. 2), in which a coupling protein links a transenvelope protein complex (a transferosome) to a nucleoprotein complex (a relaxosome), which is bound at the plasmid's origin of transfer (*oriT*). Conjugative elements in Gram-negative bacteria are easily identified by their highly conserved T4SS signature proteins[27] — the relaxase and the coupling protein[25] (FIG. 2).

T4SSs are found in all known conjugative plasmids of Gram-negative bacteria, except those of *Bacteroides* species, and in many ICEs and genomic islands[28,29]. T4SS-like type II secretion systems (T2SSs)[30] and type III secretion systems (T3SSs)[31] encode an ATPase (VirB11 in the Ti plasmid of *Agrobacterium tumefaciens*) that belongs to the TadA subfamily[32], the structure of which has been determined[33]. Members of this ATPase subfamily are associated with the assembly of extracellular filaments (such as pili[30,34], filamentous phage[35] and flagella[36]) and with the transport of DNA either into (transformation)[23,37] or out of (conjugation, DNA extrusion) the cell [38,39]. The T4SSs of plasmids can be divided into P and F subgroups on the basis of two criteria[28]. VirB11-like proteins are signature proteins of P-T4SSs, whereas F-T4SSs encode proteins with many conserved cysteines (22 in the F-TraN protein) and DsbC-like proteins with predicted thioredoxin folds. Additional T4SS proteins that are characteristic of conjugative elements include a TonB-like homologue (VirB10 in the Ti plasmid)[40] and a second ATPase (VirB4 in Ti), which is involved in pilus assembly[41]. The cyclic pilin, together with the accompanying cyclase, which has been characterized for the RP4 and the Ti plasmid, are diagnostic of P-T4SSs[42,43]. Other T4SS-associated proteins show less sequence conservation and are not as useful for detecting conjugative elements.

Low-GC Gram-positive bacteria seem to carry a limited repertoire of conjugative mechanisms[44], which include the pheromone-dependent mating systems of the enterococci and pheromone-independent systems, all of which involve a cell surface protein that initiates mating-pair formation. *Streptomyces* species, *Mycobacterium* species, and possibly the Archaea, are unusual in that they use only one essential protein that is homologous to TraG coupling proteins to transport double-stranded DNA[18–21]. Conjugative transposons, which are a form of ICEs that were first described in Gram-positive bacteria, contain characteristic phage-like integrases[2,45]. As archaeal plasmids can encode integrases, they might also form ICEs[20,46].

MOBILIZATION
Transfer by a conjugative element of a plasmid or part of the bacterial cellular chromosome that cannot effect self transfer. Mediated by the *trans*-acting proteins of the conjugative plasmid that function on cognate mobilization (*oriT*) sites in the mobilized plasmid to direct it to the conjugation pore built by the conjugative element.

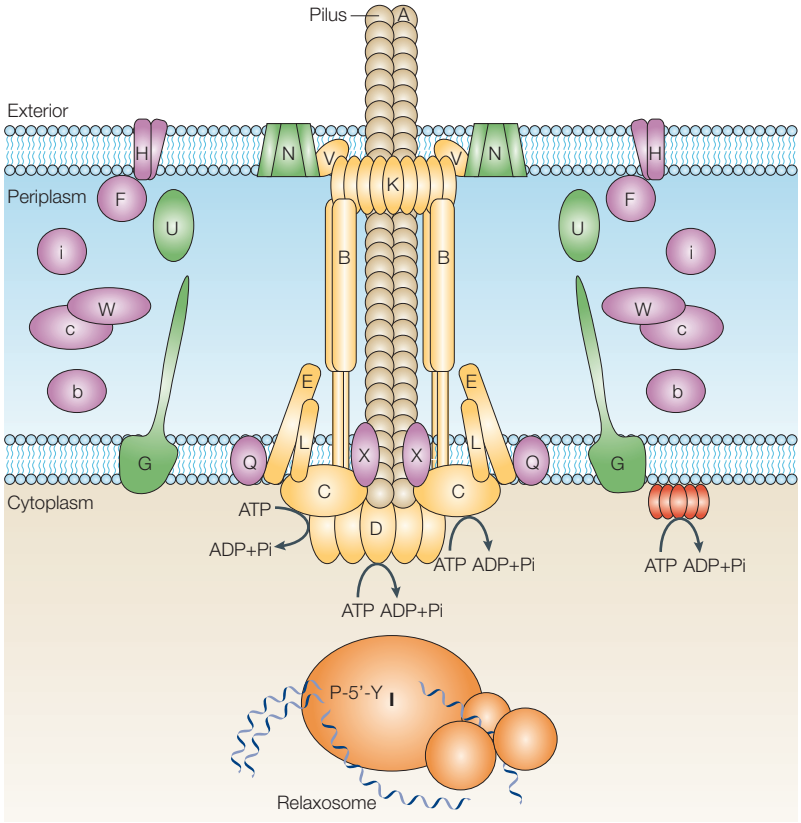

Figure 2 | **Signature proteins of conjugative systems in Gram-negative bacteria.** The F-type (F) conjugative system is shown located in the inner and outer membranes and extending through the periplasm (for further details, see REF. 28). Tra proteins are labelled with capital letters and Trb proteins with lower-case letters. Shared orthologues found in the Ti plasmid are yellow (for further details see REF. 26). Proteins that are specific for F-type systems are shown in pink for pilus assembly or in green for mating-pair stabilization. The ATPase that is characteristic of P-type systems (VirB11 in Ti) is shown in red. The pilus, which is composed of repeating subunits of TraA, is erected by plasmid-specific type IV secretion systems (T4SSs). The chaperone TraQ inserts pilin into the inner membrane where it is acetylated by TraX. Signature F-type proteins include TraF and TrbB, which have thioredoxin-like folds (T. Elton *et al.*, unpublished observations). All known conjugative systems have a 'coupling protein' (called TraD in the F plasmid), which is an inner-membrane protein with a cytoplasmic domain that links the T4SS to the relaxosome bound to the DNA during transfer. Relaxosomes have a relaxase (TraI) that nicks the DNA at *nic* within the origin of transfer (*oriT*); other mechanisms might operate in Gram-positive bacteria (see the main text for further discussion). Upon nicking, a conserved tyrosine (Y) of the relaxase forms a covalent bond with the 5' phosphate (P) of the DNA strand (P–5'–Y).

One significance of plasmid transfer in HGT lies in the fact that many plasmids and ICEs can also effect the transfer of chromosomal DNA[47] (FIG. 1), as exemplified by the high frequency of recombination (Hfr) mode of F plasmids and by the chromosome mobilization ability (Cma) of plasmids in *Streptomyces* species. Such conjugative elements integrate into the host genome and transfer large sections of the chromosome, along with parts of the conjugative element, into a recipient cell. Because of the difficulty in assessing whether host chromosome mobilization is possible for a given conjugative element, very few systems have been assayed for this property, although it has probably been a potent agent for chromosome construction. Efficient chromosome mobilization occurs in strains of *Halorubrum*[48] and Hfr-like transfer of the genes for anaerobic growth is seen in extreme thermophiles[49]. Hfrs can excise imprecisely from the chromosome, generating F' episomes, which are independently replicating plasmids that carry large adjacent portions of the chromosome and that confer stable partial diploidy for chromosomal loci[47], allowing the evolution of altered functions. Chromosome mobilization is not a well-defined trait, but seems to depend on the replicon and the presence on both the plasmid and the chromosome of transposons or other mobile elements that serve as sites of portable homology for recombination.

