

# Genesis, effects and fates of repeats in prokaryotic genomes

Todd J. Treangen<sup>1,2</sup>, Anne-Laure Abraham<sup>1,2</sup>, Marie Touchon<sup>1,2</sup> & Eduardo P.C. Rocha<sup>1,2</sup>

<sup>1</sup>UPMC Univ Paris 06, Atelier de BioInformatique, Paris, France; and <sup>2</sup>Institut Pasteur, Microbial Evolutionary Genomics, CNRS, URA2171, Paris, France

**Correspondence:** Eduardo P. C. Rocha,  
Institut Pasteur, Microbial Evolutionary  
Genomics, CNRS, URA2171, 28 rue Dr Roux,  
F-75015 Paris, France. Tel.: +33 1 44 27 65 36;  
fax: +33 1 44 27 63 12; e-mail:  
erocha@pasteur.fr

Received 15 October 2008; revised 2 February  
2009; accepted 6 February 2009.  
First published online 10 March 2009.

DOI:10.1111/j.1574-6976.2009.00169.x

Editor: Tone Tonjum

## Keywords

amplifications; recombination; comparative  
genomics; molecular evolution.

## Abstract

DNA repeats are causes and consequences of genome plasticity. Repeats are created by intrachromosomal recombination or horizontal transfer. They are targeted by recombination processes leading to amplifications, deletions and rearrangements of genetic material. The identification and analysis of repeats in nearly 700 genomes of bacteria and archaea is facilitated by the existence of sequence data and adequate bioinformatic tools. These have revealed the immense diversity of repeats in genomes, from those created by selfish elements to the ones used for protection against selfish elements, from those arising from transient gene amplifications to the ones leading to stable duplications. Experimental works have shown that some repeats do not carry any adaptive value, while others allow functional diversification and increased expression. All repeats carry some potential to disorganize and destabilize genomes. Because recombination and selection for repeats vary between genomes, the number and types of repeats are also quite diverse and in line with ecological variables, such as host-dependent associations or population sizes, and with genetic variables, such as the recombination machinery. From an evolutionary point of view, repeats represent both opportunities and problems. We describe how repeats are created and how they can be found in genomes. We then focus on the functional and genomic consequences of repeats that dictate their fate.

## Introduction

Genetic amplifications were associated with phenotypic diversification as early as the 1910s (for a review see, Taylor & Raes, 2004). Such early studies played an important role in the establishment of the association between genotypic and phenotypic variation. They also led to speculations about the association between repeated elements and functional diversification, via the creation of genetic variation (Bridges, 1935). The amplification of genetic material creates duplicated sequences that are expected to be at least partly redundant. Accordingly, selection should be relaxed on the pair of repeated elements, i.e. on the pair of copies of the repeat (Serebrovsky, 1938). How frequently are repeats fixed in populations and how such fixation occurs became a highly popular subject of research over the years, especially since Ohno (1970) suggested that most novelties in complex genomes arose from duplication and divergence of genetic elements. It is almost consensual that gene and genome duplications are the major source of functional novelties in

the genomes of animals and plants. But in prokaryotes, the amplifications of genetic material are unstable and compete with horizontal gene transfer for the major role among the mechanisms allowing the acquisition of novel functions.

Apart from their involvement in gene duplication, repeats are targeted by recombination processes and thus play extremely important roles in genome stability and plasticity. Through recombination, repeats increase the rates of rearrangement, amplification and deletion of genetic material. Deletions and amplifications of short tandem repeats are particularly frequent and often result in phase variation, i.e. in stochastic switching of gene expression. By engaging in nonreciprocal recombination (gene conversion), repeats can also lead to new combinations of pre-existing alleles. As we shall discuss, many prokaryotes use repeats and homologous recombination to generate high rates of sequence variation related to key functions. Finally, closely spaced inverted repeats (IRs) result in structured DNA or RNA molecules. Such repeats are widespread in genomes and often interfere with gene expression. Therefore, repeated elements have the

potential to increase genome plasticity and have been recruited to produce localizing sequence diversification by recombination processes.

When compared with eukaryotes, bacteria and archaea have small compact genomes with typically *c.* 85–90% of sequence coding for proteins or stable RNA molecules. Most of the remaining genome corresponds to regulatory regions. Because few chromosomal regions are devoid of functional constraints, prokaryotic genomes could be expected to be devoid of repeated elements. While prokaryotic genomes do not attain the number of repeats found in some eukaryotes, they sometimes contain hundreds or even thousands of large repeats. Thanks to existing repeat detection tools, these elements can be identified *in silico* when genome sequences are available. In this article, we will review how repeats arise in genomes, how they evolve and what are their consequences in terms of the structural and the functional evolution of genomes. This cannot be an exhaustive review on the types and effects of repeats in genomes because the literature is too vast. Therefore, we concentrate on general mechanisms and on some of the best-studied examples of the roles and effects of repeats in genomes expecting to give a bird's eye view of this very dynamic topic.

## What is a DNA repeat?

### Multiple perspectives on repeats

To precisely define a DNA repeat, one must consider mathematical and biological aspects. Mathematically, a repeat is a subsequence of a given genome that resembles another subsequence in the same genome. One is usually interested in repeats unexpected in a random assembly of the genetic text. Such repeats have not arisen by purely stochastic processes and usually uncover a relevant biological phenomenon. Biologically, repeats are interesting because they are a source of functional overlapping and sequence recombination. Overlapping arises when repeats establish an association between genetic elements. For example, two proteins with a similar protein domain share a biochemical property that allows some functional redundancy (Pasek *et al.*, 2006). Genes can also be coexpressed if they have similar copies of a repeat in their regulatory regions. In this case, repeats can effectively produce associations between genes in genetic networks. Moreover, repeats may include entire genes or operons, in which case some functional redundancy arises from the overlapping of functions between the two copies. Such overlapping may be nearly complete, as in the multiple copies of rRNA gene operons in many genomes. One should not, however, equate overlap with redundancy. For example, the multiplicity of rRNA gene operons is the fruit of selection for high gene dosage in moments of exponential growth (Klappenbach

*et al.*, 2000). Hence, even if the copies are identical they are not redundant in the sense that deletion of one copy cannot be fully compensated by the remaining copies. Finally, recombination is probably the mechanism that most frequently comes to mind when thinking about repeats in genomes. This is because most recombination processes require some level of sequence similarity between the sequences. Considering the three previous points, *i.e.* mathematics, functional overlapping and sequence recombination, can one establish a meaningful definition of a DNA repeat?

Many methods have been developed to find the significance of repeats in genomes but two problems commonly arise: the definition of a null model and the reliability of mathematical approximations. The null model, *i.e.* the definition of what one expects 'by chance alone', often assumes that genomes arise by the random assembling of a predefined oligonucleotide usage. As a result, there is no perfectly correct null model as one can build a random genome by shuffling nucleotides, dinucleotides, codons, etc. (Robin *et al.*, 2007). Because genetic texts are the result of the super-positioning of many codes, associated with different cellular processes, these models are necessarily crude approximations of the reality. Sometimes the choice of the null models is inherently simple for a given problem; for example, a codon-based model is a reasonable choice to identify intragenic repeats. In most circumstances, the choice of a model is much less obvious. For example, the analysis of an intergenic region of a mixture of coding and noncoding regions can require the use of several different models to account for different effects. Once a model is chosen, it is still necessary to assess the statistical significance of the results. Statistical analyses often rely on a number of mathematical approximations, such as regarding sequences as nearly infinite and/or oligonucleotide usage following some asymptotic statistical distribution. Furthermore, statistics for degenerate repeats with insertions and deletions are nearly always empirical (Waterman & Vingron, 1994). This can be a considerable problem when searching for repeats in small sequences, but it is rarely an issue when analyzing complete genomes.

Overlapping functions are extremely hard to uncover from biological sequences because the function of most repeated elements is poorly known. Even when such repeats concern genes of known function, it is still hazardous to quantify the degree of functional overlapping between paralogues from sequence identity alone. As a result, this parameter is rarely accounted for when defining repeats.

Repeats-mediated recombination processes usually may involve homologous recombination, slipped-strand mispairing or site-specific recombination. Intragenomic recombination between repeats has extremely important consequences in terms of sequence diversification and genome dynamics.

The frequency and extent of recombination will crucially depend on the repeat length, distance between and similarity between the copies and on the recombination mechanisms. In the next section we briefly detail these for the two most generic modes of recombination involving repeats: homologous recombination and slipped-strand mispairing.

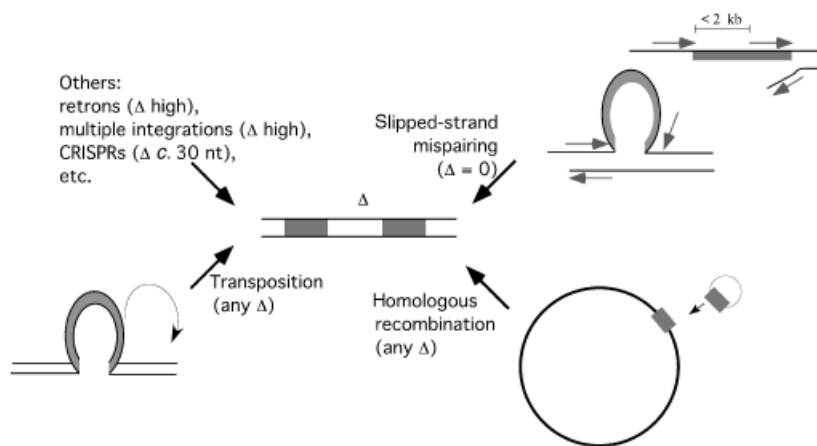
### Recombination and repeats

In the central step of homologous recombination, RecA promotes strand exchange between homologous or homeologous sequences. It is out of the scope of this review to detail recombination mechanisms. The interested reader can consult some of the excellent available reviews (Kowalczykowski, 2000; Amundsen & Smith, 2003; Wilson *et al.*, 2003; Lovett, 2004; Michel *et al.*, 2004; Rocha *et al.*, 2005). RecA is usually regarded as the key element of homologous recombination and is only lacking in very small genomes of obligatory endomutualists, such as *Buchnera*, that have lost most repair mechanisms and that typically lack repeated elements (Frank *et al.*, 2002; Gil *et al.*, 2003). Recently, it has been shown that two plasmids may recombine efficiently by a RecA-independent pathway when all cellular exonucleases are inactivated (Dutra *et al.*, 2007). It is, however, unclear as to how frequently such conditions are encountered in nature. Strand exchange by RecA occurs in two steps and its frequency crucially depends on the similarity between the sequences. In the initial step, the efficiency of recombination is very low in the presence of mismatches, especially near the 3' end of the incoming strand, and a stretch of strict identity called the minimum efficient processing segment (MEPS) is necessary for recombination to proceed at significant rates (Shen & Huang, 1986; Sagi *et al.*, 2006). The value of the MEPS seems remarkably similar among distant genomes with estimated values in the range 20–50 nt in *Escherichia coli*, *Bacillus subtilis*, *Streptococcus pneumoniae* and *Agrobacterium tumefaciens* (Costechareyre *et al.*, 2009). The number of MEPS shared by the two sequences determines the frequency of homologous recombination, suggesting that the strict homology required at the early stages of recombination is the rate-limiting step of the process (Vulic *et al.*, 1997). This leads to an exponential decay of the frequency of recombination with sequence divergence. The stringency of homologous recombination in sequence homology is provided by the action of RecA itself, but it is highly increased in the presence of mismatch repair, which will effectively forbid mismatches at the early stages of strand exchange (Rayssiguier *et al.*, 1989). Insertions and deletions between the sequences typically have very deleterious effects on the initial rates of the strand exchange process (Bucka & Stasiak, 2001). After strand exchange, branch migration by RuvABC is much less stringent regarding mismatches. RuvABC resolvases are present in the vast majority of genomes

(Rocha *et al.*, 2005), although RuvC can sometimes be replaced by phage analogues (Sharples *et al.*, 1999). The initiation of homologous recombination in *E. coli* may follow the RecBCD or the RecFOR pathway. Both pathways work to provide an ssDNA molecule coated with RecA to allow the invasion of a homologous molecule. Apart from *recR*, which has a key role in the RecFOR pathway, most other genes implicated in this step are absent in several genomes even among the ones known to engage extensively in homologous recombination (Rocha *et al.*, 2005). It is therefore difficult to predict the frequency of homologous recombination from the patterns of presence or absence of these genes.

Slipped-strand mispairing is thought to occur throughout the prokaryotic world by repeat mispairing upon replication. It is thought to be the most frequent illegitimate recombination mechanism in bacteria that does not depend upon recombinases. The frequency of illegitimate recombination through slipped-strand mispairing decreases exponentially with the distance between the copies of the repeat (Chédin *et al.*, 1994; Lovett *et al.*, 1994) and increases with the length of the repeated elements (Peeters *et al.*, 1988; Pierce *et al.*, 1991). Experimental studies have been able to identify recombination events for 8-nt repeats at a distance of up to 987 bp (Albertini *et al.*, 1982), for 18-nt repeats at a distance of 2313 bp (Chédin *et al.*, 1994) and for 24-nt repeats at a distance of 1741 bp (Singer & Westlye, 1988). Interestingly, modeling the dependence on the frequency of slippage in function of repeat size and distance suggests that the process follows the same parameters both in *E. coli* and in *B. subtilis* (Oliveira *et al.*, 2008). Repeats separated by > 2 kb recombine by slipped-strand mispairing at very low frequencies presumably because misalignment occurs at the replication fork (Lovett, 2004). In general, slipped-strand mispairing is so vastly diminished by mismatches and insertions and deletions that the frequency of recombination depends mostly on the size of the core of strict contiguous identity independently of the length of the larger degenerate repeat(s) (Petes *et al.*, 1997).

In the following, our operational definition of DNA repeat depends on three factors: (1) the distance between the copies, (2) their similarity and (3) their length. All of these variables will be intimately related with the statistical and biological significance of the repeat, which in turn will depend on genome length, sequence composition and the recombination machinery. As a guideline, nondegenerate repeats > 25 nt are statistically significant in most prokaryotic genomes ( $P < 0.01$  in *E. coli*) (Rocha *et al.*, 1999), and they are expected to engage in homologous recombination (Shen & Huang, 1986). Repeats > 12 nt and < 2 kb apart are statistically significant ( $P < 0.01$  in *E. coli*) (Rocha, 2003a), and engage in slipped-strand mispairing. Tandem repeats are unstable at lower lengths and more than five or



**Fig. 1.** Mechanisms of repeat creation.

$\Delta$  represents the length of the spacer between the two repeats expected under the different mechanisms.

six tandem motifs are typically significant, depending on genome composition (Mrazek *et al.*, 2007). These values increase for degenerate repeats, but the exact thresholds depend on the penalties for mismatches and gaps. In general, DNA repeats with low sequence identity and lacking a core of strict identity will not recombine and we will therefore ignore them. This is the case of the vast majority of paralogous genes in genomes.

## The origin of repeats

A repeat can be created in many ways (Fig. 1). Generic repeats can arise by general recombination processes, but more specific families of repeats may originate from specialized recombination processes such as site-specific integration for prophages, transposition for insertion sequences (ISs) or poorly known mechanisms for elements such as retrons or clustered regularly interspaced short palindromic repeats (CRISPR). Once created, repeats engage in general recombination processes independently of their origin.

### Generic repeats

Generic repeats may arise by horizontal gene transfer, homologous recombination between smaller or more degenerate repeats or illegitimate recombination between potentially very small and close repeats. Horizontal transfer results in the creation of repeats whenever the incoming DNA contains elements similar to information already present in the genome. Laterally transferred information may be integrated into the chromosome by site-specific recombination, for example by phage integrases, or by homologous recombination. In the latter case, if the incoming DNA contains homology with the chromosome at both extremities, then a double cross-over leads to genetic replacement (Niaudet *et al.*, 1985). Eventually, if the transferred element is (made) circular, then one single region of similarity with the chromosome allows integration of the genetic element by homologous recombination in a Camp-

bell-like manner (Duncan *et al.*, 1978). In this case, the repeats are created close together as they are spaced by the sequence of the incoming element. In *B. subtilis*, this seems to be a preponderant way of integrating genetic elements, because large repeats are often separated by c. 10 kb, which is the average size of uptake DNA segments in naturally transformable cells (Rocha *et al.*, 1999).

Homologous recombination between short or degenerate repeats may result in gene conversion and therefore effectively increase the region of similarity between the elements, thereby creating larger repeats (Santoyo & Romero, 2005). Gene conversion also allows maintaining the sequence similarity between repetitive elements, such as in rRNA gene operons (Harvey & Hill, 1990) or in the repeated *tuf* gene in *Salmonella enterica* (Hughes, 2000). Exactly how some genes manage to diverge and escape the homogenizing effect of gene conversion is not yet clear. For example, while *tuf* genes are duplicated in many enterobacteria, where they are very similar because they engage in gene conversion, in *Yersinia* these genes are much less similar. Recent analyses show that these genes have evolved independently for a long period time without gene conversion (Isabel *et al.*, 2008). The multiple rRNA gene copies show signs of gene conversion, demonstrating that the mechanism works in *Yersinia*. Thus, it may be that if sequence diverges after a given threshold, it may become free from the homogenizing effect of gene conversion because recombination frequency drops exponentially, as mentioned above.

Illegitimate recombination between small degenerate closely spaced repeats may lead to 'dislocation mutagenesis,' which results in the production of a strict repeat from a degenerate one (Schaaper, 1988). In the latter case, given slipped-strand mispairing properties, repeats must be nearby in the genome but they may only be a few nucleotides long. Such repeats can be so small that they are easily found in genetic texts (Karlin & Cardon, 1994). It has therefore been proposed that repeated elements may arise in genomes from rare slippage events between small but abundant

repeats (Edlund & Normark, 1981). These would create larger strict repeats that can further engage into illegitimate or even homologous recombination at high rates (Levinson & Gutman, 1987). Hence, large repeats can be created practically *ex nihilo* by a succession of recombination steps in what has been named the amplicon model of repeat generation (Romero & Palacios, 1997). First, illegitimate recombination leads to the creation of short close repeats. Second, these repeats are low-frequency amplicons that may amplify in tandem. Third, resulting amplifications are large and can expand and contract by homologous recombination.

### Selfish elements

The proliferation of selfish elements in genomes is one of the main sources of repeats. Among these, transposable elements (TEs) are by far the best studied and the most abundant. They are usually classified into two major groups based on their structure and transposition mechanism: retrotransposons or class I elements, which transpose by an RNA intermediate, and DNA transposons or class II elements, which use a DNA intermediate. Three kinds of prokaryotic retrotransposons have been characterized: group II introns (Ferat & Michel, 1993; Simon *et al.*, 2008), retrons (Temin, 1989; Lampson *et al.*, 2005) and diversity-generating retroelements (Medhekar & Miller, 2007). Group II introns are the only TEs distributed among all three domains of life. They are self-splicing introns that multiply via reverse transcription and they are capable of carrying out both self-splicing and retromobility reactions (Simon *et al.*, 2008). Retrtons are genetic elements that produce multicopy single-stranded DNA covalently linked to RNA (msDNA) by a reverse transcriptase. The functions of msDNA are still unknown (Lampson *et al.*, 2005). Diversity-generating retroelements are a newly discovered family of genetic elements that confer selective advantages under certain conditions by introducing vast amounts of sequence diversity into target genes (Doulatov *et al.*, 2004).