Bacteriophages and transposons

Bacteriophages were the organisms first exploited for use in molecular biology and genomics[50,51]. Phages are the most abundant (~10^{30} tailed phage particles) and the most rapidly replicating (10^{25} infections every second) life forms on earth and their genetic diversity is enormous[52–54]. Long used in genetic engineering, they have gained new attention for their potential for use in antibacterial therapy[55] and in nanotechnology[56].

The genomes of phages can be composed of either single- or double-stranded DNA or RNA and can range in size from a few to several 100 kb. Their characteristic essential genes comprise specific replicase genes, genes encoding phage components that 'hijack' the host cell replicative machinery, and genes encoding the proteins that package DNA in a protein coat (capsid). Virulent bacteriophages replicate vigorously and lyse the host bacteria. Temperate bacteriophages have an alternative, quiescent, non-lytic growth mode called lysogeny[57]. In most known cases of lysogeny, the phage genome integrates into the bacterial chromosome and replicates with it as a prophage, but in a few cases, the phage genome replicates autonomously as a circular or linear plasmid. Lysogenic conversion, which is the provision of a new phenotype such as toxin production as a result of prophage carriage, was discovered over 50 years ago[58]. Recombination with other prophages and other mobile elements that reside in the same bacterial host contributes to the well-documented mosaic structure of phages[59].

Environmental stimuli, such as DNA damaging agents, provoke a switch from quiescent to virulent replication leading to cell lysis during which host cell DNA can be accidentally packaged and later injected into a new host in a process called transduction[60] (FIG. 1). The ability to transduce host DNA seems to be limited to relatively large (50–100 kb) double-stranded DNA phages. The transduced chromosomal DNA must be able to recombine with the genome of the recipient host to survive. Therefore, similar to transformation, HGT that is mediated by transduction is limited to members of the same bacterial species.

It is now clear that intracellular DNA movement mediated by various transposons and insertion sequences is effected by enzymes that are similar to those first described for the insertion of viral genomes into chromosomes[61] (TABLE 1). Temperate phages and genomic islands go in and out of their host chromosome

Table 1 | **Nucleophiles that attack DNA and their biological functions**

Tyrosine nucleophile	Serine nucleophile	H_2O nucleophile (DDE catalytic motif)
Phage and genomic island integration and excision	Phage and genomic island integration and excision	Transposition, transposable phage integration, retroviral cDNA integration
Plasmid dimer and co-integrate resolution	Plasmid dimer and co-integrate resolution	Holliday junction resolution
DNA inversion (phase variation)	DNA inversion (phase variation)	DNA inversion* (phase variation)
Phage and plasmid rolling circle replication	Not yet observed	Replication of transposable phages
Gene expression by excision	Not yet observed	Antibody expression by VDJ recombination

*Enzymes in this family belong to the IS*110* family of transposases for which a motif related to the DDE catalytic triad has been suggested.

using site-specific tyrosine and serine recombinases. Gene cassettes go in and out of INTEGRONS using tyrosine recombinases as well. Together with relaxases, which cleave the DNA that is transferred by conjugation, site-specific recombinases generate covalent intermediates with their target DNA (for example, 5′-phosphotyrosine for tyrosine recombinases, 3′-phosphotyrosine for relaxases and 3′-phosphoserine for serine recombinases). Tyrosine recombinases account for approximately 90% of these enzymes in bacteriophages and neither class of recombinases has been found in ICEs (A.T., unpublished observations). Transposition of short transposons called insertion sequences (ISs) and other transposons is most often catalyzed by DDE transposases, which specifically recognize and introduce nicks at the ends of these elements in the first step of the transposition reaction (see REF. 61 for further mechanistic details).

MGE accessory genes

Homologous and NON-HOMOLOGOUS RECOMBINATION events (insertional events) occur between MGEs and cellular chromosomes. Therefore, the distinction between chromosome-derived and transposon-derived loci on MGEs is not always clear. However, many promiscuously recombining elements encode genes that are not commonly found among chromosomal housekeeping genes. These so-called accessory MGE genes confer clinically or economically important properties on the host cell. It is the properties conferred by these distinct loci that first drew attention to the MGEs. Some of the important roles of MGE accessory loci are described below. Although the relative amount of DNA involved is small compared to the entire chromosome of the bacterial cell, the effects on the behaviour of the host cell are dramatic. With respect to taking advantage of an unusual biochemical niche or becoming a 'hot' pathogenic strain, MGEs are 'where the action is'.

Infectious diseases. The earliest described accessory function of plasmids was antibiotic multi-resistance[62]. Now recognized as an inevitable result of the widespread use of antibiotics, this phenomenon threatens to return certain areas of medical practice (notably critical and end-of-life care) to a pre-antibiotic era in which there are no drugs to treat infected people[63]. Plasmid-borne resistance genes originate as point mutations in the target genes of susceptible bacteria and also from genes that provide antibiotic-producing bacteria with protective mechanisms. These genes can be rendered mobile when they are flanked by ISs, when they are picked up by transposons of the Tn*3* family, or as mobile cassettes by integrons[64,65] (which themselves can be part of such transposons). These configurations allow large arrays of resistance genes for most classes of antibiotics and disinfectants to be transferred together in a single conjugation event[66]. Integrons also provide a promoter to express the genes or gene fragments that they capture. In addition to drug resistance, conjugative plasmids frequently carry genes that encode toxins[67] and other virulence factors, as well as genes for cellular processes and structures required for colonization of animal[68] and plant hosts[69]. These genes can be moved around by phages and conjugative transposons. Their contribution to bacterial evolution is being revealed by bacterial-genome sequencing[53], which shows that plasmids and phages are important instruments in the divergence of closely related bacterial strains and species[52,70], especially in the emergence of new pathogens[71,72]. Gene transfer can spread rare, spontaneous resistance mutants through a bacterial population leading to potential pandemic problems. The crucial difference between a harmless commensal or soil bacterium and a deadly pathogen can be simply the presence of a plasmid (for example, in the case of anthrax[73,74]) or a bacteriophage (for example, cholera and diphtheria[67,70]).

Symbiosis and unusual metabolic traits. One of the most ecologically and agriculturally important elemental transformations on the planet — symbiotic nitrogen fixation — is mediated by plasmid-encoded genes of the genus *Rhizobium*[75]. Very large (>250 kb) conjugative plasmids in this genus carry genes for the invasion and the conversion of host-plant root cells into factories that convert atmospheric dinitrogen to ammonia, which meets the nitrogen needs of the plant. In return, the plant provides *Rhizobium* with photosynthetically generated carbohydrates. This process, which is not limited to agriculturally important plants, supplies a substantial fraction of the nitrogen requirements of all plant material on earth. In some

INTEGRON
A genetic element that encodes an integrase enzyme, which can assemble tandem arrays of genes or gene fragments and provide them with a promoter for expression. Often associated with antibiotic multi-resistance.