Class II elements constitute the most frequent and repetitive TEs in prokaryotic genomes and ISs are their most abundant members. ISs are the simplest autonomous mobile genetic elements and range from 0.7 to 3.5 kbp. They carry one or two ORFs encoding the transposase, generally (but not always) flanked by short IRs (Mahillon & Chandler, 1998). Based on their genetic organization and their host range, one can distinguish distinct families of IS, which are widely spread among the prokaryotic world (Siguier *et al.*, 2006a, b; Touchon & Rocha, 2007). Composite transposons are elements flanked by IS elements, whereas noncomposite transposons are flanked only by IRs. Transposons may carry unrelated genes such as antibiotic resistance genes, catabolic genes, virulence determinants and eventually also a resol-

vase, depending on their mechanism of transposition (for a review, see Mahillon & Chandler, 1998).

TEs move from one genomic location to another by a process that is independent of DNA homology and that follows one of two pathways: conservative to form simple insertions (also called simple insertion, nonreplicative and cut-and-paste) and replicative to form cointegrates (also known as copy-and-paste). Transposition strategies vary according to the number and the nature of cuts that sever the TE from the flanking donor DNA: one-strand break or double-strand break and the possibility of accompanying TE replication. A double-strand break effectively removes all connections with the donor DNA and thus precludes the formation of a cointegrate. In contrast, a nick allows for the maintenance of connections to both donor and target replicons after integration, and, following replication, results in the formation of a cointegrate (for a general review, see Curcio & Derbyshire, 2003). On integration into a new target site, TEs generate a repeated sequence of the target DNA in both modes of transposition. Many elements transpose exclusively by the replicative pathway, promoting the formation of a cointegrate that can be resolved either by a transposon-encoded resolvase or by the homologous recombination functions of the host. Among the replicative TEs, one finds the members of the Tn3 (Hefron, 1983) and the IS6 families (Brown *et al.*, 1984).

Simple integrations do not include any replication of the element and result in the insertion of the integral transposon within a target molecule. This cut-and-paste mechanism is used by many elements of the families of Tn7 (Bainton *et al.*, 1991), IS3 (Sekine *et al.*, 1994), IS10 (Bender & Kleckner, 1986; Bolland & Kleckner, 1996) and IS50 (Berg, 1983; Lichens-Park & Syvanen, 1988). While it does not generate cointegrate transposition products, if transposition takes place at the moment of replication from a newly replicated region to a nonreplicated region, then transposition will effectively result in the duplication of the element in one of the newly formed chromosomes. The copy number of the element does not depend exclusively on the mode of transposition. TEs are generally tightly regulated and have evolved numerous shrewd strategies to exercise this control. These regulatory mechanisms act at various moments of the transposition process: transcription of the TPase gene; its translation; TPase stability and activity; and DNA binding and catalysis (for a review, see Nagy & Chandler, 2004).

Numerous TEs have been shown to carry out both of these mechanisms. For instance, bacteriophage Mu combines properties of a temperate phage and a TE. As a TE, Mu is remarkable because its life cycle involves both transposition modes: nonreplicative, resulting in lysogeny, and replicative, leading to multiple copying of phage DNA during the lytic growth (Mizuuchi, 1992). Several ISs have also been noted to generate cointegrates and simple insertions, with

different ratios according to the element [e.g. IS1 (Turlan & Chandler, 1995); IS903/IS102 (Bernardi & Bernardi, 1991)].

Temperate phages are another class of selfish elements that lead to the creation of repeats in genomes. While phages themselves carry few repeated elements, the presence of several phages of the same family in the chromosome effectively leads to large repeated elements. Closely related phages may not always integrate their DNA at precisely the same site in the chromosome. In *E. coli*, for example, phage  $\lambda$  DNA normally integrates at only one site, phage P2 DNA can integrate into at least 10 sites (Barreiro & Haggard-Ljungquist, 1992) and phage Mu DNA integrates essentially randomly into host DNA (for a review, see Pato, 1989). Some phages have immunity mechanisms preventing super-infection. Yet, many genomes contain multiple copies of recombining closely related prophages (Nakagawa *et al.*, 2003).

Creation of repeats by selfish systems may lead to highly repetitive genomes. In these cases, repeats are created by multiple insertions of genetic elements using site-specific integration by ways of a transposase, an integrase or a related protein. ISs, prophages and closely related elements are very diverse and widespread in the prokaryotic world. Only prokaryotes refractory to lateral transfer, such as *Buchnera*, systematically lack them.

### Interspersed repeats

Specific families of repeated DNA elements have been isolated from a large variety of eukaryotic and prokaryotic organisms, but their origin and exact function(s) are rarely known (Table 1). They are widespread in bacterial and archaeal chromosomes and plasmids. These repeats could play key roles in genomic organization, genetic evolution and genome plasticity. In prokaryotes, interspersed repeated sequences are typically intergenic, although some intragenic repeats are known (Ogata *et al.*, 2000), and often have the ability to form stable secondary structures in mRNAs (Cozzuto *et al.*, 2008).

The repetitive extragenic palindromic (REP) sequences, also called palindromic units, are one of the best-described classes of intergenic repeats (Higgins *et al.*, 1982; Stern *et al.*, 1984). The bacterial interspersed mosaic elements (BIME) are a mosaic combination of REP (Gilson *et al.*, 1991). All of these repetitive elements share common features such as a short size (< 500 bp), palindromic structure, tandem organization and multiple locations in intergenic regions. In general, they do not exhibit extensive sequence similarity with known TEs, such as ISs or transposons. Many questions concerning these interspersed sequences remain unanswered. Ignoring even the very mechanisms of creation, spread and conservation in genomes, we can only offer speculations about the reasons for their exclusive location

outside structural genes and their potential role in the organization of the chromosome. An interesting property of BIMEs/REP is their functional diversity. Because BIMEs are composite REP elements, it is difficult to disentangle which effects are associated specifically to each type of elements. They have been proposed to play a role in transcription termination (Gilson *et al.*, 1986; Espeli *et al.*, 2001), mRNA stabilization (Higgins *et al.*, 1988; Khemici & Carpousis, 2004), control of translation (Stern *et al.*, 1988) and genomic rearrangements (Shyamala *et al.*, 1990). They are binding sites for DNA polymerase I (Gilson *et al.*, 1990), for DNA gyrase (Yang & Ames, 1988; Espeli & Boccard, 1997) and for integration host factor (Boccard & Prentki, 1993; Oppenheim *et al.*, 1993), all of which play a key role in prokaryotic DNA physiology. These interactions suggest that BIMEs are involved in processes as important as DNA compaction, DNA supercoiling, gene expression and chromosome replication. It is, however, unclear whether these are side effects of BIMEs or whether they are the reason for their proliferation.

REP sequences have also been proposed to be 'selfish DNA' replicating and subsisting in genomes by gene conversion (Higgins *et al.*, 1988). Some evidence associates REP elements with genomic plasticity. REP sequences have been found at the recombination junctions of  $\lambda$  bio-transducing phages (Kumagai & Ikeda, 1991) and amplification of some plasmids might be initiated by REP-REP recombination (Kofoid *et al.*, 2003). In addition, some IS elements insert specifically within REP sequences (Clement *et al.*, 1999; Tobes & Pareja, 2006). The characterization of some REP elements as hot spots for recombination and transposition suggests they may have important roles in prokaryotic evolution. As noted above, BIMEs are composite REP elements; it is difficult to disentangle which effects are associated with one or the other types of elements.

Miniature transposable elements (MITE) are mobile elements typically measuring between 100 and 400 bp, carry long terminal IRs and are flanked by target site duplication of variable lengths (for a review, see Delihias, 2008). MITE are generally thought to derive from IS by internal deletion of a part or all of the transposase but retention of both ends. They are considered to be nonautonomous TEs that can be mobilized *in trans* by the full-length transposase of the parent transposon, when it is present (Jiang *et al.*, 2004). How these mobile elements passively transpose within the chromosome is still unknown. Several MITE families have been characterized in archaea (Brugger *et al.*, 2004), and four families have been well-described in bacteria. The repeat unit of *pneumococcus* (RUP) is 107 bp in length, is spread in c. 100 copies in the genome of *S. pneumoniae* and could be mobilized by an IS-encoded transposase (Oggioni & Claverys, 1999). *Neisseria* miniature insertion sequences (NEMIS; or Correia elements) are 108–158-bp-long repeats

**Table 1.** Families of interspersed repeats

Repeats	Acronyms	Features	References
Repetitive extragenic palindromic or palindromic units	REP or PU	21–65 bp Imperfect palindrome Extragenic sequence Potential stem-loop structure Probably transcribed	Higgins <i>et al.</i> (1982), Stern <i>et al.</i> (1984)
Bacterial interspersed mosaic elements	BIME	40–500 bp Mosaic combination of REP separated by other sequence motifs	Gilson <i>et al.</i> (1991)
Clustered regularly interspaced short palindromic repeats	CRISPR	Noncontiguous direct repeats (DR, 24–47 bp) separated by stretches of similarly sized unique spacers (26–72 bp) Potential stem-loop structure Probably transcribed	Ishino <i>et al.</i> (1987), Mojica <i>et al.</i> (2000)
Miniature inverted repeat transposable elements	MITE	100–400 bp Nonautonomous element (mobilizable <i>in trans</i> by full-length transposase) Probably derived from IS by internal deletion Flanked by inverted repeat (IR, 10–40 bp) Not encode protein Extra- and intragenic sequence Potential stem-loop structure Probably transcribed	Correia <i>et al.</i> (1988), Delihas (2008)
Intergenic repeat unit or enterobacterial repetitive intergenic consensus	IRU or ERIC	69–127 bp Large palindromic sequence Potential stem-loop structure Probably transcribed	Sharples & Lloyd (1990), Hulton <i>et al.</i> (1991)
Insertion sequence	IS	0.7–3.5 kbp Autonomous element Often flanked by inverted repeat (IR, 10–40 bp) Encode transposase (carry 1 or 2 ORFs) Conservative or replicative transposition	Mahillon & Chandler (1998)
Transposons	Tn	Autonomous element Code for transposase and a number of gene products (e.g. antibiotic resistance, virulence factor) <i>Composite (Class I)</i> Flanked by two IS (identical or different, direct or inverted e.g. Tn5, Tn7, Tn10) Conservative transposition	<i>Noncomposite (Class II)</i> Flanked by two IR (e.g. Tn3, Tn9) Replicative transposition Code for resolvase
Bacteriophage elements		Phage Mu temperate phage and transposable element Lysogeny growth → conservative transposition	Pato (1989) Lytic growth → replicative transposition

that make up 1–2% of the genome in pathogenic neisseriae (Correia *et al.*, 1988). In contrast to RUP elements, which are mostly interspersed with other repeated DNA sequences, NEMIS sequences are frequently located next to *Neisseria* genes and are transcribed into mRNAs. Hairpins formed by the pairing of NEMIS terminal IRs are targeted by RNase III, which results in post-transcriptional gene expression regulation (De Gregorio *et al.*, 2003a,b). Enterobacterial repetitive intergenic consensus elements (ERIC) (Hulton

*et al.*, 1991), also known as intergenic repeat units (Sharples & Lloyd, 1990), are also a class of MITE. These elements range in size from 69 to 127 nt, are moderately abundant (20–30 copies) in the genomes of several enterobacteria (Hulton *et al.*, 1991) and are present in hundreds of copies in the genomes of *Yersinia* (De Gregorio *et al.*, 2005). Similar to NEMIS sequences, they are frequently inserted immediately downstream from ORFs and are cotranscribed with upstream and downstream genes. Hairpins formed by ERIC

mRNA are targeted by RNase E, and this interaction destabilizes the upstream transcript (De Gregorio *et al.*, 2005). The *Yersinia* palindromic element sequences share properties with both ERIC and NEMIS sequences and represent another example of MITE involved in post-transcriptional control (De Gregorio *et al.*, 2006). Other elements resembling MITE have been identified, for example *Caulobacter* CerM-associated intergenic repeat (Chen & Shapiro, 2003), *Enterococcus faecalis* repeats (Venditti *et al.*, 2007) and *Rickettsia*-specific palindromic elements (Ogata *et al.*, 2000). It has recently been proposed that REP elements are also MITE, but no equivalent full-length IS has been observed (Siguier *et al.*, 2006a, b).

CRISPR are another class of well-characterized noncoding repetitive sequences. They are hypervariable genetic loci widely distributed in bacteria and archaea (Jansen *et al.*, 2002). One or more CRISPR loci are found in 40% of the bacterial genomes sequenced so far and in most archaea (Grissa *et al.*, 2007a, b). They are composed of direct repeats, repeated up to 250 times, ranging in size from 24 to 47 nt that are separated by similarly sized nonrepetitive spacers often adjacent to *cas* (CRISPR-associated) genes (Haft *et al.*, 2005; Sorek *et al.*, 2008). While the direct repeats are similar within a single locus, they may be different between CRISPR loci of the same genome. In the large majority of cases, the direct repeats are highly conserved while the spacers are very diverse within a given locus, even among strains of the same species. Previously considered to be a simple family of noncoding repetitive elements (Ishino *et al.*, 1987; Mojica *et al.*, 2000), recent studies have established that CRISPR provide acquired resistance against foreign DNA (Barrangou *et al.*, 2007; Horvath *et al.*, 2008) and allow microbial populations to survive phage predation (Kunin *et al.*, 2007). The spacers between repeats in CRISPR are highly similar to sequences of phages and allow, possibly by an RNA-interference-like mechanism (Brouns *et al.*, 2008), to avoid infection by phages containing these sequences. Based on the detection of CRISPR spacer transcripts, it was

suggested that this system involves binding of excised spacers to target mRNAs, which are subsequently degraded by nucleases or that inhibit translation (Makarova *et al.*, 2006). However, recent data suggest that phage inactivation occurs by interaction of the CRISPR mRNA with the phage DNA (Marraffini & Sontheimer, 2008). Some compelling evidence in support of these hypotheses is the demonstration that *Streptococcus thermophilus* responds to viral predation by integrating new spacers into its CRISPR loci derived from the infecting phage genome. Addition or removal of spacers by some unknown mechanism changes the phage-resistance phenotype of the cell, confirming spacer-specific resistance (Barrangou *et al.*, 2007). CRISPR loci play a critical role in the adaptation and persistence of a microbial host in a particular ecosystem where viruses are present. As a result, the phage sequences present in CRISPR provide a historical perspective of phage exposure and an insight into the coevolution of the phage and the host genomes.

## Identification of repeats in genomes

As previously described, prokaryotic genomes contain two main classes of repeats: (1) generic repeats arising from standard recombination processes and (2) specific repeats arising from particular systems, such as TEs or CRISPR. If the sequence of the repeat is known *a priori*, both experimental approaches and simple sequence similarity searches allow its identification and analysis in a given genome. Specialized databanks contain information of many of frequent repeats (see Table 2). However, with the increasing availability of sequence data comes the need to detect repeats of which one does not necessarily know the sequence. If one wants to find all repeated elements in the genome, to pinpoint amplifications or recombination hotspots, then it is strictly necessary to use *ab initio* repeat discovery tools on the genomic sequence. Once repeats are uncovered, it is necessary to identify and classify them to

**Table 2.** Publicly available databases of annotated repetitive elements in prokaryotes

Database	Description	URL	References
ACLAME	Mobile genetic elements	<a href="http://aclame.ulb.ac.be">http://aclame.ulb.ac.be</a>	Lepae <i>et al.</i> (2004)
CBS Genome Atlas	Generic repeats	<a href="http://www.cbs.dtu.dk/services/GenomeAtlas">http://www.cbs.dtu.dk/services/GenomeAtlas</a>	Hallin & Ussery (2004)
CRISPRdb	CRISPR repeats	<a href="http://crispr.u-psud.fr/crispr/CRISPRdatabase.php">http://crispr.u-psud.fr/crispr/CRISPRdatabase.php</a>	Grissa <i>et al.</i> (2007a, b)
IS-Finder	Insertion sequences	<a href="http://www-is.biotoul.fr">http://www-is.biotoul.fr</a>	Siguier <i>et al.</i> (2006a, b)
MICdb	Prokaryote microsatellites	<a href="http://www.cdfd.org.in/micas">http://www.cdfd.org.in/micas</a>	Sreenu <i>et al.</i> (2003)
Phage DB	Bacteriophages	<a href="http://phage.sdsu.edu/~rob/cgi-bin/phage.cgi">http://phage.sdsu.edu/~rob/cgi-bin/phage.cgi</a>	Website
ProphageDB	Prophages	<a href="http://ispc.weizmann.ac.il/prophagedb">http://ispc.weizmann.ac.il/prophagedb</a>	Srividhya <i>et al.</i> (2007)
ProtRepeatsDB	Amino acid repeats	<a href="http://bioinfo.icgeb.res.in/repeats">http://bioinfo.icgeb.res.in/repeats</a>	Kalita <i>et al.</i> (2006)
Tandem Repeats DB	Tandem repeats	<a href="http://minisatellites.u-psud.fr">http://minisatellites.u-psud.fr</a>	Le Fleche <i>et al.</i> (2001)
TRIPS	Tandem repeats in proteins	<a href="http://www.ncl-india.org/trips">http://www.ncl-india.org/trips</a>	Katti <i>et al.</i> (2000)
VNTRDB	Tandem repeats	<a href="http://vntr.csie.ntu.edu.tw">http://vntr.csie.ntu.edu.tw</a>	Chang <i>et al.</i> (2007)



associate their presence with some systematic functional or genetic neighborhood.

In 1983, one of the first publicly available programs to find exact repeats in DNA sequences, QREPEATS, was made available to molecular biologists (Martinez, 1983). Since then, increasingly efficient algorithms and heuristics have been developed for finding exact and degenerate repeats (see Table 3). Although progress has been made, the task of

*ab initio* repeat family detection in genomes remains a challenging computational problem due to the diverse characteristics of repeats, the difficulties of efficiently allowing for substitutions and gaps in alignments and for the need of using a statistical framework. Fortunately, there exists a diverse variety of publicly available programs adapted to the various niches of *ab initio* repeat discovery: generic repeat detection, simple sequence repeat (SSR) detection

**Table 3.** Computational tools for *ab initio* detection of tandem and interspersed internal repeats in DNA, protein sequences and protein structures

Program	Type	L*	S†	I‡	M§	Availability/URL	References
<b>DNA</b>							
ADPLOT	GR	1	•	•		Email: taneda@si.hirosaki-u.ac.jp	Taneda (2004)
CRISPRFINDER	SpR	1	•		•	http://crispr.u-psud.fr/Server/CRISPRfinder.php	Grissa <i>et al.</i> (2007a, b)
CRT	SpR	1	•		•	http://www.room220.com/crt	Bland <i>et al.</i> (2007)
EULERALIGN	GR	1	•	•	•	http://www.stat.psu.edu/~yuzhang	Zhang & Waterman (2003)
MREPATT	SSR	1	•		•	http://algggen.lsi.upc.es/cgi-bin/search/mrepatt	Roset <i>et al.</i> (2003)
MREPS	SSR	1	•		•	http://bioinfo.lifl.fr/mreps	Kolpakov <i>et al.</i> (2003)
PATTERN LOCATOR	SSR	1	•		•	http://www.cmbi.uga.edu/software.html	Mrazek & Xie (2006)
PHOBOS	SSR	1	•	•	•	http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm	NA
PILER	GR	1	•		•	http://www.drive5.com/piler	Edgar & Myers (2005)
REAS	SpR	0	•		•	Email: ReAS@genomics.org.cn	Li <i>et al.</i> (2005)
RECON	GR	0–1	•	•	•	http://selab.janelia.org/recon.html	Bao & Eddy (2002)
REPEATFINDER	GR	0–1	•		•	http://www.cbc.umd.edu/software/RepeatFinder	Volfovsky <i>et al.</i> (2001)
REPEATOIRE	GR	1	•	•	•	http://www.abi.snv.jussieu.fr/public/Repeatoire	Treangen <i>et al.</i> (2009)
REPEATSCOUT	GR	0–1	•	•	•	http://bix.ucsd.edu/repeatscout	Price <i>et al.</i> (2005)
REPET	SpR	1	•	•	•	http://urgi.versailles.inra.fr/development/repet	NA
REPSEEK	GR	1	•	•		http://www.abi.snv.jussieu.fr/public/RepSeek	Achaz <i>et al.</i> (2007)
REPUTER	GR	1	•	•		http://bibiserv.techfak.uni-bielefeld.de/reputer	Kurtz <i>et al.</i> (2001)
SPUTNIK	SSR	1	•	•	•	http://espressoftware.com/pages/sputnik.jsp	NA
SSRIT	SSR	1	•		•	http://finder.sourceforge.net	Temnykh <i>et al.</i> (2001)
STAR	SSR	1	•	•	•	http://atgc.lirmm.fr/star	Delgrange & Rivals (2004)
TRED	SSR	1	•	•	•	http://tandem.sci.brooklyn.cuny.edu/Tandem	Sokol <i>et al.</i> (2007)
TRF	SSR	1	•	•	•	http://tandem.bu.edu/trf/trf.html	Benson (1999)
<b>Protein</b>							
HHREPID	GR	2	•	•	•	http://toolkit.tuebingen.mpg.de/hhrepid	Biegert & Soding (2008)
IRF	GR	2	•	•	•	http://nihserver.mbi.ucla.edu/Repeats	Pellegrini <i>et al.</i> (1999)
RADAR	GR	2	•	•		http://www.ebi.ac.uk/Tools/Radar/index.html	Heger & Holm (2000)
REP	GR	2	•	•	•	http://www.embl-heidelberg.de/~andrade/papers/rep	Andrade <i>et al.</i> (2000)
REPPER	SSR	2	•			http://toolkit.tuebingen.mpg.de/repper	Gruber <i>et al.</i> (2005)
SIMPLE	SSR	1–2			•	http://www.har.mrc.ac.uk/research/bioinformatics/software/simple.html	Alba <i>et al.</i> (2002)
SSR	SSR	1–2			•	http://ftp.technion.ac.il/pub/supported/biotech	Klevytska <i>et al.</i> (2001)
TRUST	GR	2	•	•		http://ibivu.cs.vu.nl/programs/trustwww	Szklarczyk & Heringa (2004)
VMATCH	GR	1–2	•	•		http://www.vmatch.de	NA
XSTREAM	SSR	1–2	•	•	•	http://jimcooperlab.mcd.ucs.edu/xstream	Newman & Cooper (2007)
<b>3D</b>							
DAVROS	GR	3	•	•		http://http://www.ebi.ac.uk/~murray/DAVROS	Murray <i>et al.</i> (2004)
MATRAS	GR	3	•	•		http://biunit.aist-nara.ac.jp/matras	Kawabata (2003)
SWELFE	GR	1–3	•	•		http://bioserv.rpbs.jussieu.fr/swelfe	Abraham <i>et al.</i> (2008)

\*L, level; level 0, unassembled sequence reads; level 1, assembled genomic regions; level 2, protein sequences, level 3, 3D protein structures.

†S, allows for substitutions.

‡I, allows for indels (gaps).

§M, outputs *multicopy* repeats.

GR, generic repeat detection.

and specific repeat detection. While the focus of this review is on repeat detection at the genome level, if one is interested in finding the functional consequences of intra-genic DNA repeats, it is also necessary to analyze their impact on protein sequence and structure. SWELFE (Abraham *et al.*, 2008) is to our knowledge the only existing program able to detect and link repeats in genes, proteins and protein structures.

Generic repeat detection tools, as the name implies, are those that do not have any predisposition for any single repeat class. One example is the REPUTER program, which was originally designed to search for all pairs of maximal *exact* repeats in genomes, but has since been modified to allow for substitutions and gaps through a banded dynamic programming seed extension algorithm (Kurtz *et al.*, 2001). Given a sequence, or a structure, most of these tools report all the similar internal repeats larger than some minimum size or score threshold. Unfortunately, many programs provide no statistical basis to evaluate the results. Providing such *P*-values is mathematically complicated (Karlin & Ost, 1985; Waterman & Vingron, 1994), and some programs use statistical methods appropriated to different ways of searching for repeats. REPSEEK has different statistical procedures for nondegenerate and for degenerate repeats and also accounts for the nucleotide composition of genomes when searching for repeats (Achaz *et al.*, 2007). The latter factor is important because A+T (G+C)-rich repeats are more abundant in A+T (G+C)-rich genomes by purely stochastic reasons and that should be accounted for.

An important distinguishing characteristic of these programs is repeat family construction. Many programs only produce pairs of repeats and the majority of programs that detect multicopy repeat families while allowing for gaps [such as RECON (Bao & Eddy, 2002)] merge pairwise repeats into families by requiring all pairs to meet or exceed some similarity threshold. These methods require an additional alignment step to build a consensus sequence or a multiple alignment of the repeat family. This process can be lengthy and result in poor accuracy due to inconsistent homology relationships among the pairs. A few existing programs [EULERALIGN (Zhang & Waterman, 2003), REPEATOIRE (Treangen *et al.*, 2009), REPEATSCOUT (Price *et al.*, 2005)] construct directly the repeat families via local multiple alignment. Because of the inherent time complexity of local multiple alignment, heuristic approaches are required, but these programs have shown good accuracy and efficiency in a wide variety of settings. Many of the generic repeat detection programs form the basis for more specific repeat detection programs. Because of their design, generic programs are best fit for discovery of unknown repeats or detecting the full repertoire of repetitive elements in a given genome.

When searching for a specific type of known repeat in new genomes, specific repeat detection tools should take

precedence over generic ones. The PILER (Edgar & Myers, 2005) suite of repeat detection programs offers detection of four specific classes of repeats: dispersed repeats (PILER-DF), tandem repeats (PILER-TA), pseudo-satellites (PILER-PS) and terminal repeats (PILER-TR). PILER-CR (Edgar, 2007), CRISPR Recognition Tool (CRT) (Bland *et al.*, 2007), and the CRISPR-FINDER (Grissa *et al.*, 2007a,b) website were all designed to identify CRISPR elements in genomes. ISCAN is a specific repeat detection (SpR) program for the identification of ISSs, but it is not really an *ab initio* repeat discovery tool because it requires the use of a pre-existing IS database and is thus unable to find nonhomologous IS elements (Wagner *et al.*, 2007). REAS was designed to search for TEs from unassembled sequence reads (Li *et al.*, 2005). Of note, while REAS was originally designed as a specific repeat detection program, a recent survey on *ab initio* repeat detection programs found REAS to be the most effective tool for analyzing repeats in unassembled sequence reads (Saha *et al.*, 2008).

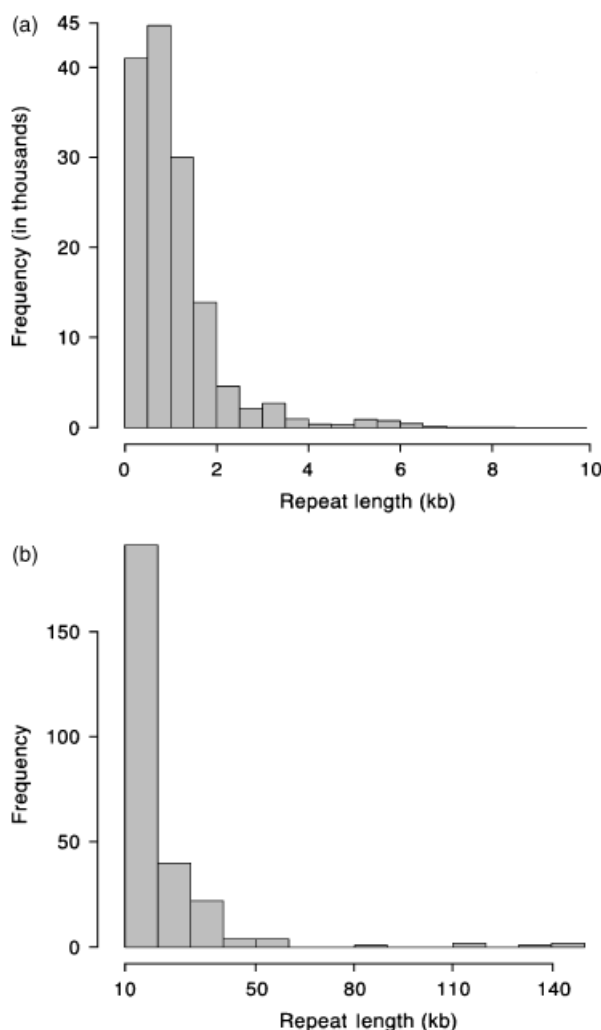
SSRs are tandem repeats of small motifs. The available programs for SSR detection can be classified by their allowance for substitutions and/or indels. For example, MREPS finds repeats of size larger than  $k+9$  independently of genome size, where  $k$  is the size of the oligonucleotide motifs, for example 12 bp for trinucleotide motifs (Kolpakov *et al.*, 2003). TANDEM REPEAT FINDER (TRF) (Benson, 1999) is arguably the most commonly used program for SSR detection, but recent comparisons of SSR programs concluded that SPUTNIK was the most sensitive method for SSR detection of perfect and imperfect microsatellites (Leclercq *et al.*, 2007; Merkel & Gemmell, 2008). In prokaryotes, active SSR typically lack mismatches, which makes the problem easier. However, SSRs also tend to be of smaller size than in eukaryotes, which demands for accurate statistical criteria for the minimal size of SSR given a genome size and composition. Most SSR-finding programs lack such statistical analyses. Mrazek *et al.* (2007) have proposed a series of statistical models implemented in software that also searches for SSR.

In summary, there exist a large variety of publicly available programs adapted to the various niches of *ab initio* repeat discovery: generic repeat detection, SSR detection and specific repeat detection. We have presented and discussed a representative collection of commonly used tools for each of these tasks. Unfortunately, few studies are available to compare the accuracy of the different programs. Further progress will inevitably be required to better capture diverse characteristics of repeats and to allow for a statistical framework to improve the sensitivity and accuracy of repeat detection tools, especially when considering highly divergent sequences. In the next section we will apply an *ab initio* repeat detection tool to take an in-depth look at the repeat landscape in nearly 700 bacterial and archaeal genomes.

## The abundance of repeats in genomes

Several works have previously described the amount of repeats found in genomes (Rocha *et al.*, 1999; Bayliss *et al.*, 2001; Metzgar *et al.*, 2001; Achaz *et al.*, 2002, 2003; Frank *et al.*, 2002; Hancock, 2002; Aras *et al.*, 2003; Rocha, 2003a, b; Mrazek *et al.*, 2007). Given the exponential growth of the databanks, this is an exercise permanently lagging behind the data. While a large analysis of SSR in genomes has recently been published (Mrazek *et al.*, 2007), there is no equivalent recent survey for large repeats. We therefore illustrate the literature on the subject with a genome-wide survey for repetitive DNA. We first used general measures of genome repetitiveness in terms of long repeats and SSR. We then extended the analysis to the identification of long repeats in 659 prokaryotic genomes, for a total of 720 chromosomes. For these illustrative purposes, we searched for all large (> 300 nt, extended from 25 nt seed matches of strict identity), interspersed, non-overlapping repeats using REPEATTOIRE, an *ab initio* tool for detecting degenerate repeats in DNA sequences. In total, we found 144 118 repeats classified into 56 196 repeat families (c. 220 repeats and 85 repeat families per genome) (Fig. 2). If one searches for smaller repeats *a priori* still capable of engaging into homologous recombination in *E. coli* (with > 25 nt of a core of contiguous strict identity), we find 324 096 repeats contained in 111 268 repeat families. The majority of the > 300 nt repeats are part of multicopy repeat families (60%), and the multiplicity of the repeat families ranged from 2 to 230 but contained on average approximately three repeat copies per repeat family. The mean length of repeats > 300 nt is 1135 nt (640 nt including repeats > 25 nt). While nearly all of the repeats (95%) are < 3000 nt, the few repeats > 3000 nt account for c. 23% of the total length of all repeats.

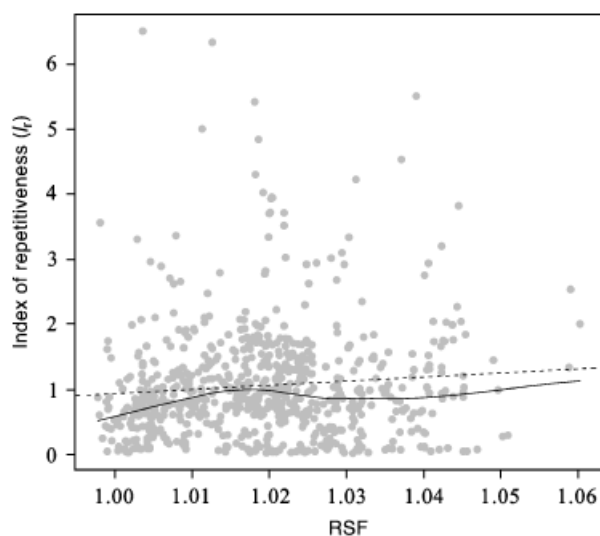
Several measures of genome repetitiveness have been proposed previously. These include the index of repetitiveness ( $I_r$ ) (Haubold & Wiehe, 2006), which measures the extent to which subsequences of the genome are repeated elsewhere, and the relative simplicity factor (RSF) (Hancock, 2002), which measures the degree of accumulation of SSR. Both measures vary widely in genomes (Fig. 3), but they are only very weakly correlated ( $r=0.015$ ,  $P=0.02$ ). Hence, genomes with many long repeats do not necessarily have many SSR, and vice versa. This is hardly surprising, as the former will recombine mostly by homologous recombination and the latter by slipped-strand mispairing. They must therefore be considered separately. The variable  $I_r$  is biased toward large-scale duplications. For example, the most repetitive genome found using  $I_r$  is *Methylobacillus flagellatus* KT ( $I_r=6.51$ ) because it contains the largest repeat found in our analysis, a 143-kb two-copy repeat. The two copies of the repeat are identical, suggesting a very



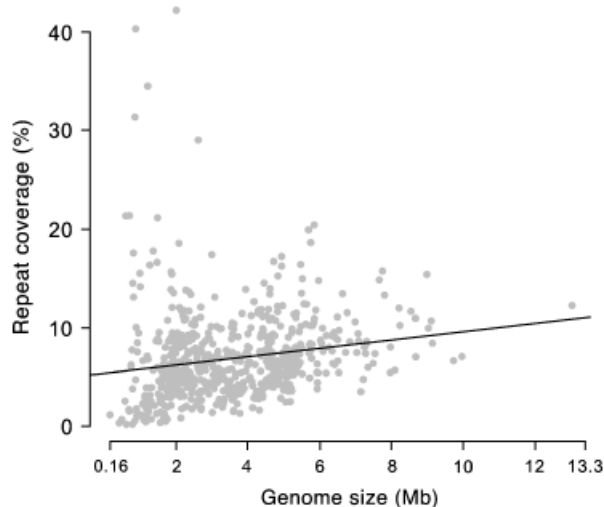
**Fig. 2.** Distribution of the size of 144 118 repeats families (containing repeats > 300 nt) found in 720 chromosomes. (a) Over 99% of the repeats detected were < 10 kb, and more than 80% were 2 kb or shorter. (b) There is evidence of large segmental duplications because 267 repeat families spanned 10 kb or larger, varying in size from 10 to 143 kb.

recent duplication event. This large segmental duplication accounts for 95% of the overall repeat density of the genome and spans 288 genes (Chistoserdova *et al.*, 2007). This shows the limits of simple indexes of repetitiveness. To understand why some genomes contain large number of repeats whereas others lack them almost entirely, one has to precisely identify and class repeats in families.

For our analysis, we computed genome repetitiveness as the total length of non-overlapping repeats divided by the genome size (Achaz *et al.*, 2002), also referred as repeat coverage. We found a very weak positive correlation between repeat coverage and genome size (Spearman's  $\rho=0.31$ ,  $P=0.03$ ,  $R^2=0.03$ ) (Fig. 4). However, we did find a strong correlation between the number of repeat families in a



**Fig. 3.** Lack of significant association between  $I_r$  (Haubold & Wiehe, 2006) and RSF (Hancock, 2002). We compared two indices of repetitiveness based on large repeats ( $I_r$ ) and SSRs. The association between the two is significant but very weak (Pearson's coefficient = 0.015,  $P = 0.02$ ). The solid line represents a spline interpolation showing that the weak correlation is consistent throughout the plot. The dashed regression line shows that the association of the two indices is slightly positive ( $R^2 = 0.006$ ).