NON-HOMOLOGOUS RECOMBINATION
DNA recombination that requires little or no similarity between the DNA segments involved. This process is carried out by specialized enzymes that are encoded by transposons and phages.

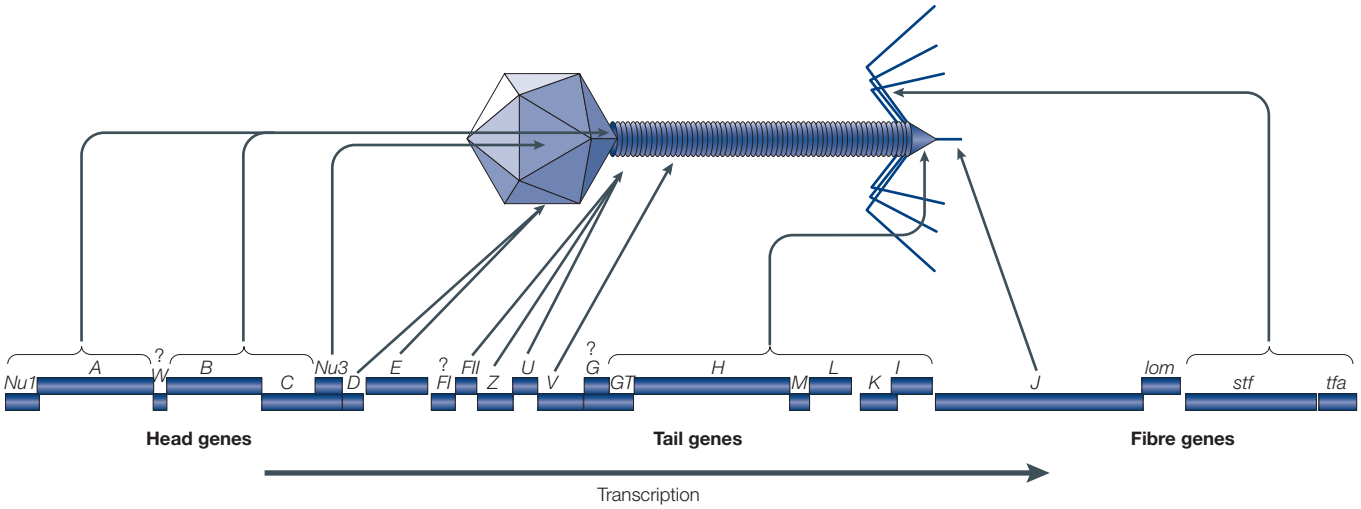

Figure 3 | **Genes and components of tailed phages.** The main components of the lambdoid phage family and the conserved order of the genes for these components is shown. Heads (or capsids), tails and fibres are composed of several proteins, some of which interact tightly with a defined chronology during morphogenesis. This is probably why the order of some genes is conserved although their sequences display no detectable similarity. Global analysis of phage genome sequences should allow the identification of the building blocks and the rules that govern their combination into functional phages. Component genes and their protein products include: *Nu1*, terminase small subunit; *A*, terminase large subunit; *W*, head stabilization; *B*, portal protein; *C*, capsid component; *Nu3*, scaffold protein; *D*, head-DNA stabilization; *E*, major head protein; *FI*, DNA maturation; *FII*, head stabilization; *Z–H*, tail components; *M*, *L*, *K* and *I*, tail assembly proteins; *J*, tail tip, host specificity; *lom*, outer membrane protein; *stf* and *tfa*, tail fibre proteins. Modified with permission from REF. 88 © 1992 Academic Press.

genera, the genes for symbiosis and nitrogen fixation are encoded on large (>500 kb) chromosomal 'symbiosis islands', which also include genes encoding a T4SS, a relaxase and a coupling protein for transfer by conjugation[76]. MGEs have recently been found to carry genes that participate in another important global elemental cycle, that of carbon. Phages of photosynthetic, marine blue-green bacteria (cyanophages) carry genes involved in photosynthesis as well as other genes that might help their host bacteria to survive the relatively nutrient-poor conditions of the ocean[77].

Bioremediation. Naturally occurring plasmid-encoded genes for the biotransformation of hydrocarbons were central to the very first patent awarded for a living organism[78,79]. Fears over the release of bacteria that carry genetically engineered plasmids delayed exploitation of similar natural microorganisms that are readily found in soil or water contaminated with hydrocarbons and other toxic compounds, including heavy metals. Nevertheless, substantial advances have been made in our understanding of the genetic and biochemical basis of plasmid-mediated remediation processes in polluted environments[80]. Both *in vitro* and *in vivo* techniques have been used to enrich for consortia of strains that possess combinations of organic and metal bioremediation activities for use in contained, pump-and-treat industrial primary treatment waste reduction applications[81]. Such plasmid-encoded genes are usually organized in large operons, but are also found on genomic islands, some of which are conjugative and can be engineered by similar methods[82].

Limitations in the capacity of MGEs. Because there are no known constraints on their size, plasmids encode a greater number and variety of accessory genes when compared with phages, in which the genome length is limited by the volume of their capsid. Nevertheless, all phages can accommodate foreign DNA and can carry small antibiotic-resistance transposons or toxin and invasion genes. The so-called broad-host-range plasmid replicons (for example, the Inc P1 group) can encode antibiotic-resistance genes and hydrocarbon-degradation genes depending on the environmental niche of their host microorganism[83]. Although the currently known actively transmissible conjugative elements seem to range in size from approximately 40 to 250 kb and are of low copy number, no limit has been identified for the number or size of independently replicating conjugative agents. These range in size from multicopy, mobilizable plasmids of one kb to single copy, 'mini-chromosomes' of a few Mb. Moreover, the promiscuous recombination capabilities of MGEs allow them to transfer many accessory genes of known and unknown function to chromosomes[84]. Therefore, the type and number of genes transferred seems to be limited only by the selective pressures of the host or niche, not by any characteristic of the core replicative genes of the MGEs. Gene loss is now recognized as important in sculpting chromosomes[85,86] and, although as yet unstudied, the loss rates for MGE genes are probably high. Only by sequencing many members of related plasmid or phage families will we begin to understand the dynamics and biology of MGE gene acquisition and maintenance.