**Fig. 4.** Association between repeat coverage and genome size. We found a positive but very weak correlation (Spearman's  $\rho = 0.31$ ,  $P = 0.03$ ,  $R^2 = 0.03$ ).

genome and its size (Spearman's  $\rho = 0.81$ ,  $P < 2.2 \times 10^{-16}$ ;  $R^2 = 0.65$ ), suggesting that although a larger genome implies more repeat families and repeats, it does not necessarily imply increased relative repeat coverage. The mean coverage among prokaryotes is 6.9% (q.d. 2.2%). *Orientia tsutsugamushi* stands out as the most repetitive genome, containing 42.2% of its 2.0-Mb chromosome covered by repeats. This

**Table 4.** Top 10 genomes with highest repeat coverage

Rank	Genome	Type	Size (Mb)	Coverage (%)
1	<i>Orientia tsutsugamushi</i> str. Ikeda	Circular	2.0	42.2
2	<i>Candidatus Phytoplasma australiense</i>	Circular	0.9	40.3
3	<i>Mycoplasma mycoides mycoides</i> SC	Circular	1.2	35.5
4	<i>Onion yellows phytoplasma</i> <i>Phytoplasma asteris</i>	Circular	0.9	31.4
5	<i>Bartonella tribocorum</i> CIP 105476	Circular	2.6	29.0
6	<i>Candidatus Phytoplasma mali</i>	Linear	0.6	21.4
7	<i>Aster yellows witches' broom</i> <i>phytoplasma</i> AYWB	Circular	0.7	21.4
8	<i>Wolbachia pipientis</i>	Circular	1.6	21.2
9	<i>Microcystis aeruginosa</i> NIES-843	Circular	5.8	20.4
10	<i>Photorhabdus luminescens</i> <i>laumondii</i> TTO1	Circular	5.7	20.0

Repeat coverage was calculated by summing the lengths of all of the repeats in the genome divided by the genome size. *Orientia tsutsugamushi* has the highest coverage (42.2%) of all genomes analyzed. Four of the top ten genomes are *Phytoplasma* (shaded in gray).

astonishing repeat density in *O. tsutsugamushi* results from over 1000 copies of 770-bp (mean length) repeats found in the genome. These include 359 tra genes for conjugative type IV secretion systems, over 400 transposases, 60 phage integrases and 70 reverse transcriptases (Cho *et al.*, 2007). Also worth noting is that the eight most repetitive genomes are small sized and of obligatory symbionts (Table 4), including four species of *Phytoplasma*.

*Microcystis aeruginosa* NIES-843 (Frangeul *et al.*, 2008) and *Trichodesmium erythraeum* IMS101 also contain > 10 times more repeats than the average genome (> 1000 repeats each). These two cyanobacteria contain large numbers of TEs. *Bordetella pertussis* genome, a strict human pathogen that has endured an expansion of TEs in its 4-Mb genome (Parkhill *et al.*, 2003), contains the repeat family with the highest copy number with 230 copies of a 407-bp repeat. Five genomes – *E. coli* 0157:H7, *E. coli* K-12 DH10B, *Mycobacterium flagellatus*, *Mycobacterium smegmatis* MC2 155 and *Streptomyces griseus griseus* NBRC 13350 – contain repeats spanning > 50 kb. We also investigated which genomes were devoid of large repeats or were severely lacking them (Table 5). The four genomes found to be completely devoid of repeats were all obligate endosymbiotic bacteria and nine of the 10 genomes with lower repeat coverage were endosymbionts. *Prochlorococcus marinus* MIT 9312 was the lone exception.

When using RSF to gauge genome repetitiveness based on SSR, *Candidatus Phytoplasma australiense* is the most repetitive (RSF = 1.0602) and *Nocardioideis* sp. JS614 is the least repetitive (RSF = 0.9978). *Phytoplasma australiense* is also one of the most repetitive genomes with a coverage by large repeats of 40.2%. This genome houses potential mobile

**Table 5.** Top 10 genomes with lowest repeat coverage

Rank	Genome	Type	Size (Mb)	Coverage (%)
1	<i>Buchnera aphidicola</i> Sg	Circular	0.6	0.0
1	<i>Buchnera</i> sp. APS	Circular	0.6	0.0
1	Candidatus <i>Sulcia muelleri</i> GWSS	Circular	0.6	0.0
1	Candidatus <i>Blochmannia floridanus</i>	Circular	0.7	0.0
5	<i>Buchnera aphidicola</i> Baizongia pistaciae	Circular	0.6	0.2
6	Candidatus <i>Blochmannia pennsylvanicus</i> BPEN	Circular	0.8	0.2
7	<i>Buchnera aphidicola</i> str. Cc ( <i>Cinara cedri</i> )	Circular	0.4	0.3
8	<i>Polynucleobacter necessarius</i> STIR1	Circular	1.6	0.4
9	Candidatus <i>Ruthia magnifica</i> str. Cm	Circular	1.1	0.5
10	<i>Prochlorococcus marinus</i> MIT 9312	Circular	1.7	0.5

Repeat coverage was calculated by summing the lengths of all of the repeats in the genome divided by the genome size. We found 4 of the 659 genomes to have no significant repeats > 300 nt. Of the genomes listed, all are intracellular symbionts except for *Prochlorococcus marinus* MIT 9312.

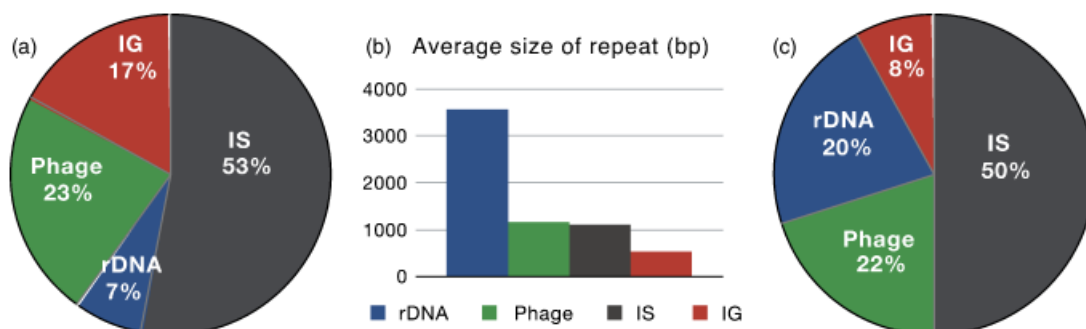
units containing clusters of DNA repeats that account for 12.1% of the genome (Tran-Nguyen *et al.*, 2008). However, *Nocardioideis* sp. JS614 has an average repeat coverage (7%), suggesting once again the lack of a correlation between various measures to gauge genome repetitiveness.

In prokaryotes, individual SSR are small, compared with eukaryotic microsatellites, and rarely occupy > 100 nt. Most studies on SSR focused on the identification of SSR of small motifs (1–4 nt), which are more abundant in small genomes. In the current databases, these often correspond to host-adapted bacteria. SSR of larger motifs (5–11) are more frequent in larger genomes, typically free-living prokaryotes (Mrazek *et al.*, 2007). There is no significant correlation between the numbers of the two types of SSR among genomes, suggesting that they occur independently. While SSR of shorter motifs have been found to be less stable, it is unclear whether this difference between genomes corresponds to different evolutionary strategies, recombination mechanisms or both (or none). The most recent global assessment shows that some clades are prone to having more and larger SSR (Mrazek *et al.*, 2007). Among the very small genomes (< 1 Mb), the phytoplasmas and the mycoplasmas show large numbers of SSR and large repeats. Inversely, obligatory endosymbionts lack SSR as well as large repeats. *Firmicutes*, excluding *Mollicutes*, also show few SSR. Most of the genomes with many SSR of large motifs belong to the *Proteobacteria* or to the *Cyanobacteria* (Mrazek *et al.*, 2007). Among the former, the number of SSR in genomes can vary by orders of magnitude, with maxima among *Burkholderia mallei* and *Burkholderia pseudomallei* (Holden *et al.*, 2004; Nierman *et al.*, 2004). In general, SSR of motifs that are not multiples of three nucleotides are under-represented inside

coding sequences most likely because recombination leads to frameshifts and thus gene inactivation (Field & Wills, 1998; Ackermann & Chao, 2006). Some genomes have SSR elements of particularly large motifs. *Ehrlichia ruminantium* has one of the lowest coding densities among prokaryotic genomes partly because of a highly active process of contraction and expansion of 150 nt tandem motifs in intergenic regions (Frutos *et al.*, 2006). In some genomes, one finds lipoproteins that are composed almost exclusively of large tandems of amino acid motifs (Bhugra & Dybvig, 1992).

Structured repeats, i.e. repeats that can form stable RNA structures when transcribed, are most frequently intergenic. Larger generic repeats can be as large as to implicate several genes. In our survey, 43% of the repeats contained one or more genes, 39% of the repeats were located inside a coding sequence and 9% partially overlapped one coding sequence and an intergenic sequence. The remaining 9% of the repeats were intergenic. There is an abundant literature on genes often found in multiple copies in genomes such as IS elements (Touchon & Rocha, 2007), phages (Brussow *et al.*, 2004) and rRNA gene operons (Acinas *et al.*, 2004). Indeed IS, phage and rRNA gene represent 46% of the repetitive data in the 659 genomes (see Fig. 5 for the relative contribution of each category). ISs are the single most abundant category including half of all the classified repeats.

The distance between two copies of a repeat, the spacer, can provide further insight into the causes and the consequences of repetitive DNA. For example, an average spacer distance close to 0 is indicative of segmental duplications and/or gene amplification, while larger spacer distances may indicate an abundance of interspersed repeat elements (e.g. IS). For circular chromosomes, the largest spacer distance between two repeats is  $L/2$  ( $\leq 50\%$  of the chromosome length), where  $L$  is the chromosome length. For linear chromosomes, the maximum distance is  $L$ . For circular chromosomes, we found that the average spacer distance between two adjacent repeats in a repeat family is 13.8% of the chromosome length. In linear chromosomes, on average, repeats were separated by 21% of the chromosome length. When the repeat size increases for repeats in circular chromosomes, the spacer distance rapidly decreases to 0 (data not shown). This strongly suggests that large amplifications, which are unstable and therefore recent, are created in tandem and may be separated by rearrangements while they diverge (Achaz *et al.*, 2000, 2002). The opposite effect is observed in linear chromosomes; spacer size increases with the repeat size because the edges of such chromosomes are often unstable and prone to recombination involving the two chromosome arms (Volff & Altenbuchner, 1998). It has been suggested that small genomes use a more intense intragenomic recombination between close repeat sequences to modulate gene content (Aras *et al.*, 2003). Yet, this is not



**Fig. 5.** Identification of repeats classed as IS, phage, rRNA gene and intergenic regions (IG). Fifty percent of the identified repeats could not be confidently classified and thus are absent from the graph. (a) Repeats frequency; (b) average size; (c) repeat coverage. Repeat coverage is the fraction of the summed lengths of all repeats in a given category. We were able to class c. 50% of the 144 118 repeats into the four categories by mining the annotations for rRNA gene and IS, followed by a BLASTX query of the repeats against ISFINDER (Siguier *et al.*, 2006a, b) and BLASTN queries against the predicted phage regions by PHAGE\_FINDER (Fouts, 2006).

found either among close repeats capable of engaging in illegitimate recombination (Rocha, 2003a) or among our current assessments of long repeats. Indeed, larger genomes (> 6 Mb) have more repeats than smaller ones (< 2 Mb), but also a larger fraction of large repeats with spacers < 2 kb (13.6% and 9.8% for repeats > 300 nt at < 2 kb, and 11.1% and 6.8% for repeats > 25 nt at < 2 kb).

In summary, the global repeat content in nearly 700 prokaryote genomes is rich and varied. There are various measures for classifying a given genome's repetitiveness (index of repetitiveness, RSF and repeat coverage) but inherently each provides a different perspective of the repeat landscape. When defining repetitiveness as repeat coverage and focusing on large repeats (> 300 nt), we found prokaryote genome repetitiveness to range from 0% (no repeats) to over 42% (nearly half of the genome covered by repeats). On taking a closer look at both ends of the repetitiveness spectrum in these genomes, we found that obligate endomutualistic bacteria (such as *Buchnera* spp.) are typically the genomes with the fewest repeats, while obligate endoparasitic bacteria (such as *Phytoplasma* spp.) were found to contain the highest density of repeats. Both have undergone reductive genome evolution (Shigenobu *et al.*, 2000; Gil *et al.*, 2003; Oshima *et al.*, 2004). However, pathogens typically require repeated elements to generate genetic variability and their horizontal transmission implicates susceptibility to phages and plasmids. Vertically inherited endomutualists such as *Buchnera* have little need to generate genetic variability and are protected from parasites. Additional investigation will be required to exhaustively class repeats into families to further elucidate why some genomes contain a large number of repeats and why others are almost entirely devoid of repeats.

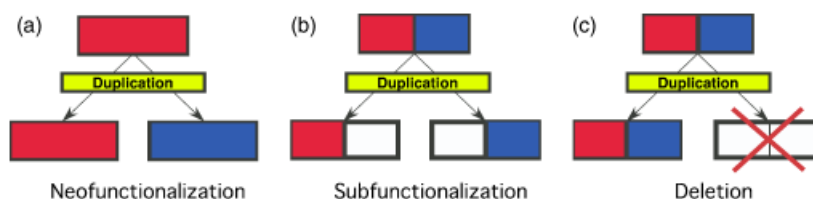
The entire repertoire of the repeats found in the 659 prokaryote genomes used in this study can be found at <http://www.wabi.snv.jussieu.fr/public/Repeatoire>.

## The functional consequences of repeats

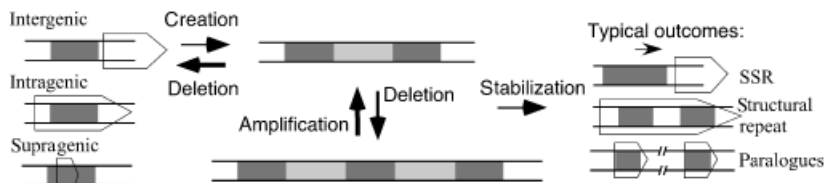
### Gene genesis by duplication and divergence

Intragenomic gene duplications are commonly seen as the bread and butter of gene creation. The genesis of a functional sequence *ex nihilo* seems an unlikely event, although it must have occurred at some time point in early evolution, and therefore most evolutionary novelty is thought to arise from the duplication and divergence of pre-existing elements (Ohno, 1970; Taylor & Raes, 2004). Gene duplication creates redundancy, thus freeing the copies from some functional constraints. If the copies are effectively redundant, then one of them is likely to be lost by lack of selection against inactivating mutations. However, in rare cases, the accumulation of mutations may lead either to neofunctionalization, one of the copies acquires a new function, or subfunctionalization, the two copies specialize in subfunctions of the primitive gene (Fig. 6) (Lynch & Katju, 2004). Both processes are expected to rely heavily on the accumulation of point mutations. This is a slow process that rarely allows enough time for adaptive evolution before the gene is inactivated in some way. On the other hand, if there is strong selection for both copies at the time of amplification, then the amplification will be stable and mutational processes could have much more time to diversify the function of the genes (Francino, 2005; Bergthorsson *et al.*, 2007). For example, if a gene confers a secondary function with weak efficiency and if the need for that function increases, then amplification of the gene results in an excess of genes for the primary function, but on just about what is needed to perform the secondary function. In this case, there is selection for the amplification because the secondary function is adaptive and required in amounts larger than can be provided by a single copy of the original gene. The amplified gene accumulates mutations in the different copies. Mutations in

**Fig. 6.** Fates of duplicate genes after a duplication event: (a) neofunctionalization, (b) subfunctionalization and (c) deletion. Red and blue represent different functions, white represents no function.



**Fig. 7.** The dynamics of repeats. Repeats are created at low rates, but once created, they may amplify or delete at high rates. Amplification results in the creation of very large tandem repeats. After stabilization, if it occurs, repeats are often found as SSR in intergenic regions, structural repeats in proteins or as paralogous genes.



one of the copies of the gene allowing an increase in efficiency of the secondary function become highly adaptive. Once such a mutation is acquired, either all other copies of the gene will be deleted, if the adaptive change has not affected the primary function, or one other copy will remain, if adaptation in this gene led to lower efficiency of the primary function. In this model, initial selection for amplification of secondary functions may lead to gene duplication by subfunctionalization (Francino, 2005). This might be a frequent event because recent work has shown that 20% of auxotrophs could be rescued by overexpression of noncognate genes (Patrick *et al.*, 2007).

While transient amplifications are frequent in both prokaryotes and eukaryotes (see Transient amplifications and amplicons), it is unclear whether gene duplication has a lasting contribution to the creation of novel functions. First, gene duplication in prokaryotes competes with horizontal transfer for the creation of novel functions. The latter is a very quick avenue for acquiring functions that have already endured selection in other contexts and there is now ample evidence that it brings new genes into most genomes at very high rates (Hao & Golding, 2006). Second, there is evidence for a deletion over amplification bias in prokaryotic genomes (see The fate of repeats) (Lawrence *et al.*, 2001; Mira *et al.*, 2001). Once repeats are created by recombination, typically in tandem, they are then deleted at high rates (Fig. 7). Third, gene conversion will frequently homogenize the two copies of the gene, thereby reducing sequence divergence and slowing subfunctionalization and neofunctionalization (Santoyo & Romero, 2005). Fourth, the architecture of prokaryotic genomes is more conserved than that of eukaryotes, i.e. large rearrangements tend to be highly deleterious (Rocha, 2008). Repeats can be stabilized if, after creation in tandem, they are physically separated by chromosome rearrangements (Achaz *et al.*, 2000). The higher stability of prokaryotic genomes implicates that the probability of repeat stabilization by a rearrangement physically separating the two copies is lower than in eukaryotes.

Finally, tandem amplifications of single genes in prokaryotes are complicated by the high coding density of the genomes and by the widespread existence of operons, which place a set of genes under the action of the same regulatory sequences. In short, there is ample evidence for gene amplifications in genomes, demonstrated by the abundance of repeated elements and some large families of paralogues (Snel *et al.*, 2002; Gevers *et al.*, 2004; Alm *et al.*, 2006; Cho *et al.*, 2007), especially among the largest and most complex genomes (Goldman *et al.*, 2006; McLeod *et al.*, 2006). However, it is unclear whether gene amplifications last long enough to be responsible for the creation of these large gene families.

Few studies have attempted to discern the relative roles of horizontal gene transfer and intrachromosomal gene duplication in creating gene families. Snel *et al.* (2002) analyzed gene repertoires of 21 prokaryotes and inferred the origin of groups of homologues (including paralogues) with respect to the phylogeny. They estimated that at most 20% of paralogues arose by horizontal transfer and 65% by gene duplication. A subsequent broader study estimated that gene duplication contributed to the creation of paralogues at least twice as frequently as horizontal transfer (Kunin & Ouzounis, 2003). An analysis of 87 866 paralogues in 106 prokaryotic genomes also suggested that most of the paralogues were created by intrachromosomal gene duplication (Gevers *et al.*, 2004). It should be noted that most comparisons in these studies concerned distant genomes and that two genes were regarded as paralogues if they shared some low, but significant, sequence similarity. In fact, the majority of these paralogues are not repeats in the sense that they have diverged excessively to engage recombination processes. The long distance between the genomes makes single-gene phylogenetic reconstruction hazardous. When the genes of the core genome were used to define a phylogeny among the 13 *Gammaproteobacteria* species, they provided a reference phylogeny against which to compare whether new paralogues arose by duplication or



by horizontal transfer (Lerat *et al.*, 2005). These results contradicted the previous ones, suggesting that gene amplification rarely leads to stable paralogy. Further investigations are required to disentangle this controversy, but the importance of gene duplication in the creation of new genes cannot yet be taken for granted in prokaryotes.

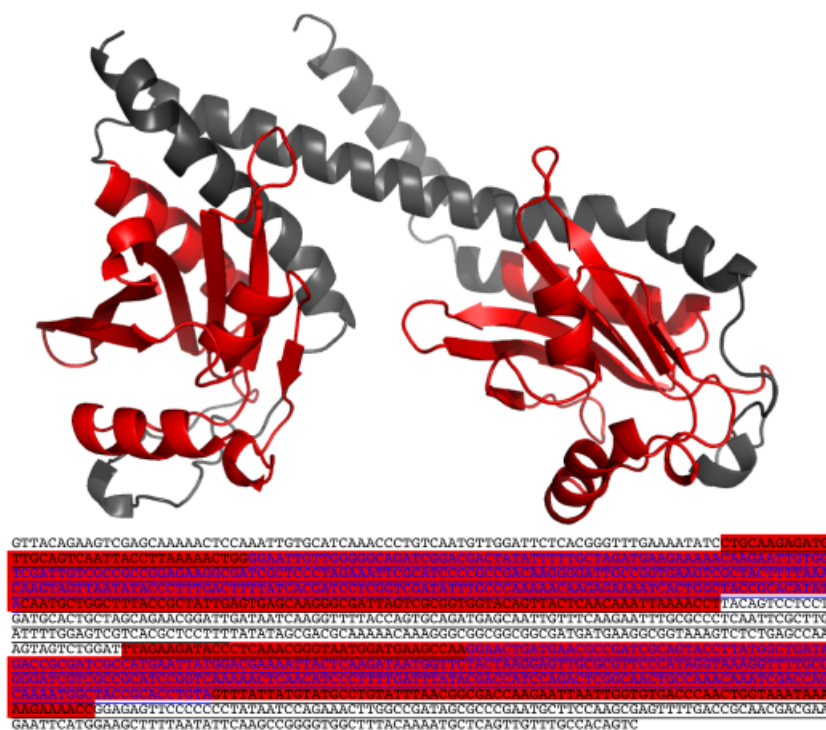
### Intragenic repeats and the evolution of protein function

As we have shown, many repeats can be found in genes and a significant number of them correspond to intragenic amplifications of genetic material. The nucleotide sequence of genes contains the information necessary to synthesize the amino acid sequences of proteins, which determines the protein folding. Several different codons are translated to the same amino acid and many different amino acid sequences code for similar protein structures. As a result protein structures evolve slower than protein sequences, and DNA sequences evolve faster than proteins. The direct influence of DNA sequences on protein sequences and of these on protein structures establish a causality link between intragenic DNA repeats and their functional consequences (Fig. 8). Therefore, intragenic amplification will have consequences at the three levels and will often have deleterious consequences. The creation of the repeat can lead to a change in the reading frame, in which case the amino acid sequence will change; the structure of the protein will be modified and the resulting protein will not be functional.

In fact, intragenic repeats are rarely observed in different coding frames. If the reading frame is conserved, the repeat can still modify or suppress the protein function, for example by leading to misfolding.

In spite of these constraints, > 14% of proteins contained repeats in an analysis of the 1998 Swissprot database (Marcotte *et al.*, 1998). Repeats are present both in prokaryotic and eukaryotic proteins, but in eukaryotic proteins are three times more likely to contain repeats, with archaeal proteins having more repeats than bacterial ones (Marcotte *et al.*, 1998; Lavorgna *et al.*, 2001). This could result from the larger size of eukaryotic proteins, because the longer proteins have more repeats. Yet, even among eukaryotes, multicellular organisms have more repeats in their proteins than unicellular organisms (Lavorgna *et al.*, 2001). The comparison of sequences between eukaryotic and prokaryotic proteins shows a small overlap (4%), and most of the conserved fragments code for ATP-binding cassettes (Marcotte *et al.*, 1998). This suggests that either the majority of repeats emerged after the divergence between prokaryotes and eukaryotes or they have diverged beyond recognition. As expected, there is a positive correlation between the number of repeats in genomes and in proteomes. Some bacteria such as *Aquifex aeolicus* or *Thermotoga maritima* show evidence of massive gene exchange from archaea and this could explain why their proteins have more repeats than the ones of other bacteria (Lavorgna *et al.*, 2001).

Most studies have analyzed intragenic repeats from the point of view of protein domains. Protein domains can be



**Fig. 8.** A protein containing a repeat in the DNA sequence, the amino acid sequence and the protein structure (PDB: 1 ykd, chain A, RefSeq: NP\_485944, GAF domain of an adenyl cyclase from the cyanobacteria *Anabaena* sp.). The structural repeat is shown in red, the amino acid repeat is in black and underlined and the nucleotide repeat is shown in blue and underlined. The amino acid repeat found is longer than the three-dimensional (3D) repeat, and the 3D repeat is longer than the nucleotide repeat.



defined in several ways and are usually relatively large (> 100 amino acids). A structural domain is defined as an independently folding unit, and an evolutionary domain as an independently evolving unit. These two definitions often result in the same domain families (Elofsson & Sonnhammer, 1999). Depending on the estimates, between 40% and 70% of prokaryotic proteins are composed of several domains (Liu & Rost, 2004; Ekman *et al.*, 2005), and domain duplication is one of the main sources of new domains (Vogel *et al.*, 2004). More protein-coding genes in a given genome result in a higher frequency of duplicated domains (Bornberg-Bauer *et al.*, 2005): according to some estimations, > 58% of *Mycoplasma genitalium* domains (Teichmann *et al.*, 1998), c. 68% in *Helicobacter pylori* and c. 85% in *E. coli* are duplicated (Muller *et al.*, 2002). However, the duplicated domains are rarely in the same protein. In prokaryotes, the proteins with more repeated domains are from the thiolase-like superfamily (found in proteins of degradation pathways such as fatty acid  $\beta$ -oxidation) and has only two repeats on average (Muller *et al.*, 2002).

Some proteins contain multiple repeated elements in tandem (from 20 to 50 amino acids depending on the type of the repeat) (Andrade *et al.*, 2001). Families of repeats present among prokaryotes include tetratrico-peptide repeats (Das *et al.*, 1998), ankyrin repeats (Sedgwick & Smerdon, 1999) and leucine-rich repeats (Kobe & Deisenhofer, 1995). Some families are specific to prokaryotes, for example the ones of the bacterial glycosyl transferase (Wren, 1991) and of the ice nucleation protein (Gurian-Sherman & Lindow, 1993). These repetitive proteins are frequently found in regular arrangements such as a linear array or a superhelix, with repeats around a common axis, and are theoretically not limited in the number of repeated elements. The number of repeats tends to vary between homologous sequences. For example, homologous proteins from two strains of *Wolbachia pipientis* contain a variation in the number of ankyrin repeats (Iturbe-Ormaetxe *et al.*, 2005). This suggests that these proteins evolve rapidly by gain or loss of a repeated element. Very repetitive proteins are not necessarily disabled by a supplementary repeat. In fact, increasing the number of repeat of the ankyrin domain leads to an increase in the stability of the protein (Tripp & Barrick, 2004). Proteins with multiple repeated elements can have a large surface area available for ligand binding, and are often involved in protein–protein or protein–DNA interactions (Andrade *et al.*, 2001).

Tandem repeats can lead to misfolding and are typically avoided by proteins. Only 5% of sequences in archaea and 4% in bacteria contain domain tandem repeats, and slightly over 10% of domain families are found repeated in tandem (Apic *et al.*, 2001). This is somewhat surprising because tandem amplifications are mechanistically the most frequent and given the above-mentioned highly repetitive

proteins. When repeats are found in tandem, some residues named ‘structural gatekeepers’ may prevent the proteins from misfolding. These residues are conserved, and include charged residues, proline, glycine or disulfide bonds (Steward *et al.*, 2002; Parrini *et al.*, 2005; Han *et al.*, 2007). In Eukaryotes, domain aggregation is important for sequences with > 70% of sequence identity, but not detectable for sequences with < 40% identity (Wright *et al.*, 2005). Sequence identity between adjacent domain pairs is significantly lower than sequence identity between nonadjacent domains in the same protein. Adjacent domains could thus be under evolutionary pressure to evolve faster. Avoidance of aggregation could also explain why duplication of combinations of domains is more frequent than duplication of a single domain (Bjorklund *et al.*, 2006).

While most intragenic amplification events are expected to be deleterious, there are many protein sequences and structures with repeats. This shows that protein structures can sometimes adjust to the repeat and thereby allow proteins to acquire advantageous traits from internal amplifications. Some proteins are extremely repetitive. These families of proteins can withstand a large number of tandem repeats. While these provide striking examples of adaptive internal amplifications, they are relatively rare among crystallized proteins in prokaryotic genomes. Yet, most published studies have analyzed protein repeats from the point of view of protein domains. Additional studies using *ab initio* repeat-searching approaches are necessary to complement such analyses. More importantly, further structural studies are necessary in relation to protein sequence analyses to understand the evolution and function of intragenic repeats.

## The evolutionary impact of repeats

### Repeats, recombination and adaptation

One often considers living organisms as well adapted to their environments. However, environments change and thereby challenge the organisms’ fitness. Schematically, environmental challenges can be classified according to the frequency with which they arise and according to the possibility of building appropriate physiological responses to tackle them. The evolutionary patterns associated with adaptation to these challenges will dramatically depend on these two variables. Stresses that can be tackled by a physiological response based on the regulated expression of a set of proteins, such as protection from oxygen radicals by superoxide dismutase, can evolve if they occur frequently enough and if they can be discriminated by the cell to develop an appropriate regulatory response. In many cases, stresses are too complex, too elusive or too quickly lethal to be handled by a well-regulated stress response. In this case,

selection will purge the elements in the population that are unable to find a qualified solution to the standing challenge.

Point mutations lead to a slow generation of genetic novelty. This difficulty can be removed by increasing the mutation rates in times of stress (Bjedov *et al.*, 2003). While mutator subpopulations are frequently encountered in bacterial populations, this leads to a substantial burden of deleterious mutations (Sniegowski *et al.*, 2000). Hence, when stresses are frequent and can be tackled by the generation of variability at a reduced number of loci, there is often selection for the maintenance of repeats engaging in recombination at those loci. In this case, repeats are under second-order selection because they generate at high rates the adaptive changes with which they hitchhike to fixation (Tenaillon *et al.*, 2001). This does not suppose any sort of directionality; repeats will recombine randomly, resulting in sequence reassortment or gene conversion. When such events have a high probability of leading to adaptive changes, then the associated repeats will propagate in the population. On the other hand, repeats generating deleterious changes with a high frequency will in general be purged by the action of natural selection.

Programmed hotspots of variation by recombination are often found to be associated with functions managing antagonistic social interactions. When arms races develop between the two partners, for example a pathogenic bacterium and a human host, both partners are continuously changing to keep the other in check (Van Valen, 1973). These functions are often under negative frequency-dependent selection, i.e. a given variant is highly adaptive only when rare. This is the case when a variation arises in antigens. The fittest elements in the population are typically the new rare alleles that have not yet been discriminated by the immune system. As a result of their high fitness, these rare variants become more frequent in the population. However, this leads to a decrease in their fitness relative to new rare variants. Hence, such functions are constantly varying. Host-associated prokaryotes are constantly being faced with the host modulation of their behavior, but in general prokaryotes are also faced with parasites, such as phages, and predators, such as protozoa (Brussow, 2007). Most of the interactions of prokaryotic cells with other organisms involve molecules exposed at the cell surface or secreted into the environment. The corresponding genes are thus frequently under selection for diversification.

Many forms of diversifying selection share the peculiarity that they hardly select for higher efficiency, but instead for sequences that are different while performing an equivalent function. In this case, adaptation does not need to proceed by the invention of novel features and may result simply from reassortment of resident genetic information. This can be achieved by homologous recombination between a master gene and repeats scattered in the chromosome. It

can also be achieved by varying the expression of part of a pool of functionally redundant genes by transient gene deletions or by transient silencing of gene expression.

### Repeats as evolutionary reservoirs

Homologous recombination between one (or a few) master gene and copies of it, or parts of it, dispersed in the genome allows sequence variation without loss of function. Such repeats may arise naturally from tandem amplifications or horizontal transfer from closely related bacteria. As the system evolves, the repeats leading to recombination events with better results, in that they result in a large fraction of diverse but functional variants, will be selected for. This creates evolutionary reservoirs in the form of repeats scattered in the chromosome (Palmer & Brayton, 2007). With a dozen large degenerate repeats, such a system can provide for a huge amount of different combinations of protein sequences providing an endless source of genetic variation.

A particularly remarkable example of generation of genetic variation by homologous recombination is antigenic variation in *Borrelia*. These genomes have a large number of plasmids containing the vast majority of repeated elements of the genome (Casjens, 1999). Both in *Borrelia burgdorferi*, the agent of Lyme disease, and in *Borrelia hermsii*, the agent of relapsing fever, antigenic variation is under strong selection and proceeds by recombination processes within the repeats present in the plasmids (Norris, 2006). In *B. hermsii*, antigenic variation occurs mainly by gene conversion involving recombination at upstream and downstream homology sequences that result in replacement of the master genes *vlp* or *vsp*. There are over 60 such silent gene segments grouped into several clusters in the plasmids (Dai *et al.*, 2006). In the closely related *B. burgdorferi*, antigenic variation results from replacement of internal segments of varied lengths of the master gene by gene conversion resulting from its recombination with the repeated elements (Zhang *et al.*, 1997). In both cases, replacements occur at high frequencies during infection, thereby allowing continuous escape from the immune system (McDowell *et al.*, 2002).

While many cases have been found where repeats provide evolutionary reservoirs for antigenic variation, it must still be stressed that the immune system is just one of the many reasons why frequency-dependent selection occurs in pathogenic prokaryotes. For example, in the closely related *Mycoplasma pneumoniae* and *M. genitalium* genomes, many parts of the gene coding for the major adhesin, P1 in the former and MgPa in the latter, are repeated (Peterson *et al.*, 1995; Himmelreich *et al.*, 1996). Recombination between these repeats leads to the generation of new variants of the adhesins (Kenri *et al.*, 1999; Iverson-Cabral *et al.*, 2006, 2007). Although initial reports suggested that this variation

was linked to antigen variation, the analysis of genome data in the light of experimental analysis of patients' antibodies and the low rate of sequence variation suggests that the variations proceed for other reasons, for example to vary niche tropism (Rocha & Blanchard, 2002). Similarly, there is evidence that variation of genes involved in *S. enterica* O-antigens, polysaccharides decorating the cell surface, does not result from selection to escape the host immune system, but to escape protozoa grazing in the gut (Wildschutte *et al.*, 2004). As a case in point, the most abundant commensal in the human gut, *Bacteroides thetaiotaomicron*, has a very large number of repeat elements engaging in frequent recombination even though there are no described cases of pathogenicity for this species (Xu *et al.*, 2003). The most expanded classes of paralogues in its genome involve environmental sensing and polysaccharide uptake and biosynthesis, showing how recombination processes implicate other traits besides avoidance of the host immune system. Interestingly, even in the aforementioned *B. hermsii*, where antigenic variation seems to occur at high rates, there is evidence that expression of different *vsp* genes correlates with different niche tropism, because some variants are found more frequently in specific tissues (Cadavid *et al.*, 2001).

To assess the evolutionary role of repeats engaging in frequent gene conversion events, it is therefore important to account for the different ecological variables affecting fitness, which are rarely limited to antigen variation. It is also important to quantify the frequency of recombination. This will depend on the recombination machinery, but also on other factors. For example, in *E. coli*, the frequency of recombination is much higher for repeats within the same macrodomain because of chromosome folding (Valens *et al.*, 2004). Hence, the distribution of repeats could also have a direct impact on the rate of sequence variation by recombination processes.

### Transient gene inactivation

Gene deletions may be highly adaptive in contexts of host–pathogen associations or for the adoption of new ecological niches (Sokurenko *et al.*, 1999), even though they are expected to lower the cell fitness in the original habitat. Gene deletions may arise from recombination between repeats (Gaudriault *et al.*, 2008). While losing genes might appear to be a hazardous adaptive strategy, it does occur at high rates among mismatch repair genes, *mutS* and *mutL*, in enterobacteria (Bjedov *et al.*, 2003). The loss of these genes increases mutations rates 100- to 1000-fold. This typically leads to a high load of deleterious changes, but in maladapted populations, the rate of adaptive mutations may be enough to allow mutators to increase in frequency (Taddei *et al.*, 1997). It has been proposed that such deletions arise frequently because these genes are particularly rich in close

repeats (Rocha *et al.*, 2002), and indeed *mutS* deletions in mutator populations of *Pseudomonas aeruginosa* show close small repeats at the edges (Oliver *et al.*, 2002). Natural populations of *E. coli* contain 1–5% of mutator strains that, upon acquisition of an advantageous mutation, increase in frequency because they hitchhike with it. The newly well-adapted strain is better off if it loses its mutator phenotype, because in well-adapted populations the frequency of adaptive over deleterious mutations is lower. It is therefore adaptive for the new variant to revert to the nonmutator phenotype. This occurs at high frequencies because the loss of mismatch repair renders homologous recombination more permissive, thereby facilitating the recovery of the intact genes by horizontal transfer. Indeed, mismatch repair genes show very high rates of recombination (Denamur *et al.*, 2000). Gene deletion can be only partial and thereby allow intermediate mutation rates, as is the case in some isolates of *Pseudomonas putida*, where a third of the *mutS* gene is missing (Kurusu *et al.*, 2000). Experimental deletion analysis of the *mutS* gene of *E. coli* indicates that several large regions of the protein can be deleted, affecting the DNA binding and dimerization properties of the protein to varying degrees (Wu & Marinus, 1999). Therefore, different mutation rates may result from different recombination-mediated deletions between repeats.

It is unclear whether this model can be extended to many other functions. Losing a mismatch gene increases the likelihood of receiving one by horizontal transfer because recombination barriers are diminished. Most genes have no such feedback effect. Furthermore, this strategy is limited to prokaryotes with high rates of horizontal gene transfer such as *E. coli*, *Pseudomonas* or *Neisseria*. For genomes under sexual isolation, gene deletion is irreversible. In this case, the larger the number of repeats in a gene, the more likely its irreversible deletion from the genome. It has been suggested that reductive evolution of obligate intracellular symbionts resulted from deletions mediated by recombination between repeats (Frank *et al.*, 2002). An example of this is provided by *Buchnera*, which are sexually isolated endosymbiotic bacteria proliferating in the very protected intracellular environment of aphids (Baumann *et al.*, 1995). The comparison of two completely sequenced genomes of these bacteria showed that genes with a higher frequency of closely spaced repeats were more likely to be partially or totally deleted (Rocha, 2003a). Pseudogenization in these genomes was then confirmed to often result from deletions between close repeats (Gomez-Valero *et al.*, 2004). Such a scenario of reductive genome evolution by recombination-mediated deletions between close repeats will only unfold in the absence of horizontal transfer, counterbalancing gene loss. Accordingly, small genomes undergoing horizontal transfer, such as mollicutes (Sirand-Pugnet *et al.*, 2007), have many repeated elements involved in the evolution of virulence

(Razin *et al.*, 1998; Rocha & Blanchard, 2002) and do not show systematic evidence of further reductive evolution. One should note that gene deletions may occur even in the absence of repeats, especially in mutator populations (Nilsson *et al.*, 2005).