Table 2 | **Web resources for MGE genomics**

Mobilome Resource	URL	Ref.
ACLAME	http://aclame.ulb.ac.be/Classification/description.html	114
Artemis	http://www.sanger.ac.uk/Software/Artemis	111
Gene Ontology	http://www.geneontology.org	113
IS Finder DB	http://www-is.biotoul.fr	122
Islander DB	http://129.79.232.60/cgi-bin/islander/islander.cgi	123
MUMmer	http://www.tigr.org/software/mummer	112
PlasMapper	http://wishart.biology.ualberta.ca/PlasMapper/index.html	124
Plasmid Genome Database	http://genomics.nerc-oxford.ac.uk/plasmiddb	121
T4-like Phage Genomes	http://phage.bioc.tulane.edu	125

The current state of MGE genomics

Although sequencing of cellular genomes first revealed a more prominent role for HGT in bacterial evolution than previously recognized, until recently, genome sequencing and analysis of the agents of HGT themselves — plasmids and bacteriophages — has remained sparse and haphazard. Many more phages and plasmids have been sequenced than bacterial genomes. However, the total size of all of these phage sequences is only approximately 30 Mb (that is, the size of approximately 6 bacterial genomes). Similarly, up to July 2005, the total size of all sequenced plasmids is only approximately 61 Mb and these sequences are derived from only about 40 different bacterial genera. Moreover, only approximately 20% of sequenced plasmids are of the large conjugative type that move significant amounts of DNA other than their own core genes. Many of these plasmid sequences are by-products of cellular-genome sequencing and, as such, are biased towards pathogenic microorganisms. Given the ubiquity of MGEs and their roles in important environmental processes and clinical problems as well as in bacterial evolution, it is remarkable that, more than a decade into the genomic era, so little attention has been directed towards them.

There are two main challenges in the genomic analysis of any organism. First, developing suitable information management and computational analysis resources and second, generating sufficient new sequence data so that the important scientific questions that face the field can be addressed. The genomic analysis of cellular organisms, both prokaryotic and eukaryotic, has passed these hurdles and the study of most of them is now moving towards coherent, if not entirely settled, bioinformatic infrastructures. Neither of these challenges has yet been addressed on any large scale for MGEs. In the following sections, we consider the roadblocks and requisite advances in more detail.

Challenges of MGE bioinformatics

Gene identification. There are many difficulties with annotating MGE sequences and consequently most sequenced MGEs are poorly annotated, especially when they are part of a bacterial-genome sequencing project. Only about a dozen phages are well characterized. Limited sequence similarity between functionally equivalent phage-encoded proteins makes assignment of functions to sequenced phage proteins difficult. Although there is a long-recognized tendency for conserved organization of functions in phage genomes[54,87–89] (FIG. 3), this feature is rarely exploited. Consequently, prophages or their remnants are not readily identifiable in bacterial genomes. There is a similar dearth of information on the 785 (from July 2005) plasmids listed in the Plasmid Genome Database at Oxford University (which is updated by regular automated parsing of the National Center for Biotechnology Information (NCBI) GenBank repository) (TABLE 2). Most of the entries in this database are small, cryptic plasmids or partially annotated, larger plasmids.

Although methods for effective automated gene prediction have received considerable attention during the past decade[90,91], these efforts have been based on cellular chromosomes and, in the case of MGEs, there is considerable room for improvement[92]. Newly developed stochastic approaches have improved gene prediction, but programs such as Gene-ID[93], GENMARK[94], GeneParser[95], GENSCAN[96], GeneMark.hmm[97] and Glimmer2 (REF. 98), need large datasets of annotated genes from the same or a closely related species. Plasmids and phages are much smaller than cellular chromosomes and can carry a larger fraction of genes from many different organisms, which makes them ineffective as training sets. The use of alternative training sets (for example, groups of related plasmids) or adjustment of parameters are not effective in predicting genes in MGEs, which often contain DNA segments with very different GC content and codon preferences.

Consequently, there are still open questions concerning the many so-called 'ORFan' genes[99] that lack significant similarity to any known sequence and that constitute a substantial fraction of MGE sequences. ORFan genes comprise approximately 30% of sequenced phage genomes compared with approximately 15% for the typical cellular genome (G. Lima-Mendez *et al.*, unpublished observations). Are ORFans unique genes, pseudogenes[100] or rapidly evolving genes[101] with a tendency to be AT rich[102]?

The study of their commonalities, regardless of their similarities to defined genes, could be an important component of the next generation of gene-prediction methods. More recent methods have been developed as alternatives to the single sequence input methods to make use of evolutionary relationships between genomic sequences. Functional regions of genomic sequences are more conserved than non-functional regions and regions of strong sequence conservation often correspond to protein-coding regions[103]. Homology-based gene-finding approaches include GeneWise[104] , ExoFish[105], CRITICA[106] and SGP-1 (REF. 107). There are no reports of the effectiveness of these alternatives in the analysis of MGEs.

Currently, most MGE annotation is performed manually, even at large sequencing centres (L. Hauser and J. Parkhill, personal communications). As the number of sequenced MGEs increases, the quality of the automated annotation methods might increase owing to larger training sets and smaller evolutionary gaps between the sequenced MGEs. However, with the huge diversity in the accessory genes of MGEs, this increase in quality might be limited without some important conceptual advances. Specialized databases with high quality (manually curated) annotations and classifications can fill in some blanks, as has been the case for the annotation of proteins with Pfam[108], Superfamily[109] and SCOP[110]. Tools for the pair-wise comparison of plasmids, such as Artemis[111] and MUMmer[112] allow visualization of homologous regions and rearrangements and can facilitate annotation of existing and newly generated sequences. The generally conserved order of genes in many phage genomes (FIG. 3) makes these agents an especially appropriate test bed for such strategies.

Nomenclature of MGEs and their genes. MGEs have a mosaic structure because of their recombinational promiscuity and replicative flexibility. Consequently, there are minimal physical or functional criteria for defining various types, and a confusing nomenclature plagues their classification. The most immediate need is to develop a bioinformatics system for MGEs that recognizes their unique features but that is also well integrated with other bioinformatics systems. However, there are some categories in standard bacterial-sequence databases that are not applicable to MGEs. For example, the organism name must first be documented when submitting a nucleotide sequence to a database; however, the 'natural host' of an MGE might not be known. Indeed, given the peripatetic nature of the so-called broad-host-range plasmids, a 'natural host' might not be an operative concept.

Also, the naming of naturally occurring plasmids and transposons has no universally agreed standards and no central nomenclature authority similar to the International Committee on Systematic Bacteriology, which names newly discovered bacteria. Phages and eukaryotic viruses are named by the International Committee for Taxonomy of Viruses. However, this taxonomy is based on virion morphology and nucleic-acid content, which do not necessarily correlate with sequence relationships. There are proposals for a unified phage nomenclature (see Bacteriophage Names 2000 in Online links box), which has rules that are similar to those proposed for IS sequences — a three letter designation for the host followed by a serial number; for example, phage Bcep6 from *Burkholderia cepacia* or ISEch4 from *Erwinia chrysanthemi*. Unfortunately, journals do not yet specify this format and few authors use it. For plasmids, a semiformal naming system operated among researchers during the mid 1970s, but has fallen into disuse. As a result, even common plasmids are not easily located in databases. The importance of establishing a rational nomenclature for MGEs and a central system for assigning unique identifiers to them cannot be over-emphasized. Presently, such an initiative exists only for ISs in the IS-Finder database, which also provides a site for new IS-sequence submissions and names (see TABLE 2).

A further complication is that there is no consensus ontology for many MGE-specific functions, even for phage head and tail genes or for plasmid conjugation genes, most of which are not included in the standard Gene Ontology (GO) database[113], which is the reference for functional annotations of eukaryotic and also prokaryotic genomes (TABLE 2). The ACLAME database project (TABLE 2; REF 114) was initiated as a resource for analysis of proteins encoded by prokaryotic MGEs. As MGEs share many common functions, having a single database facilitates functional assignment and comparative genomics. Other databases that are related to MGEs are listed in TABLE 2 and in the Online links box. Such initiatives serve as a platform for improved annotation and analysis of the prokaryotic mobilome and will also be important for defining MGE–host interactions.

It is now essential to establish standard formats for MGE sequence deposition and an ontology for MGE genes for use by all annotators and curators. Any MGE ontology will have to adopt a format compatible with that of the GO, ideally as an extension of it. Such an arrangement would serve as an archive of record and a research tool. So far, there is no consensus on how independently replicating MGEs should be taken into account in bacterial phylogeny. This will require the concerted effort of both the 'chromosome-centric' and the MGE research communities and is a *sine qua non* of understanding the biology of MGEs, their historical and contemporary evolution, and their role in the evolution and population biology of host bacteria.

Challenges to acquiring MGE sequence data

Despite the daunting state of nomenclature and annotation of existing MGE sequences, it remains important to gather more MGE-sequence information. The current dataset is biased towards MGEs from human pathogens and vastly under-represents the huge diversity of prokaryotic MGEs. Also, as for cellular chromosomes, with further sequence data,

explanatory patterns will emerge for core and accessory genes that are not yet discernible. Such patterns will improve annotation, assembly and functional analysis of MGEs, of chromosomal islands, and of direct shotgun sequencing of aggregate natural ecosystems (METAGENOMICS). Just as for organismal chromosomes, there are many MGEs that need to be sequenced and annotated, but the task is not impossible[115]. However, the production of high quality sequence from MGEs presents some unique problems that are not encountered in the sequencing of cellular genomes.

DNA preparation. Bacteriophage-DNA isolation is relatively easy if a suitable host exists, as it is conveniently packaged in the virion particle and can usually be acquired from a few ml of infected cells in sufficient quantity and purity for use in sequencing. By contrast, plasmids, of which there can be several in a natural isolate, must be physically segregated from each other and from chromosomal DNA before sequencing. The common occurrence of DNA repeats in plasmids makes the shotgun assembly of individual plasmids from pooled caesium chloride (CsCl) supercoiled bands very difficult. Moreover, few laboratories have access to instruments for CsCl density gradient ultracentrifugation or pulse field electrophoresis for the preparation of DNA. The recovery of large, low copy plasmids using alkaline-detergent lysis methods and commercial kits is poor. Direct lysis-in-the-well methods[116], which were originally designed as analytical, not preparative, tools have low and erratic DNA yields. As publicly supported sequencing services require microgram quantities of pure DNA, these technical drawbacks have caused a significant bottleneck in the first step in sequence determination. Recent improvements have been made in the recovery of pure, library-quality plasmid DNA using kits designed for the purification of bacterial artificial chromosomes (BACs), which are similar in size and copy number to large conjugative plasmids[117]. Nevertheless, there is no technique that is presently effective for the routine, robust, direct recovery of low copy plasmids that are larger than 250 kb. As sequencing libraries that provide good coverage of plasmids with sizes of approximately 100 kb can be obtained from a few micrograms of plasmid DNA, advances in separation sciences including microfluidics could be of great benefit in the high-throughput preparation of pure plasmid DNA, especially of the very large, conjugative plasmids that are central players in HGT. Finally, transposons and genomic islands do not produce abundant physically independent forms. As such, they can only be sequenced when the host replicon is sequenced.

Assembly. Once raw sequence data are obtained, MGEs present other problems at the assembly stage. Natural plasmids can contain genes that are present in the commonly used cloning vectors (for example, replication, mobilization and antibiotic-resistance genes). Therefore, using the entire vector sequence during the screening step before assembly can prevent scaffold (supertig) assembly by masking similar sequences in the natural plasmid. Employing different vectors can avoid this problem, as can the judicious choice of vector sub-sequences for screening.

As noted, MGEs frequently contain repeat sequences, which are the universal bane of shotgun assembly programs. Typical repeats include the ISs at the ends of transposons or the iterons that are involved in plasmid replication control. Depending on the size of the repeat unit and the depth of coverage in the library, restriction mapping or PCR walking might allow the placement of each repeat in a unique environment or the determination of the number of repeats in a tandem array.

As with any genome, there will be DNA that cannot be cloned in standard high-copy library vectors, leading to gaps in the sequence. Depending on the plasmid size, restriction mapping can help to identify unclonable regions, which can then be obtained by sequencing PCR products that cover the gaps. The new technique of optical mapping of restriction fragments[118] has proved valuable for whole-genome sequencing and will be valuable for MGE genomics in this regard as well as in the identification of repeated regions.

Conclusions and future perspectives

As each new prokaryotic genome sequence is completed, the scientific community is both amazed and frustrated by the number of genes for which there is no information on function or evolutionary history. To complicate the situation, much of the genome-sequencing effort has focused on laboratory strains that, in many cases, do not reflect the diversity of natural isolates of the same species. As noted above, the important phenotypic properties of natural isolates are often distributed among a half dozen or more plasmids. Most wild strains will also carry distinct repertoires of chromosomally inserted prophages, transposons and genomic islands. Making sense of this genetic fluidity to reveal the underlying regularities, and devising accurate mathematical descriptions of the biological processes involved, are two of the main challenges facing microbial genomics research. Meeting these challenges will require abundant and accurate genomic information on the different types of MGE.

Understanding the agents of HGT is essential for taking advantage of the opportunities provided by MGEs for bioremediation and genetic engineering as well as for avoiding the problems they cause as carriers of virulence and resistance genes. Genomic information will provide a springboard for MGE transcriptomics, proteomics and metabolomics, which will lead to an understanding of how the cell and its assortment of plasmids adapt to each other and exploit their environment. A new emphasis on laboratory and molecular epidemiology research will be required to understand the rates and extents of

METAGENOMICS
Sequencing of a clone library derived from the total DNA purified from a complex microbial ecosystem. This is followed by computer assembly of the reads into multiple linkage groups assumed to represent the organisms present in the community, including those that cannot be cultured.

physical gene transfer in the real (that is, geographic and zoonotic) world, the fraction of DNA that is incorporated in a new host, what types of genes are successfully transferred, and which bacterial hosts are most successful in making use of peripatetic genetic information. Ultimately, these efforts will lead to a better understanding of the core MGE genes and the accessory genes that are relevant to existing bacterial infectious diseases, essential bacterial symbioses in plant and animal biology, and the metabolic versatility of bacteria in bioremediation processes.