### Transient gene silencing

A reversible way of function gain and loss is naturally by gene expression regulation. Functions are not lost when gene expression can be turned on or off under specific conditions. Yet, handling a stress by regulating the expression of an appropriate response is impossible when both the stress and its solution are unforeseeable or they are lethal in such a short period that gene regulation cannot operate efficiently. This is often the case for pathogens tracked by the immune system (Moxon *et al.*, 2006). In this case, cells recollect no signal that their exposed proteins are being targeted by the immune system (when they are effectively targeted, it is too late). In this case, a classical gene regulation strategy is not efficient. Instead, if microbial cells are constantly changing the repertoire of proteins at the surface, then there is a constant generation of new elements in the population with variants that may escape the immune system. Contingency loci are elements that are stochastically turned on and off resulting in random variation, phase variation, of the expression of the proteins under diversifying selection (Moxon *et al.*, 1994). They can result from different mechanisms. In a manner similar to antigenic variation in *Borrelia*, they may involve homologous recombination between one master gene and repeats in the chromosomes. Yet, contrary to the previous examples, in this case, recombination leads to gene inactivation or silencing. For example, the expression of the type IV pilin of *Neisseria gonorrhoeae* can be silenced by recombination, followed by gene conversion with silent repeats in the genome (Swanson *et al.*, 1986).

In some cases, site-specific recombination systems are activated stochastically to produce chromosomal rearrangements leading to phase variation (Komano, 1999). Yet, the most frequently found elements supporting contingency loci are repeats engaged in strand-slippage (Henderson *et al.*, 1999), most notably SSR located either at the regulatory regions or in the genes themselves. SSR are unstable and thus endure frequent expansions and contractions. An SSR between the –10 and the –35 boxes of promoters will, by slippage, alter the spacing between the two motifs and thus abolish gene expression. An SSR in genes with a motif that is not a multiple of 3 may, by slippage, break the reading frame and lead to the production of an inactive protein. As a result of their instability, SSR are less frequent than expected in prokaryotic genomes, especially inside genes (Field & Wills, 1998; Ackermann & Chao, 2006), but, when present, they are often associated with genes under diversifying selection.

Hence, while SSR are typically purged from genomes, their presence can be selected for when it leads to the creation of contingency loci for functions under diversifying or negative frequency-dependent selection. There is ample evidence for the roles of SSR under such circumstances, especially regarding the evolution of pathogenic bacteria to escape the immune system (reviewed in van Belkum *et al.*, 1998; Moxon *et al.*, 2006).

An important property of SSR is that changes in copy number are heritable and reversible. By subsequent contraction and expansions, it is possible for a gene to oscillate between expression and silencing. This is unattainable for deletions between spaced repeats, which require hazardous recovery of the missing sequence by horizontal transfer. Switching rates of contingency loci are generally around  $10^{-2}$  to  $10^{-5}$  per generation. This is five to eight orders of magnitude higher than mutation rates in *E. coli*. For example, the *vsa* locus of *Mycoplasma pulmonis* contains a set of genes that are large tandem repeats of motifs of 36–57 nt, i.e. 12–19 amino acids (Shen *et al.*, 2000). The instability of the loci is so large that no single clone could be obtained and the genome sequence displays the most frequent alleles (Chambaud *et al.*, 2001). While considerable emphasis has been placed on the role of SSR in genes associated with pathogenesis, they also exist in nonpathogens; for example in *B. subtilis* strains, motility is abolished by variation in a tandem of eight adenines (Kearns *et al.*, 2004).

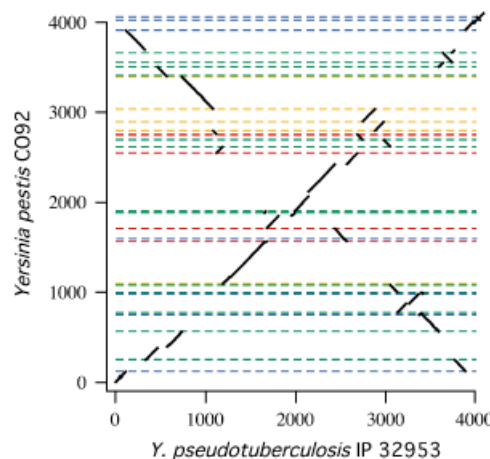
Interestingly, the loss of mismatch repair genes leads to SSR instability when motifs are < 4 nt (Richardson & Stojiljkovic, 2001; Bayliss *et al.*, 2004). Hence, gene deletion in *mutS* amplifies mutation rates in general, but more intensely so in SSR. This shows the tight associations that can be created within the different ways of generating variability in times of stress by means of recombination between repeats. While recombination between repeats may inactivate genes associated with mismatch repair, the loss of the latter leads to higher and more permissive recombination processes that result in further generation of genetic diversity.

It is important to note that the study of repeats as recombination reservoirs and as promoters of gene deletion or silencing has been carried out almost exclusively in obligatory or facultative pathogenic bacteria and in light of antigenic variation. Yet, repeats are frequent in many different genomes, and as noticed above they are widespread in free-living bacteria. Also, even within the studies on pathogenic bacteria, there is a large diversity of ways used to generate variability. As we have described, these include gene conversion, gene silencing and gene inactivation. While there are exceptions, large repeats engaging in homologous recombination or site-specific recombination tend to be associated with protein sequence variation, while close repeats are involved in gene inactivation and SSR in phase

variation. It remains a challenge to understand the roles of these different repeats from genome analysis alone. It is also largely unknown what are the roles of repeats in free-living bacteria, both in the type of targeted functions and in the frequency with which variation occurs.

### The impact of selfish DNA

Upon insertion, TEs may affect the expression of nearby genes, either by interrupting/silencing them or by enhancing their expression. TE elements also provide a structural basis to duplications, deletions, rearrangements and the incorporation of foreign DNA either by active transposition or by homologous recombination between the multiple copies present in a genome (Naas *et al.*, 1995; Alokam *et al.*, 2002; Kothapalli *et al.*, 2005; Redder & Garrett, 2006). TE can mediate the transfer of genetic information between genomes or between replicons of the same genome. They have thus been found to shuttle the transfer of adaptative traits, such as antibiotic resistance (Boutoille *et al.*, 2004), virulence (Kunze *et al.*, 1991; Lichter *et al.*, 1996) and new metabolic capabilities (Schmid-Appert *et al.*, 1997). TE are believed to undergo frequent horizontal transfer and cycles of expansion and extinction within a given species, most likely as a consequence of transfer between genomes and plasmids (Wagner, 2006). Their expansion, genome location and diversity may differ among closely related species, representing an important source of genomic diversity and plasticity (Schneider *et al.*, 2002; Brugger *et al.*, 2004; Nascimento *et al.*, 2004; Yang *et al.*, 2005). For example, *Bordetella bronchiseptica* lacks IS elements, which is in stark contrast to the 261 and the 112 ISs in *B. pertussis* and *Bordetella parapertussis* genomes, respectively (Parkhill *et al.*, 2003). Also, only 20 ISs elements were found in the *Yersinia pseudotuberculosis* chromosome, whereas two different strains of *Yersinia pestis* have 117 and 138 copies of ISs (Chain *et al.*, 2004). As a result of IS expansion, the genome of *Y. pestis* has endured many recent rearrangements (Fig. 9). IS may induce rearrangements either during homologous recombination or when transposing. It has been proposed that IS expansions are associated with a large number of factors, for example pathogenicity (Parkhill *et al.*, 2003), selection to generate variability (Shapiro, 1999), human or ecological associations (Mira *et al.*, 2006) or the frequency of horizontal gene transfer (Wagner, 2006). While horizontal transfer seems necessary for the existence of IS, genome size is to date the only statistically significant determinant of IS abundance (Touchon & Rocha, 2007). Indeed, the numbers and density of IS increase rapidly with genome size, which is the exact inverse trend found for the density of genes under strong selection such as essential genes, suggesting that ISs are as strongly counter-selected as the average gene is under strong selection.



**Fig. 9.** Dot-plot showing the relative position of orthologous genes in closely related *Yersinia* genomes. The dashed lines correspond to the rearrangement breakpoints flanked by repeated ISs. Each colour corresponds to an IS family.

Because the activity of TE has a direct impact on cell survival, control of transposition is tightly regulated at the transcriptional and translational level (for a review, see Nagy & Chandler, 2004). In addition, several host proteins have been identified as part of the transpososome, whose assembly may be controlled by host factors, thus integrating transposition activity into the host physiology (Gueguen *et al.*, 2005). Some equilibrium must be found between the impact of transposition events for the successful maintenance of the element and host viability. How the differential expansion occurs even among closely related species and strains remains to be fully understood, but is likely the result of changes in the efficiency of purifying selection and also due to uncontrolled proliferation of some elements (Wagner, 2006; Touchon & Rocha, 2007).

### Transient amplifications and amplicons

As mentioned previously (see Generic Repeats), two repeats separated by a spacer have the capacity to serve as amplicons, i.e. units that may amplify in tandem. If the flanking repeats are long and distant, the amplification will engage homologous recombination. If the repeats are close they will engage slipped-strand mispairing. If they are both long and close they will engage both. While we have mentioned the doubts surrounding the relevance of such structures for the creation of stable paralogues, many results show that these structures accelerate adaptation by transient amplifications of genetic material (Anderson & Roth, 1977). Tandem amplifications can affect almost any region of genomes because small close repeats are abundant and any pair of copies of a repeat can lead to amplification by either homologous, if they are large, or illegitimate recombination, if they are close. Some genomes, for example the ones very

G+C or A+T rich, are more likely to contain small repeats that can start the process of amplicon formation (Achaz *et al.*, 2002). They may thus be more prone to the creation of amplicons by recombination processes. TEs or rRNA genes may also be efficient starting points to the creation of amplicons and thus create hotspots of amplicon formation in genomes. In some rare cases, amplicons are extremely deleterious because they duplicate genetic elements that are needed in strictly single copies, but in the majority of cases they are neutral, i.e. without effect on fitness, beneficial or only mildly deleterious.

Amplifications show a high diversity of occurrence between regions because the frequency of recombination depends on the repeat size and similarity, their relative distance, the natural selection of the ensuing amplification and the probability of recombination between such chromosomal locations. In *S. enterica*, the frequencies of amplification can be as high as 3% per generation when they involve rRNA gene operons (Anderson & Roth, 1981). The amplification of an amplicon in the genome generates two large tandem repeats often spanning regions > 10 kb (Edlund & Normark, 1981; Janni re *et al.*, 1985; Smith *et al.*, 1988), and sometimes even 100 kb (Romero *et al.*, 1991; Kugelberg *et al.*, 2006). Such large repeats unbalance simultaneously the dosage of many genes and retard replication. They are therefore unlikely to be strictly neutral. For example, it has been observed that smaller amplicons resulting from deletion of larger ancestral ones amplify to larger numbers with lower growth consequences (Kugelberg *et al.*, 2006). Because repeats are abundant in genomes, a given locus can be under the action of several amplicons.

Most amplicons are very stable and amplify at low rates. Yet, amplification produces large tandem repeats that recombine at high frequencies. The very large repeats resulting from amplifications are remarkably unstable, and homologous recombination quickly leads to either deletion or further amplification (Romero & Palacios, 1997). Therefore, the rate-limiting step of the process corresponds to the initial amplification. Most experiments stimulating amplifications in genomes proceed by the inclusion of a gene with insufficient expression or efficiency. Amplifications covering this region will then be fixed if the gene function is under strong selection. Tandem arrays of drug resistance determinants can number over 50 copies and thus occupy significant fractions of the chromosome, up to 7.5% in *B. subtilis* (Janni re *et al.*, 1985) or 10% in *Deinococcus radiodurans* (Smith *et al.*, 1988). It has been proposed that some amplicons may attain particularly high amplification numbers, 45 copies in 2 h, by mechanisms that involve rolling-circle replication (Petit *et al.*, 1992). Several amplifications with important consequences are known. In *Vibrio cholerae*, the amplification of the cholera toxin proceeds by amplicons of 7 and 9.7 kb and the level of amplifications correlates with

virulence (Mekalanos, 1983). Similarly, the Shiga toxin genes in *Shigella dysenteriae* endure tandem amplification with a concomitant increase in toxin production (McDonough & Butters, 1999). In *Haemophilus influenzae* the capsule formation genes, which are stably duplicated, undergo further amplifications in children with meningitis and this is thought to improve invasiveness (Corn *et al.*, 1993). Tandem amplifications have been associated with resistance to drugs also among natural isolates of *E. coli* (Nichols & Guay, 1989) and *Streptococcus agalactiae* (Brochet *et al.*, 2008). It has even been proposed that clusters of functional neighboring genes, such as operons, are maintained to allow their coamplification (Reams & Neidle, 2004).

Many plasmids insert in the chromosome by duplication insertion, involving the production by homologous recombination of two copies of a repeat at the edges of the integrated element (Fig. 1). This is exactly the structure of an amplicon. Elements integrated in this way are immediately ready to be further amplified (Haldenwang *et al.*, 1980; Smith *et al.*, 1988). Laterally transferred genes are typically maladapted for expression on the host because their regulatory genes do not allow optimized expression. If they are amplified in the chromosome in high copy number, they may then adapt more rapidly by the joint action of three effects. First, the amplified regions offer a larger mutation target, i.e. even at stable mutation rates there will be more mutations accumulating in the amplified set of genes than in a single-gene locus. Second, recombination between repeats may engage the action of an error-prone polymerase and thus become locally mutagenic. Third, these repeats will accumulate changes and allele reassortment may occur within chromosomes by recombination between the copies, accelerating the combination of favorable changes in one gene. Transient amplifications have been suggested to explain the velocity with which leaky *lac* mutants revert to *lac*<sup>+</sup> more rapidly under conditions of stress (Cairns & Foster, 1991; Roth & Andersson, 2004). While the subject is still highly controversial (Stumpf *et al.*, 2007; Gonzalez *et al.*, 2008), amplification of the leaky *lac* locus occurs at high rates and leads to higher than expected reversions as predicted by the selective gene amplification model (Hendrickson *et al.*, 2002). Amplification can also result in higher evolutionary rates by leading to increased mutagenesis. The duplication of the DinB error-prone DNA polymerase leads to increased mutagenesis, by producing a transient mutator phenotype (Slecht *et al.*, 2003). Accordingly, it has been proposed that amplification increases significantly in times of stress (Slack *et al.*, 2006). For example, starved *E. coli* cells adapt through a small 46-bp duplication in the *rpoS* gene, the stationary phase-associated  $\sigma$  factor, because it results in a 'reduced function' phenotype (Zambrano *et al.*, 1993). Hence, by allowing the formation of amplicons, repeats can accelerate evolutionary processes in multiple ways.

## The fate of repeats

It is the fate of repeats to disappear. This can take very diverse time frames depending on the repeats. Genes under prolonged selection for gene dosage effects will last in genomes until such selection ceases. rRNA gene operons can thus remain in multiple copies for hundreds of millions of years as shown by the conservation of rRNA gene multiplicity in *Escherichia*, *Salmonella* and *Yersinia*. Some repeats will fade away because they accumulate point mutations. Even if they may remain in genomes, they will end up by losing the ability to recombine and therefore cease to be considered as repeats. Finally, repeats carrying no significant lasting advantage will disappear quickly either by deletion or by counter-selection, i.e. elimination of the lineage from the population. As a case in point, most of the large-scale duplications we found in our survey are contiguous and the copies of the repeats are strictly identical. This suggests that they are very recent segmental duplications. Because one does not observe large ancient segmental duplications in genomes, it follows that most very large amplifications are quickly deleted.

The creation of a repeat impacts genomes because it changes the genetic information and because it creates a recombination hotspot. The fitness difference associated with the repeat will result from the functional effect of the change plus from the average effects of recombination events involving the repeats weighted by their probability of occurrence. If the net fitness effect is negative, the individuals with the repeat will be removed from the population. Recombination events creating large deletions or inversions of genetic material are expected to be particularly deleterious. The effects of deletions are explained by the loss of associated functions, whereas the effects of large inversions result from selection for chromosome organization (Rocha, 2008). The rearrangement rate between large repeats in *E. coli* is high ( $10^{-5}$  per generation for pairs of rRNA gene operons) (Hill & Gray, 1988), but the number of fixed rearrangements is extremely low, a handful between *E. coli* and *S. enterica*, which diverged by an estimated *c.*  $10^{10}$  generations. Hence, while genomic rearrangement rates are close to genomic mutation rates, rearrangements are much more rarely fixed in populations than point substitutions because of their deleterious effects. Because the stability of genomes is largely a result of the rearrangements between repeats and ensuing natural selection, this means that (1) selection for organization is extremely strong and the vast majority of rearrangements are purged (Rocha, 2006); (2) genomes with many repeats exhibit high rearrangement rates and are less fit regarding chromosome organization and gene repertoire stability (Rocha, 2003b). As an example, the aforementioned highly repetitive genomes of *O. tsutsugamushi*, *B. pertussis* and *Y. pestis* show extremely large

numbers of rearrangements and pseudogenes when compared with closely related species. Interestingly, the careful analysis of the rearrangements between the strains of *Y. pestis* shows that successive rearrangements accumulated following paths that were at each time point less deleterious than expected by chance (Darling *et al.*, 2008). In all the three bacteria, such rearrangements seem to have been mediated by sudden expansions of IS (Fig. 9). These patterns are not circumscribed to rearrangements between IS elements because similar qualitative patterns of repeat-mediated rearrangements are found between rRNA gene operons (Hill & Harnish, 1981). However, no other repeated element is found in such large amounts as IS in genomes, which can span hundreds of identical copies.

The size of genomes is largely the result of the balance of forces increasing genetic information, horizontal transfer and amplifications, and forces removing genetic information, selection and genetic deletions. DNA lacking adaptive value is very quickly deleted, which has led to suggestions that there is a strong deletion bias in prokaryotic genomes (Lawrence *et al.*, 2001; Mira *et al.*, 2001), i.e. that recombination creates more deletions than amplifications. One should note that the dynamics of the system does not require a rate of deletion higher than that of amplification. As we mentioned previously, the limiting step in the formation of the amplicon is the creation of the flanking repeats. A deletion of these repeats will result in a sequence that will lead to a new amplicon at very low rates (Fig. 1). On the other hand, large tandems of amplicons are expected to be deleterious under most circumstances because they are a focus of genetic instability and they create a severe imbalance in the gene dosage. Hence, the absorption state of the deleted amplicon, together with counter-selection of large amplifications, will result in a deletion bias (Koch, 1979). Yet, given the high rates of gene deletion there have been attempts to identify mechanistic biases in the absence of selection. Some have argued that there is no mechanistic evidence for such a bias (Lovett, 2004). Others have found experimental evidence that in SSR supposedly under neutral evolution, i.e. without selection for or against SSR, deletions are more frequently observed than insertions at a rate of 2 : 1 in *H. influenzae* (De Bolle *et al.*, 2000), 3 : 1 in *Mycoplasma gallisepticum* (Metzgar *et al.*, 2002) and 10 : 1 in *E. coli* (Morel *et al.*, 1998). These biases appear to be excessively high to be explained without a deletion bias in slipped-strand mispairing. Interestingly, among the three cited genomes, *E. coli* is the one with fewer SSR, in line with the observed stronger deletion bias. If deletions are mechanistically more frequent than amplifications, then the short life of many repeats in genomes could simply result from stochastic effects and might explain why large tandem amplifications occur at high rates, but are rarely found in genome sequences.