From a larger perspective, there are few subjects with greater implications for the understanding of evolution than these remarkably protean elements. MGEs afford their prokaryotic hosts access to vast genetic resources, which can be tried in a given niche–host combination, improved on and made available to other microorganisms, much like modern public domain software development. Although incomplete, current knowledge clearly holds promise that there are patterns, ripe for refinement and exploitation, even in this apparently chaotic welter of genetic phenomena.

1. Burrus, V. & Waldor, M. K. Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol.* **155**, 376–386 (2004).
A useful review of the role of ICEs in bacterial evolution.
2. Scott, J. R. & Churchward, G. G. Conjugative transposition. *Annu. Rev. Microbiol.* **49**, 367–397 (1995).
3. Toussaint, A. & Merlin, C. Mobile elements as a combination of functional modules. *Plasmid* **47**, 26–35 (2002).
4. Lawrence, J. G. & Hendrickson, H. Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50**, 739–749 (2003).
A succinct framing of important questions in horizontal genomics research.
5. Chaconas, G. & Chen, C. W. in *The Bacterial Chromosome* (ed. Higgins, P. N.) 525–539 (ASM Press, Washington DC, 2004).
6. Stewart, P. E., Byram, R., Grimm, D., Tilly, K. & Rosa, P. A. The plasmids of *Borrelia burgdorferi*: essential genetic elements of a pathogen. *Plasmid* **53**, 1–13 (2005).
7. Hinnebusch, J. & Tilly, K. Linear plasmids and chromosomes in bacteria. *Mol. Microbiol.* **10**, 917–922 (1993).
8. Lilley, A., Young, P. & Bailey, M. J. in *The Horizontal Gene Pool: Bacterial Plasmids and Gene Spread* (ed. Thomas, C. M.) 287–300 (Harwood Academic, Amsterdam, Netherlands, 2000).
9. Dahlberg, C. & Chao, L. Amelioration of the cost of conjugative plasmid carriage in *Eschericha coli* K12. *Genetics* **165**, 1641–1649 (2003).
A discussion of the cost of maintaining plasmids: why do bacteria tolerate them?
10. Bradley, D. E., Taylor, D. E. & Cohen, D. R. Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12. *J. Bacteriol.* **143**, 1466–1470 (1980).
11. Novick, R. P. Plasmid incompatibility. *Microbiol. Rev.* **51**, 381–395 (1987).
12. Couturier, M., Bex, F., Bergquist, P. L. & Maas, W. K. Identification and classification of bacterial plasmids. *Microbiol. Rev.* **52**, 375–395 (1988).
13. Helinski, D.R. in *The Horizontal Gene Pool: Bacterial Plasmids and Gene Spread* (ed. Thomas, C. M.) 1–21 (Harwood Academic, Amsterdam, Netherlands, 2000).
14. Redfield, R. J. *et al*. A novel CRP-dependent regulon controls expression of competence genes in *Haemophilus influenzae*. *J. Mol. Biol*. **347**, 735–747 (2005).
15. Gomis-Ruth, F. X., Sola, M., de la Cruz, F. & Coll, M. Coupling factors in macromolecular type-IV secretion machineries. *Curr. Pharm. Des*. **10**, 1551–1565 (2004).
16. Schroder, G. *et al.* TraG-like proteins of DNA transfer systems and of the *Helicobacter pylori* type IV secretion system: inner membrane gate for exported substrates? *J. Bacteriol.* **184**, 2767–2779 (2002).
17. Frost, L. S. in *Conjugation* (ed. Clewell, D. B.) 189–221 (Plenum, New York, 1993).
18. Possoz, C., Ribard, C., Gagnat, J., Pernodet, J. L. & Guerineau, M. The integrative element pSAM2 from *Streptomyces*: kinetics and mode of conjugal transfer. *Mol. Microbiol.* **42**, 159–66 (2001).
19. Pettis, G. S. & Cohen, S. N. Unraveling the essential role in conjugation of the Tra protein of *Streptomyces lividans* plasmid pIJ101. *Antonie Van Leeuwenhoek* **79**, 247–250 (2001).
20. Garret, R. A. *et al*. in *Plasmid Biology* (eds Funnell, B. E. & Phillips, G. J.) 377–392 (ASM Press, Washington DC, 2004).
21. Wang, J., Parsons, L. M. & Derbyshire, K. M. Unconventional conjugal DNA transfer in mycobacteria. *Nature Genet.* **34**, 80–84 (2003).
22. Lybarger, S. R. & Sandkvist, M. Microbiology. A hitchhiker's guide to type IV secretion. *Science* **304**, 1122–1123 (2004).
23. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nature Rev. Microbiol.* **2**, 241–249 (2004).
24. Gomis-Ruth, F. X. *et al.* The bacterial conjugation protein TrwB resembles ring helicases and F1-ATPase. *Nature* **409**, 637–641 (2001).
25. Zechner, E. L. *et al*. in *The Horizontal Gene Pool: Bacterial Plasmids and Gene Spread* (ed. Thomas, C. M.) 87–174 (Harwood Academic, Amsterdam, Netherlands, 2000).
26. Cascales, E. & Christie, P. J. The versatile bacterial type IV secretion systems. *Nature Rev. Microbiol.* **1**, 137–149 (2003).
27. Lawley, T. D., Wilkins, B. M. & Frost, L. S. in *Plasmid Biology* (eds Funnell, B. E. & Phillips, G. J.) 203–226 (ASM Press, Washington DC, 2004).
28. Lawley, T. D., Klimke, W. A., Gubbins, M. J. & Frost, L. S. F factor conjugation is a true type IV secretion system. *FEMS Microbiol. Lett*. **224**, 1–15 (2003).
Defines the relationship of F-like T4SSs to P-like T4SSs.
29. Boltner, D. & Osborn, A. M. Structural comparison of the integrative and conjugative elements R391, pMERPH, R997, and SXT. *Plasmid* **51**, 12–23 (2004).
30. Peabody, C. R. *et al.* Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149**, 3051–3072 (2003).
31. He, S. Y., Nomura, K. & Whittam, T. S. Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim. Biophys. Acta* **1694**, 181–206 (2004).
32. Planet, P. J., Kachlany, S. C., DeSalle, R. & Figurski, D. H. Phylogeny of genes for secretion NTPases: identification of the widespread *tadA* subfamily and development of a diagnostic key for gene classification. *Proc. Natl Acad. Sci. USA* **98**, 2503–2508 (2001).
33. Savvides, S. N. *et al.* VirB11 ATPases are dynamic hexameric assemblies: new insights into bacterial type IV secretion. *EMBO J.* **22**, 1969–1980 (2003).
34. Kim, S. R. & Komano, T. The plasmid R64 thin pilus identified as a type IV pilus. *J. Bacteriol.* **179**, 3594–3603 (1997).
35. Model, P. & Russel, M. Prokaryotic secretion. *Cell* **61**, 739–741 (1990).
36. Macnab, R. M. Type III flagellar protein export and flagellar assembly. *Biochim. Biophys. Acta* **1694**, 207–217 (2004).
37. Averhoff, B. DNA transport and natural transformation in mesophilic and thermophilic bacteria. *J. Bioenerg. Biomembr.* **36**, 25–33 (2004).
38. Cascales, E. & Christie, P. J. Definition of a bacterial type IV secretion pathway for a DNA substrate. *Science* **304**, 1170–1173 (2004).
Immunoprecipitation of Vir protein–DNA complexes defines the path of the DNA through the conjugative pore.
39. Hamilton, H. L., Dominguez, N. M., Schwartz, K. J., Hackett, K. T. & Dillard, J. P. *Neisseria gonorrhoeae* secretes chromosomal DNA via a novel type IV secretion system. *Mol. Microbiol.* **55**, 1704–1721 (2005).
40. Cascales, E. & Christie, P. J. *Agrobacterium* VirB10, an ATP energy sensor required for type IV secretion. *Proc. Natl Acad. Sci. USA* **101**, 17228–17233 (2004).
VirB10 has TonB-like properties indicating that it is involved in signalling between the outer and inner membranes.
41. Christie, P. J. Type IV secretion:the *Agrobacterium* VirB/D4 and related conjugation systems. *Biochim. Biophys. Acta* **1694**, 219–234 (2004).
42. Kalkum, M., Eisenbrandt, R. & Lanka, E. Protein circlets as sex pilus subunits. *Curr. Protein Pept. Sci.* **5**, 417–424 (2004).
43. Lai, E. M., Eisenbrandt, R., Kalkum, M., Lanka, E. & Kado, C. I. Biogenesis of T pili in *Agrobacterium tumefaciens* requires precise VirB2 propilin cleavage and cyclization. *J. Bacteriol.* **184**, 327–330 (2002).
44. Clewell, D. B. & Francia, M. V. in *Plasmid Biology* (eds. Funnell, B. E. & Phillips, G. J.) 227–256 (ASM Press, Washington DC, 2004).
45. Salyers, A. A., Shoemaker, N. B., Stevens, A. M. & Li, L. Y. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiol. Rev.* **59**, 579–90 (1995).
46. Charlebois, R. L., She, Q., Sprott, D. P., Sensen, C. W. & Garrett, R. A. *Sulfolobus* genome: from genomics to biology. *Curr. Opin. Microbiol.* **1**, 584–588 (1998).
47. Wilkins, B. M. & Frost, L. S. in *Molecular Medical Microbiology* (ed. Sussman, M.) 355–400 (Academic, London, 2001).
48. Papke, R. T., Koenig, J. E., Rodriguez-Valera, F. & Doolittle, W. F. Frequent recombination in a saltern population of *Halorubrum*. *Science* **306**, 1928–1929 (2004).
49. Ramirez-Arcos, S., Fernandez-Herrero, L. A., Marin, I. & Berenguer, J. Anaerobic growth, a property horizontally transferred by an Hfr-like mechanism among extreme thermophiles. *J. Bacteriol.* **180**, 3137–3143 (1998).
50. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).
The first published genome sequence, which predates the advent of DNA sequencing techniques.
51. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162**, 729–773 (1982).
52. Canchaya, C., Fournous, G. & Brussow, H. The impact of prophages on bacterial chromosomes. *Mol. Microbiol.* **53**, 9–18 (2004).
53. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brussow, H. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
A good perspective on how bacterial genomics reveals the main impact of phages on bacterial chromosome evolution.
54. Pedulla, M. L. *et al.* Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171–182 (2003).
55. Merril, C. R., Scholl, D. & Adhya, S. L. The prospect for bacteriophage therapy in Western medicine. *Nature Rev. Drug Discov.* **2**, 489–497 (2003).
56. Zhang, S. Fabrication of novel biomaterials through molecular self-assembly. *Nature Biotechnol.* **21**, 1171–1178 (2003).
57. Lwoff, A. Lysogeny. *Bacteriol. Rev.* **17**, 269–337 (1953).
58. Freeman, V. J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol.* **61**, 675–688 (1951).
59. Hendrix, R. W. Bacteriophage genomics. *Curr. Opin. Microbiol.* **6**, 506–511 (2003).
60. Zinder, N. D. & Lederberg, J. Genetic exchange in *Salmonella*. *J. Bacteriol.* **64**, 679–699 (1952).
61. Mizuuchi, K. & Baker, T. in *Mobile DNA II* (eds. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A.) 12–23 (ASM press, Washington DC, 2002).
62. Hughes, V. M. & Datta, N. Conjugative plasmids in bacteria of the pre-antibiotic era. *Nature* **302**, 725–726 (1983).
63. Mazel, D. & Davies, J. Antibiotic resistance in microbes. *Cell. Mol. Life Sci.* **56**, 742–754. (1999).

64. Bennett, P. M. Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methods Mol. Biol.* **266**, 71–113 (2004).
65. Hall, R. M. Mobile gene cassettes and integrons: moving antibiotic resistance genes in Gram-negative bacteria. *Ciba Found. Symp.* **207**, 192–202; discussion 202–205 (1997).
66. Liebert, C. A., Hall, R. M. & Summers, A. O. Transposon Tn*21*, flagship of the floating genome. *Microbiol. Mol. Biol. Rev.* **63**, 507–522 (1999).
67. Novick, R. P. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of *Staphylococcus aureus*. *Plasmid* **49**, 93–105 (2003).
68. Shipley, P. L., Gyles, C. L. & Falkow, S. Characterization of plasmids that encode for the K88 colonization antigen. *Infect. Immun.* **20**, 559–566 (1978).
Early recognition of role for plasmids in the bacterial colonization of animal hosts.
69. Schell, J. *et al.* Interactions and DNA transfer between *Agrobacterium tumefaciens*, the Ti-plasmid and the plant host. *Proc. R. Soc. Lond., B, Biol. Sci.* **204**, 251–266 (1979).
Early demonstration of pathogenesis that involves plasmid-directed transfer of DNA from a bacterium to a plant.
70. Brussow, H., Canchaya, C. & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602, (2004).
71. Banks, D. J., Beres, S. B. & Musser, J. M. The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends Microbiol.* **10**, 515–521 (2002).
72. Boyd, E. F. & Brussow, H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol.* **10**, 521–529 (2002).
73. Koehler, T. M. *Bacillus anthracis* genetics and virulence gene regulation. *Curr. Top. Microbiol. Immunol.* **271**, 143–164 (2002).
74. Okinaka, R. T. *et al.* Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harbouring the anthrax toxin genes. *J. Bacteriol.* **181**, 6509–6515 (1999).
75. Crossman, L. C. Plasmid replicons of *Rhizobium*. *Biochem. Soc. Trans.* **33**, 157–158 (2005).
76. Sullivan, J. T. *et al.* Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. *J. Bacteriol.* **184**, 3086–3095 (2002).
77. Paul, J. H. & Sullivan, M. B. Marine phage genomics: what have we learned? *Curr. Opin. Biotechnol.* **16**, 299–307 (2005).
78. Wade, N. Court says lab-made life can be patented. *Science* **208**, 1445 (1980).
79. Kellogg, S. T., Chatterjee, D. K. & Chakrabarty, A. M. Plasmid-assisted molecular breeding: new technique for enhanced biodegradation of persistent toxic chemicals. *Science* **214**, 1133–1135 (1981).
80. Lindstrom, J. E. *et al.* Microbial populations and hydrocarbon biodegradation potentials in fertilized shoreline sediments affected by the T/V Exxon Valdez oil spill. *Appl. Environ. Microbiol.* **57**, 2514–2522 (1991).
81. von Canstein, H., Li, Y. & Wagner-Dobler, I. Long-term performance of bioreactors cleaning mercury-contaminated wastewater and their response to temperature and mercury stress and mechanical perturbation. *Biotechnol. Bioeng.* **74**, 212–219 (2001).
82. van der Meer, J. R. & Sentchilo, V. Genomic islands and the evolution of catabolic pathways in bacteria. *Curr. Opin. Biotechnol.* **14**, 248–254 (2003).
83. Schluter, A. *et al.* The 64,508 bp IncP-1b antibiotic multiresistance plasmid pB10 isolated from a waste-water treatment plant provides evidence for recombination between members of different branches of the IncP-1b group. *Microbiology* **149**, 3139–3153 (2003).
84. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
85. Frank, A. C., Amiri, H. & Andersson, S. G. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica* **115**, 1–12 (2002).
86. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596 (2001).
87. Botstein, D. A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.* **354**, 484–490 (1980).
A seminal paper on the mosaic nature of lambdoid phages, which is now clearly applicable to several other phage families.
88. Casjens, S., Hatfull, G. & Hendrix, R. Evolution of the dsDNA tailed-bacteriophage genomes. *Semin. Virol.* **3**, 383–397 (1992).
89. Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276 (2003).
90. Burge, C. B. & Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346–354 (1998).
91. Claverie, J. M. Computational methods for exon detection. *Mol. Biotechnol.* **10**, 27–48 (1998).
92. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
93. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226**, 141–157 (1992).
94. Borodovsky, M. & McIninch, J. Recognition of genes in DNA sequence with ambiguities. *Biosystems* **30**, 161–171 (1993).
95. Snyder, E. E. & Stormo, G. D. Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* **248**, 1–18 (1995).
96. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
97. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
98. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
99. Fischer, D. & Eisenberg, D. Finding families for genomic ORFans. *Bioinformatics* **15**, 759–762 (1999).
100. Amiri, H., Davids, W. & Andersson, S. G. Birth and death of orphan genes in *Rickettsia*. *Mol. Biol. Evol.* **20**, 1575–1587 (2003).
101. Domazet-Loso, T. & Tautz, D. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* **13**, 2213–2219 (2003).
102. Daubin, V. & Ochman, H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* **14**, 1036–1042 (2004).
103. Morgenstern, B. *et al.* Exon discovery by genomic sequence alignment. *Bioinformatics* **18**, 777–787 (2002).
104. Meyer, I. M. & Durbin, R. Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics* **18**, 1309–1318 (2002).
105. Crollius, H. R. *et al.* Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* **10**, 939–949 (2000).
106. Badger, J. H. & Olsen, G. J. CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**, 512–524 (1999).
107. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. & Guigo, R. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11**, 1574–1583 (2001).
108. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
109. Gough, J. & Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30**, 268–272 (2002).
110. Andreeva, A. *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**, D226–D229 (2004).
111. Berriman, M. & Rutherford, K. Viewing and annotating sequence data with Artemis. *Brief. Bioinformatics* **4**, 124–132 (2003).
112. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
113. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
114. Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32** (Database issue), D45–D49 (2004).
115. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**, e10 (2005).
116. Eckhardt, T. A rapid method for the identification of plasmid desoxyribonucleic acid in bacteria. *Plasmid* **1**, 584–588 (1978).
117. Williams, L., Miller, D., Summers, A. O. & Detter, C. Fast, cheap, and easy preparation of library-quality DNA from 100+ kb, low copy eubacterial plasmids. *Plasmid* **53**, 45–46 (2005).
118. Guo, X. H., Huff, E. J. & Schwartz, D. C. Sizing of large DNA molecules by hook formation in a loose matrix. *J. Biomol. Struct. Dyn.* **11**, 1–10 (1993).
119. Funnell, B. E. & Phillips, G. J. (eds) *Plasmid Biology* (ASM Press, Washington DC, 2004).
120. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds) *Mobile DNA II* (ASM Press, Washington DC, 2002).
121. Molbak, L. *et al.* The plasmid genome database. *Microbiology* **149**, 3043–3045 (2003)
122. Chandler, M. & Mahillon, J. in *Mobile DNA II* (eds. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 305–366 (ASM Press, Washington DC, 2003).
123. Mantri, Y. & Williams, K. P. Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.* **32** (Database issue), D55–D58 (2004).
124. Dong, X., Stothard, P., Forsythe, I. J. & Wishart, D. S. PlasMapper: a web server for drawing and auto-annotating plasmid maps. *Nucleic Acids Res.* W365–W371 (2004).
125. Galperin, M.Y. The molecular biology database collection. *Nucleic Acids Res.* **33**, (Database issue) Entry no. 750 (2005).