For the same number of repeats, certain chromosomal distributions of the copies are less disruptive upon deletion or inversion than others. For example, close repeats will affect smaller fractions of the genome. Indeed, repeats tend to be distributed nonrandomly in genomes in such a way as to exert less negative effects on the chromosome upon recombination than expected by chance (Achaz *et al.*, 2003). In our dataset, there are 1.5 times more direct than inverted repeats ( $\geq 25$  nt), which might result from selection against inversions mediated by inverted repeats. They are also distributed more symmetrically around the origin of replication than expected, suggesting counter-selection of repeats that lead to asymmetric inversions (Achaz *et al.*, 2003).

In summary, repeats are opportunities, in the sense that they allow the generation of novel functions from pre-existing ones by evolutionary tinkering (Jacob, 1977), but they are also a problem, in that they challenge chromosome integrity and organization. As a consequence, understanding their role in genomes must account for the rate at which they are created and for their consequences from a functional and evolutionary perspective. The initial steps of such studies, repeat identification, contextualization and evolution, are now much simpler, thanks to genome sequences and bioinformatics. Because we ignore the role of many of the 111 268 repeat families we found in prokaryotes, there are fine perspectives for further experimental, computational and theoretical work.

## Acknowledgements

Work in our lab is supported by ANR contract ANR-07-GMGE-0 Flavogenomics. A.-L.A. was a recipient of a PhD grant from Region Ile-de-France. We are greatly indebted to the people with whom we have had discussions and collaborations over the years relating to repeats and recombination in genomes, notably Guillaume Achaz, Joel Pothier, Alain Blanchard, Bénédicte Michel, Ivan Matic, François Taddei, Erick Denamur, Eric Coissac, Pierre Netter and Alain Viari.

## Statement

Manuscript submitted to *FEMS Microbiology Reviews* based on the GMM3 meeting in Norway.

## References

- Abraham AL, Rocha EP & Pothier J (2008) Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics* **24**: 1536–1537.
- Achaz G, Coissac E, Viari A & Netter P (2000) Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. *Mol Biol Evol* **17**: 1268–1275.
- Achaz G, Rocha EPC, Netter P & Coissac E (2002) Origin and fate of repeats in bacteria. *Nucleic Acids Res* **30**: 2987–2994.
- Achaz G, Coissac E, Netter P & Rocha EPC (2003) Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* **164**: 1279–1289.
- Achaz G, Boyer F, Rocha EPC, Viari A & Coissac E (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics* **23**: 119–121.
- Acinas SG, Marcelino LA, Klepac-Ceraj V & Polz MF (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J Bacteriol* **186**: 2629–2635.
- Ackermann M & Chao L (2006) DNA sequences shaped by selection for stability. *PLoS Genet* **2**: e22.
- Alba MM, Laskowski RA & Hancock JM (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* **18**: 672–678.
- Albertini AM, Hofer M, Calos MP & Miller JH (1982) On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell* **29**: 319–328.
- Alm E, Huang K & Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* **2**: e143.
- Alokam S, Liu SL, Said K & Sanderson KE (2002) Inversions over the terminus region in *Salmonella* and *Escherichia coli*: IS200s as the sites of homologous recombination inverting the chromosome of *Salmonella enterica* serovar typhi. *J Bacteriol* **184**: 6190–6197.
- Amundsen SK & Smith GR (2003) Interchangeable parts of the *Escherichia coli* recombination machinery. *Cell* **112**: 741–744.
- Anderson P & Roth J (1981) Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (*rrn*) cistrons. *P Natl Acad Sci USA* **78**: 3113–3117.
- Anderson RP & Roth JR (1977) Tandem genetic duplications in phage and bacteria. *Annu Rev Microbiol* **31**: 473–505.
- Andrade MA, Ponting CP, Gibson TJ & Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol* **298**: 521–537.
- Andrade MA, Perez-Iratxeta C & Ponting CP (2001) Protein repeats: structures, functions, and evolution. *J Struct Biol* **134**: 117–131.
- Apic G, Gough J & Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* **310**: 311–325.
- Aras RA, Kang J, Tschumi AI, Harasaki Y & Blaser MJ (2003) Extensive repetitive DNA facilitates prokaryotic genome plasticity. *P Natl Acad Sci USA* **100**: 13579–13584.
- Bainton R, Gamas P & Craig NL (1991) Tn7 transposition *in vitro* proceeds through an excised transposon intermediate generated by staggered breaks in DNA. *Cell* **65**: 805–816.



- Bao Z & Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res* **12**: 1269–1276.
- Barrangou R, Fremaux C, Deveau H *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–1712.
- Barreiro V & Haggard-Ljungquist E (1992) Attachment sites for bacteriophage P2 on the *Escherichia coli* chromosome: DNA sequences, localization on the physical map, and detection of a P2-like remnant in *E. coli* K-12 derivatives. *J Bacteriol* **174**: 4086–4093.
- Baumann P, Baumann L, Lai CH & Rouhbakhsh D (1995) Genetics, physiology, and evolutionary relationships of the genus *Buchnera*: intracellular symbionts of aphids. *Annu Rev Microbiol* **49**: 55–94.
- Bayliss CD, Field D & Moxon ER (2001) The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *J Clin Invest* **107**: 657–662.
- Bayliss CD, Sweetman WA & Moxon ER (2004) Mutations in *Haemophilus influenzae* mismatch repair genes increase mutation rates of dinucleotide repeat tracts but not dinucleotide repeat-driven pilin phase variation rates. *J Bacteriol* **186**: 2928–2935.
- Bender J & Kleckner N (1986) Genetic evidence that Tn10 transposes by a nonreplicative mechanism. *Cell* **45**: 801–815.
- Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.
- Berg DE (1983) Structural requirement for IS50-mediated gene transposition. *P Natl Acad Sci USA* **80**: 792–796.
- Bergthorsson U, Andersson DI & Roth JR (2007) Ohno's dilemma: evolution of new genes under continuous selection. *P Natl Acad Sci USA* **104**: 17004–17009.
- Bernardi F & Bernardi A (1991) Inter- and intramolecular transposition of Tn903. *Mol Gen Genet* **227**: 22–27.
- Bhugra B & Dybvig K (1992) High-frequency rearrangements in the chromosome of *Mycoplasma pulmonis* correlate with phenotypic switching. *Mol Microbiol* **6**: 1149–1154.
- Biegert A & Soding J (2008) *De novo* identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **24**: 807–814.
- Bjedov I, Tenaillon O, Gerard B *et al.* (2003) Stress-induced mutagenesis in bacteria. *Science* **300**: 1404–1409.
- Bjorklund AK, Ekman D & Elofsson A (2006) Expansion of protein domain repeats. *PLoS Comput Biol* **2**: e114.
- Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC & Hugenholtz P (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**: 209.
- Boccard F & Prentki P (1993) Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units. *EMBO J* **12**: 5027.
- Bolland S & Kleckner N (1996) The three chemical steps of Tn10/IS10 transposition involve repeated utilization of a single active site. *Cell* **84**: 223–233.
- Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA & Weiner J III (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci* **62**: 435–445.
- Boutoille D, Corvec S, Caroff N *et al.* (2004) Detection of an IS21 insertion sequence in the *mexR* gene of *Pseudomonas aeruginosa* increasing beta-lactam resistance. *FEMS Microbiol Lett* **230**: 143–146.
- Bridges C (1935) Salivary chromosome maps. *J Hered* **26**: 60–64.
- Brochet M, Couve E, Zouine M, Poyart C & Glaser P (2008) A naturally occurring gene amplification leading to sulfonamide and trimethoprim resistance in *Streptococcus agalactiae*. *J Bacteriol* **190**: 672–680.
- Brouns SJ, Jore MM, Lundgren M *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**: 960–964.
- Brown AM, Coupland GM & Willetts NS (1984) Characterization of IS46, an insertion sequence found on two IncN plasmids. *J Bacteriol* **159**: 472–481.
- Brugger K, Torarinsson E, Redder P, Chen L & Garrett RA (2004) Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements. *Biochem Soc T* **32**: 179–183.
- Brussow H (2007) Bacteria between protists and phages: from antipredation strategies to the evolution of pathogenicity. *Mol Microbiol* **65**: 583–589.
- Brussow H, Canchaya C & Hardt WD (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol R* **68**: 560–602.
- Bucka A & Stasiak A (2001) RecA-mediated strand exchange traverses substitutional heterologies more easily than deletions or insertions. *Nucleic Acids Res* **29**: 2464–2470.
- Cadavid D, Pachner AR, Estanislao L, Patalapati R & Barbour AG (2001) Isogenic serotypes of *Borrelia turicatae* show different localization in the brain and skin of mice. *Infect Immun* **69**: 3389–3397.
- Cairns J & Foster PL (1991) Adaptive reversion of a frameshift mutation in *Escherichia coli*. *Genetics* **128**: 695–701.
- Casjens S (1999) Evolution of the linear DNA replicons of the *Borrelia* spirochetes. *Curr Opin Microbiol* **2**: 529–534.
- Chain PS, Carniel E, Larimer FW *et al.* (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *P Natl Acad Sci USA* **101**: 13826–13831.
- Chambaud I, Heilig R, Ferris S *et al.* (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res* **29**: 2145–2153.
- Chang CH, Chang YC, Underwood A, Chiou CS & Kao CY (2007) VNTRDB: a bacterial variable number tandem repeat locus database. *Nucleic Acids Res* **35**: D416–D421.
- Chédin F, Dervyn E, Ehrlich SD & Noirot P (1994) Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Mol Microbiol* **12**: 561–569.
- Chen SL & Shapiro L (2003) Identification of long intergenic repeat sequences associated with DNA methylation sites in

- Caulobacter crescentus* and other alpha-proteobacteria. *J Bacteriol* **185**: 4997–5002.
- Chistoserdova L, Lapidus A, Han C *et al.* (2007) Genome of *Methylobacillus flagellatus*, molecular basis for obligate methylotrophy, and polyphyletic origin of methylotrophy. *J Bacteriol* **189**: 4020–4027.
- Cho NH, Kim HR, Lee JH *et al.* (2007) The *Orientia tsutsugamushi* genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. *P Natl Acad Sci USA* **104**: 7981–7986.
- Clement JM, Wilde C, Bachelier S, Lambert P & Hofnung M (1999) IS1397 is active for transposition into the chromosome of *Escherichia coli* K-12 and inserts specifically into palindromic units of bacterial interspersed mosaic elements. *J Bacteriol* **181**: 6929–6936.
- Corn PG, Anders J, Takala AK, Kayhty H & Hoiseth SK (1993) Genes involved in *Haemophilus influenzae* type b capsule expression are frequently amplified. *J Infect Dis* **167**: 356–364.
- Correia FF, Inouye S & Inouye M (1988) A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae*. *J Biol Chem* **263**: 12194–12198.
- Costechareyre D, Bertolla F & Nesme X (2009) Homologous recombination in *Agrobacterium*: potential implications for the genomic species concept in bacteria. *Mol Biol Evol* **26**: 167–176.
- Cozzuto L, Petrillo M, Silvestro G, Di Nocera PP & Paoletta G (2008) Systematic identification of stem-loop containing sequence families in bacterial genomes. *BMC Genomics* **9**: 20.
- Curcio MJ & Derbyshire KM (2003) The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Bio* **4**: 865–877.
- Dai Q, Restrepo BI, Porcella SF, Raffel SJ, Schwan TG & Barbour AG (2006) Antigenic variation by *Borrelia hermsii* occurs through recombination between extragenic repetitive elements on linear plasmids. *Mol Microbiol* **60**: 1329–1343.
- Darling AE, Miklos I & Ragan MA (2008) Dynamics of genome rearrangement in bacterial populations. *PLoS Genet* **4**: e1000128.
- Das AK, Cohen PW & Barford D (1998) The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *EMBO J* **17**: 1192–1199.
- De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW & Moxon ER (2000) The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol Microbiol* **35**: 211–222.
- De Gregorio E, Abrescia C, Carlomagno MS & Di Nocera PP (2003a) Ribonuclease III-mediated processing of specific *Neisseria meningitidis* mRNAs. *Biochem J* **374**: 799–805.
- De Gregorio E, Abrescia C, Carlomagno MS & Di Nocera PP (2003b) Asymmetrical distribution of *Neisseria* miniature insertion sequence DNA repeats among pathogenic and nonpathogenic *Neisseria* strains. *Infect Immun* **71**: 4217–4221.
- De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS & Di Nocera PP (2005) Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J Bacteriol* **187**: 7945–7954.
- De Gregorio E, Silvestro G, Venditti R, Carlomagno MS & Di Nocera PP (2006) Structural organization and functional properties of miniature DNA insertion sequences in yersiniae. *J Bacteriol* **188**: 7876–7884.
- Delgrange O & Rivals E (2004) STAR: an algorithm to search for tandem approximate repeats. *Bioinformatics* **20**: 2812–2820.
- Delhas N (2008) Small mobile sequences in bacteria display diverse structure/function motifs. *Mol Microbiol* **67**: 475–481.
- Denamur E, Lecomte G, Darlu P *et al.* (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**: 711–721.
- Doulavov S, Hodes A, Dai L *et al.* (2004) Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**: 476–481.
- Duncan CH, Wilson GA & Young FE (1978) Mechanism of integrating foreign DNA during transformation of *Bacillus subtilis*. *P Natl Acad Sci USA* **75**: 3664–3668.
- Dutra BE, Suter VA Jr & Lovett ST (2007) RecA-independent recombination is efficient but limited by exonucleases. *P Natl Acad Sci USA* **104**: 216–221.
- Edgar RC (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**: 18.
- Edgar RC & Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (suppl 1): i152–i158.
- Eklund T & Normark S (1981) Recombination between short DNA homologies causes tandem duplication. *Nature* **292**: 269–271.
- Ekman D, Bjorklund AK, Frey-Skott J & Eklund A (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* **348**: 231–243.
- Eklund A & Sonnhammer EL (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* **15**: 480–500.
- Espeli O & Boccard F (1997) *In vivo* cleavage of *Escherichia coli* BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site. *Mol Microbiol* **26**: 767–777.
- Espeli O, Moulin L & Boccard F (2001) Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *J Mol Biol* **314**: 375–386.
- Ferat JL & Michel F (1993) Group II self-splicing introns in bacteria. *Nature* **364**: 358–361.
- Field D & Wills C (1998) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae* and the different distributions of microsatellites in 8 prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *P Natl Acad Sci USA* **95**: 1647–1652.

- Fouts DE (2006) Phage\_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* **34**: 5839–5851.
- Francino MP (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet* **37**: 573–577.
- Frangoul L, Quillardet P, Castets AM *et al.* (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* **9**: 274.
- Frank AC, Amiri H & Andersson SG (2002) Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica* **115**: 1–12.
- Frutos R, Viari A, Ferraz C *et al.* (2006) Comparative genomic analysis of three strains of *Ehrlichia ruminantium* reveals an active process of genome size plasticity. *J Bacteriol* **188**: 2533–2542.
- Gaudriault S, Pages S, Lanois A, Laroui C, Teyssier C, Jumas-Bilak E & Givaudan A (2008) Plastic architecture of bacterial genome revealed by comparative genomics of *Photobacterium* variants. *Genome Biol* **9**: R117.
- Gevers D, Vandepoele K, Simillon C & Van de Peer Y (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* **12**: 148–154.
- Gil R, Silva FJ, Zientz E *et al.* (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *P Natl Acad Sci USA* **100**: 9388–9393.
- Gilson E, Rousset JP, Clement JM & Hofnung M (1986) A subfamily of *E. coli* palindromic units implicated in transcription termination? *Ann Inst Pasteur Mic* **137B**: 259–270.
- Gilson E, Perrin D & Hofnung M (1990) DNA polymerase I and a protein complex bind specifically to *E. coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res* **18**: 3941–3952.
- Gilson E, Saurin W, Perrin D, Bachellier S & Hofnung M (1991) Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res* **19**: 1375–1383.
- Goldman BS, Nierman WC, Kaiser D *et al.* (2006) Evolution of sensory complexity recorded in a myxobacterial genome. *P Natl Acad Sci USA* **103**: 15200–15205.
- Gomez-Valero L, Latorre A & Silva FJ (2004) The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola*. *Mol Biol Evol* **21**: 2172–2181.
- Gonzalez C, Hadany L, Ponder RG, Price M, Hastings PJ & Rosenberg SM (2008) Mutability and importance of a hypermutable cell subpopulation that produces stress-induced mutants in *Escherichia coli*. *PLoS Genet* **4**: e1000208.
- Grissa I, Vergnaud G & Pourcel C (2007a) CRISPRfinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52–W57.
- Grissa I, Vergnaud G & Pourcel C (2007b) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**: 172.
- Gruber M, Soding J & Lupas AN (2005) REPPER – repeats and their periodicities in fibrous proteins. *Nucleic Acids Res* **33**: W239–W243.
- Gueguen E, Rousseau P, Duval-Valentin G & Chandler M (2005) The transpososome: control of transposition at the level of catalysis. *Trends Microbiol* **13**: 543–549.
- Gurian-Sherman D & Lindow SE (1993) Bacterial ice nucleation: significance and molecular basis. *FASEB J* **7**: 1338–1343.
- Haft DH, Selengut J, Mongodin EF & Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**: e60.
- Haldenwang WG, Banner CD, Ollington JF, Losick R, Hoch JA, O'Connor MB & Sonenshein AL (1980) Mapping a cloned gene under sporulation control by insertion of a drug resistance marker into the *Bacillus subtilis* chromosome. *J Bacteriol* **142**: 90–98.
- Hallin PF & Ussery DW (2004) CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* **20**: 3682–3686.
- Han JH, Batey S, Nickson AA, Teichmann SA & Clarke J (2007) The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Bio* **8**: 319–330.
- Hancock JM (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica* **115**: 93–103.
- Hao W & Golding GB (2006) The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res* **16**: 636–643.
- Harvey S & Hill CW (1990) Exchange of spacer regions between rRNA operons in *Escherichia coli*. *Genetics* **125**: 683–690.
- Haubold B & Wiehe T (2006) How repetitive are genomes? *BMC Bioinformatics* **7**: 541.
- Hefron F (1983) Tn3 and its relatives. *Mobile Genetic Elements, Vol. 1* (Shapiro J, ed), pp. 223–260. American Society for Microbiology, Washington, DC.
- Heger A & Holm L (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**: 224–237.
- Henderson IR, Owen P & Nataro JP (1999) Molecular switches – the ON and OFF of bacterial phase variation. *Mol Microbiol* **33**: 919–932.
- Hendrickson H, Slechts ES, Bergthorsson U, Andersson DI & Roth JR (2002) Amplification-mutagenesis: evidence that “directed” adaptive mutation and general hypermutability result from growth with a selected gene amplification. *P Natl Acad Sci USA* **99**: 2164–2169.
- Higgins CF, Ames GF, Barnes WM, Clement JM & Hofnung M (1982) A novel intercistronic regulatory element of prokaryotic operons. *Nature* **298**: 760–762.
- Higgins CF, McLaren RS & Newbury SF (1988) Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. *Gene* **72**: 3–14.
- Hill CW & Gray JA (1988) Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. *Genetics* **119**: 771–778.
- Hill CW & Harnish B (1981) Inversions between ribosomal RNA genes of *E. coli*. *P Natl Acad Sci USA* **78**: 7069–7072.