Acknowledgements

We thank M. Syvanen for the public domain software metaphor for MGEs and the reviewers for their thoughtful critiques. Work in the laboratories of R.L. and A.T. is supported by the Fonds National de la Recherche Scientifique, the Université Libre de Bruxelles, Belgium, and the European Space Agency. L.S.F. acknowledges support from the Canadian Institutes of Health Research and Natural Sciences and Engineering Research Council of Canada. A.O.S. acknowledges support from the US Department of Energy (DOE) Genomes-To-Life Program, the assistance of L. Williams, the staff of the DOE Joint Genome Institute, Walnut Creek, California and Oak Ridge National Laboratory, Tennessee.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:
ACLAME: http://aclame.ulb.ac.be/Classification/description.html
Artemis: http://www.sanger.ac.uk/Software/Artemis
Gene Ontology: http://www.geneontology.org
MUMmer: http://www.tigr.org/software/mummer
National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov
Plasmid Genome Database: http://genomics.nerc-oxford.ac.uk/plasmiddb
Swiss-Prot: http://www.expasy.ch
VirB4 | VirB10 | VirB11

FURTHER INFORMATION

Laura Frost's laboratory: http://www.biology.ualberta.ca/faculty/laura_frost/index.php
Raphael Leplae's laboratory: http://www.scmbb.ulb.ac.be/~raphael
Anne Summers' laboratory: http://www.uga.edu/mib/people/summers.htm
Ariane Toussaint's laboratory: http://www.ulb.ac.be/rechinventaire/chercheurs/3/CH2403.htm
Bacteriophage Names 2000: http://www.mansfield.ohio-state.edu/~sabedon/names_introduction.htm
Database of Plasmid Replicons: http://www.essex.ac.uk/bs/staff/osborn/DPR_home.htm
Access to this interactive links box is free online.