- Himmelreich R, Hilbert H, Plagens H, Pirki E, Li BC & Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**: 4420–4449.
- Holden MT, Titball RW, Peacock SJ *et al.* (2004) Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *P Natl Acad Sci USA* **101**: 14240–14245.
- Horvath P, Romero DA, Coute-Monvoisin AC *et al.* (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**: 1401–1412.
- Hughes D (2000) Co-evolution of the *tuf* genes links gene conversion with the generation of chromosomal inversions. *J Mol Biol* **297**: 355–364.
- Hulton CS, Higgins CF & Sharp PM (1991) ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. *Mol Microbiol* **5**: 825–834.
- Isabel S, Leblanc E, Boissinot M *et al.* (2008) Divergence among genes encoding the elongation factor Tu of *Yersinia* species. *J Bacteriol* **190**: 7548–7558.
- Ishino Y, Shinagawa H, Makino K, Amemura M & Nakata A (1987) Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol* **169**: 5429–5433.
- Iturbe-Ormaetxe I, Burke GR, Riegler M & O'Neill SL (2005) Distribution, expression, and motif variability of ankyrin domain genes in *Wolbachia pipientis*. *J Bacteriol* **187**: 5136–5145.
- Iverson-Cabral SL, Astete SG, Cohen CR, Rocha EP & Totten PA (2006) Intrastrain heterogeneity of the *mgpB* gene in *Mycoplasma genitalium* is extensive *in vitro* and *in vivo* and suggests that variation is generated via recombination with repetitive chromosomal sequences. *Infect Immun* **74**: 3715–3726.
- Iverson-Cabral SL, Astete SG, Cohen CR & Totten PA (2007) *mgpB* and *mgpC* sequence diversity in *Mycoplasma genitalium* is generated by segmental reciprocal recombination with repetitive chromosomal sequences. *Mol Microbiol* **66**: 55–73.
- Jacob F (1977) Evolution and tinkering. *Science* **196**: 1161–1166.
- Janni  re L, Niaudet B, Pierre E & Ehrlich SD (1985) Stable gene amplification in the chromosome of *Bacillus subtilis*. *Gene* **40**: 47–55.
- Jansen R, van Embden JD, Gaastra W & Schouls LM (2002) Identification of a novel family of sequence repeats among prokaryotes. *Omic* **6**: 23–33.
- Jiang N, Feschotte C, Zhang X & Wessler SR (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr Opin Plant Biol* **7**: 115–119.
- Kalita MK, Ramasamy G, Duraisamy S, Chauhan VS & Gupta D (2006) ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics* **7**: 336.
- Karlin S & Cardon LR (1994) Computational DNA sequence analysis. *Annu Rev Microbiol* **48**: 619–654.
- Karlin S & Ost F (1985) Maximal segmental match length among random sequences from a finite alphabet. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. I* (Cam LML & Olsen RA, eds), pp. 225–243. Wadsworth Inc., New York.
- Katti MV, Sami-Subbu R, Ranjekar PK & Gupta VS (2000) Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci* **9**: 1203–1209.
- Kawabata T (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res* **31**: 3367–3369.
- Kearns DB, Chu F, Rudner R & Losick R (2004) Genes governing swarming in *Bacillus subtilis* and evidence for a phase variation mechanism controlling surface motility. *Mol Microbiol* **52**: 357–369.
- Kenri T, Taniguchi R, Sasaki Y *et al.* (1999) Identification of a new variable sequence in the P1 cytoadhesin gene of *Mycoplasma pneumoniae*: evidence for the generation of antigenic variation by DNA recombination between repetitive sequences. *Infect Immun* **67**: 4557–4562.
- Khemici V & Carpousis AJ (2004) The RNA degradosome and poly(A) polymerase of *Escherichia coli* are required *in vivo* for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Mol Microbiol* **51**: 777–790.
- Klappenbach JA, Dunbar JM & Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microb* **66**: 1328–1333.
- Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J & Keim P (2001) Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. *J Clin Microbiol* **39**: 3179–3185.
- Kobe B & Deisenhofer J (1995) Proteins with leucine-rich repeats. *Curr Opin Struc Biol* **5**: 409–416.
- Koch AL (1979) Selection and recombination in populations containing tandem multiplet genes. *J Mol Evol* **14**: 273–285.
- Kofoed E, Bergthorsson U, Slechts ES & Roth JR (2003) Formation of an F' plasmid by recombination between imperfectly repeated chromosomal Rep sequences: a closer look at an old friend (F'(128) pro lac). *J Bacteriol* **185**: 660–663.
- Kolpakov R, Bana G & Kucherov G (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* **31**: 3672–3678.
- Komano T (1999) Shufflons: multiple inversion systems and integrons. *Annu Rev Genet* **33**: 171–191.
- Kothapalli S, Nair S, Alokam S *et al.* (2005) Diversity of genome structure in *Salmonella enterica* serovar Typhi populations. *J Bacteriol* **187**: 2638–2650.
- Kowalczykowski SC (2000) Initiation of genetic recombination and recombination-dependent replication. *Trends Biochem Sci* **25**: 156–165.
- Kugelberg E, Kofoed E, Reams AB, Andersson DI & Roth JR (2006) Multiple pathways of selected gene amplification during adaptive mutation. *P Natl Acad Sci USA* **103**: 17319–17324.

- Kumagai M & Ikeda H (1991) Molecular analysis of the recombination junctions of lambda bio transducing phages. *Mol Gen Genet* **230**: 60–64.
- Kunin V & Ouzounis CA (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res* **13**: 1589–1594.
- Kunin V, Sorek R & Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**: R61.
- Kunze ZM, Wall S, Appelberg R, Silva MT, Portaels F & McFadden JJ (1991) IS901, a new member of a widespread class of atypical insertion sequences, is associated with pathogenicity in *Mycobacterium avium*. *Mol Microbiol* **5**: 2265–2272.
- Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J & Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* **29**: 4633–4642.
- Kurusu Y, Narita T, Suzuki M & Watanabe T (2000) Genetic analysis of an incomplete mutS gene from *Pseudomonas putida*. *J Bacteriol* **182**: 5278–5279.
- Lampson BC, Inouye M & Inouye S (2005) Retrons, msDNA, and the bacterial genome. *Cytogenet Genome Res* **110**: 491–499.
- Lavorgna G, Patthy L & Boncinelli E (2001) Were protein internal repeats formed by “bricolage”? *Trends Genet* **17**: 120–123.
- Lawrence JG, Hendrix RW & Casjens S (2001) Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9**: 535–540.
- Leclercq S, Rivals E & Jarne P (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics* **8**: 125.
- Le Fleche P, Hauck Y, Onteniente L *et al.* (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol* **1**: 2.
- Leplae R, Hebrant A, Wodak SJ & Toussaint A (2004) ACLAME: a CLAssification of mobile genetic elements. *Nucleic Acids Res* **32**: D45–D49.
- Lerat E, Daubin V, Ochman H & Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* **3**: e130.
- Levinson G & Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**: 203–221.
- Li R, Ye J, Li S *et al.* (2005) ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol* **1**: e43.
- Lichens-Park A & Syvanen M (1988) Cointegrate formation by IS50 requires multiple donor molecules. *Mol Gen Genet* **211**: 244–251.
- Lichter A, Manulis S, Valinsky L, Karniol B & Barash I (1996) IS1327, a new insertion-like element in the pathogenicity-associated plasmid of *Erwinia herbicola* pv. *gypsophilae*. *Mol Plant Microbe In* **9**: 98–104.
- Liu J & Rost B (2004) Sequence-based prediction of protein domains. *Nucleic Acids Res* **32**: 3522–3530.
- Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* **52**: 1243–1253.
- Lovett ST, Gluckman TJ, Simon PJ, Sutera VA & Drapkin PT (1994) Recombination between repeats in *E. coli* by a recA-independent, proximity-sensitive mechanism. *Mol Gen Genet* **245**: 294–300.
- Lynch M & Katju V (2004) The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20**: 544–549.
- Mahillon J & Chandler M (1998) Insertion sequences. *Microbiol Mol Biol R* **62**: 725–774.
- Makarova KS, Grishin NV, Shabalina SA, Wolf YI & Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**: 7.
- Marcotte EM, Pellegrini M, Yeates TO & Eisenberg D (1998) A census of protein repeats. *J Mol Biol* **293**: 151–160.
- Marraffini LA & Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**: 1843–1845.
- Martinez HM (1983) An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res* **11**: 4629–4634.
- McDonough MA & Butters JR (1999) Spontaneous tandem amplification and deletion of the shiga toxin operon in *Shigella dysenteriae* 1. *Mol Microbiol* **34**: 1058–1069.
- McDowell JV, Sung SY, Hu LT & Marconi RT (2002) Evidence that the variable regions of the central domain of VlsE are antigenic during infection with lyme disease spirochetes. *Infect Immun* **70**: 4196–4203.
- McLeod MP, Warren RL, Hsiao WW *et al.* (2006) The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse. *P Natl Acad Sci USA* **103**: 15582–15587.
- Medhekar B & Miller JF (2007) Diversity-generating retroelements. *Curr Opin Microbiol* **10**: 388–395.
- Mekalanos JJ (1983) Duplication and amplification of toxin genes in *Vibrio cholerae*. *Cell* **35**: 253–263.
- Merkel A & Gemmell N (2008) Detecting short tandem repeats from genome data: opening the software black box. *Brief Bioinform* **9**: 355–366.
- Metzgar D, Thomas E, Davis C, Field D & Wills C (2001) The microsatellites of *Escherichia coli*: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. *Mol Microbiol* **39**: 183–190.
- Metzgar D, Liu L, Hansen C, Dybvig K & Wills C (2002) Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. *Genome Res* **12**: 408–413.
- Michel B, Grompone G, Florès MJ & Bidnenko V (2004) Multiple pathways process stalled replication forks. *P Natl Acad Sci USA* **101**: 12783–12788.
- Mira A, Ochman H & Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**: 589–596.
- Mira A, Pushker R & Rodriguez-Valera F (2006) The Neolithic revolution of bacterial genomes. *Trends Microbiol* **14**: 200–206.

- Mizuuchi K (1992) Transpositional recombination: mechanistic insights from studies of mu and other elements. *Annu Rev Biochem* **61**: 1011–1051.
- Mojica FJ, Diez-Villasenor C, Soria E & Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**: 244–246.
- Morel P, Reverdy C, Michel B, Ehrlich SD & Cassuto E (1998) The role of SOS and flap processing in microsatellite instability in *Escherichia coli*. *P Natl Acad Sci USA* **95**: 10003–10008.
- Moxon ER, Rainey PB, Nowak MA & Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* **4**: 24–33.
- Moxon R, Bayliss C & Hood D (2006) Bacterial contingency Loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* **40**: 307–333.
- Mrazek J & Xie S (2006) Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences. *Bioinformatics* **22**: 3099–3100.
- Mrazek J, Guo X & Shah A (2007) Simple sequence repeats in prokaryotic genomes. *P Natl Acad Sci USA* **104**: 8472–8477.
- Mudunuri SB & Nagarajaram HA (2007) IMEx: imperfect microsatellite extractor. *Bioinformatics* **23**: 1181–1187.
- Muller A, MacCallum RM & Sternberg MJ (2002) Structural characterization of the human proteome. *Genome Res* **12**: 1625–1641.
- Murray KB, Taylor WR & Thornton JM (2004) Toward the detection and validation of repeats in protein structure. *Proteins* **57**: 365–380.
- Naas T, Blot M, Fitch WM & Arber W (1995) Dynamics of IS-related genetic rearrangements in resting *Escherichia coli* K-12. *Mol Biol Evol* **12**: 198–207.
- Nagy Z & Chandler M (2004) Regulation of transposition in bacteria. *Res Microbiol* **155**: 387–398.
- Nakagawa I, Kurokawa K, Yamashita A et al. (2003) Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res* **13**: 1042–1055.
- Nascimento AL, Ko AI, Martins EA et al. (2004) Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. *J Bacteriol* **186**: 2164–2172.
- Newman AM & Cooper JB (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**: 382.
- Niaudet B, Jannière L & Ehrlich SD (1985) Integration of linear, heterologous DNA molecules into the *Bacillus subtilis* chromosome: mechanism and use in induction of predictable rearrangements. *J Bacteriol* **163**: 111–120.
- Nichols BP & Guay GG (1989) Gene amplification contributes to sulfonamide resistance in *Escherichia coli*. *Antimicrob Agents Ch* **33**: 2042–2048.
- Nierman WC, DeShazer D, Kim HS et al. (2004) Structural flexibility in the *Burkholderia mallei* genome. *P Natl Acad Sci USA* **101**: 14246–14251.
- Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JC & Andersson DI (2005) Bacterial genome size reduction by experimental evolution. *P Natl Acad Sci USA* **102**: 12112–12116.
- Norris SJ (2006) Antigenic variation with a twist – the *Borrelia* story. *Mol Microbiol* **60**: 1319–1322.
- Ogata H, Audic S, Barbe V, Artiguenave F, Fournier PE, Raoult D & Claverie JM (2000) Selfish DNA in protein-coding genes of rickettsia. *Science* **290**: 347–350.
- Oggioni MR & Claverys JP (1999) Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology* **145**: 2647–2653.
- Ohno S (1970) *Evolution by Gene Duplication*. Springer-Verlag, Berlin.
- Oliveira PH, Lemos F, Monteiro GA & Prazeres DM (2008) Recombination frequency in plasmid DNA containing direct repeats – predictive correlation with repeat and intervening sequence length. *Plasmid* **60**: 159–165.
- Oliver A, Baquero F & Blázquez J (2002) The mismatch repair system (mutS, mutL, and uvrD genes) in *Pseudomonas aeruginosa*: molecular characterisation of naturally occurring mutants. *Mol Microbiol* **43**: 1641–1650.
- Oppenheim AB, Rudd KE, Mendelson I & Teff D (1993) Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli*. *Mol Microbiol* **10**: 113–122.
- Oshima K, Kakizawa S, Nishigawa H et al. (2004) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat Genet* **36**: 27–29.
- Palmer GH & Brayton KA (2007) Gene conversion is a convergent strategy for pathogen antigenic variation. *Trends Parasitol* **23**: 408–413.
- Parkhill J, Sebaihia M, Preston A et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35**: 32–40.
- Parrini C, Taddei N, Ramazzotti M, Degl'Innocenti D, Ramponi G, Dobson CM & Chiti F (2005) Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure* **13**: 1143–1151.
- Pasek S, Risler JL & Brezellec P (2006) The role of domain redundancy in genetic robustness against null mutations. *J Mol Biol* **362**: 184–191.
- Pato M (1989) Bacteriophage Mu. *Mobile DNA* (Berg D & Howe M, eds), pp. 23–52. ASM Press, Washington, DC.
- Patrick WM, Quandt EM, Swartzlander DB & Matsumura I (2007) Multicopy suppression underpins metabolic evolvability. *Mol Biol Evol* **24**: 2716–2722.
- Peeters BP, de Boer JH, Bron S & Venema G (1988) Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. *Mol Gen Genet* **212**: 450–458.

- Pellegrini M, Marcotte EM & Yeates TO (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* **35**: 440–446.
- Peterson SN, Bailey CC, Jensen JS, Borre MB, King ES, Bott KF & Hutchison CA (1995) Characterisation of repetitive DNA in the *Mycoplasma genitalium* genome: possible role in the generation of antigenic variation. *P Natl Acad Sci USA* **92**: 11829–11833.
- Petes TD, Greenwell PW & Dominska M (1997) Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. *Genetics* **146**: 491–498.
- Petit MA, Mesas JM, Noirot P, Morel-Deville F & Ehrlich SD (1992) Induction of DNA amplification in the *Bacillus subtilis* chromosome. *EMBO J* **11**: 1317–1326.
- Pierce JC, Kong D & Masker W (1991) The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7. *Nucleic Acids Res* **19**: 3901–3905.
- Price AL, Jones NC & Pevzner PA (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl 1): i351–i358.
- Rayssiguier C, Thaler DS & Radman M (1989) The barrier to recombination between *E. coli* and *S. typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342**: 396–401.
- Razin S, Yogev D & Naot Y (1998) Molecular biology and pathogenicity of Mycoplasmas. *Microbiol Mol Biol R* **62**: 1094–1165.
- Reams AB & Neidle EL (2004) Selection for gene clustering by tandem duplication. *Annu Rev Microbiol* **58**: 119–142.
- Redder P & Garrett RA (2006) Mutations and rearrangements in the genome of *Sulfolobus solfataricus* P2. *J Bacteriol* **188**: 4198–4206.
- Richardson AR & Stojiljkovic I (2001) Mismatch repair and the regulation of phase variation in *Neisseria meningitidis*. *Mol Microbiol* **40**: 645–655.
- Robin S, Schbath S & Vandewalle V (2007) Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics* **8**: 84.
- Rocha EPC (2003a) An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res* **13**: 1123–1132.
- Rocha EPC (2003b) DNA repeats lead to the accelerated loss of gene order in Bacteria. *Trends Genet* **19**: 600–604.
- Rocha EPC (2006) Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol* **23**: 513–522.
- Rocha EPC (2008) The organisation of the bacterial genome. *Annu Rev Genet* **42**: 211–233.
- Rocha EPC & Blanchard A (2002) Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. *Nucleic Acids Res* **30**: 2031–2042.
- Rocha EPC, Danchin A & Viari A (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. *Mol Biol Evol* **16**: 1219–1230.
- Rocha EPC, Matic I & Taddei F (2002) Over-representation of close repeats in stress response genes: a strategy to increase versatility under stressful conditions? *Nucleic Acids Res* **30**: 1886–1894.
- Rocha EPC, Cornet E & Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* **1**: e15.
- Romero D & Palacios R (1997) Gene amplification and genomic plasticity in prokaryotes. *Annu Rev Genet* **31**: 91–111.
- Romero D, Brom S, Martinez-Salazar J, Girard ML, Palacios R & Davila G (1991) Amplification and deletion of a nod-nif region in the symbiotic plasmid of *Rhizobium phaseoli*. *J Bacteriol* **173**: 2435–2441.
- Roset R, Subirana JA & Messeguier X (2003) MREPATT: detection and analysis of exact consecutive repeats in genomic sequences. *Bioinformatics* **19**: 2475–2476.
- Roth JR & Andersson DI (2004) Amplification-mutagenesis-how growth under selection contributes to the origin of genetic diversity and explains the phenomenon of adaptive mutation. *Res Microbiol* **155**: 342–351.
- Sagi D, Tlusty T & Stavans J (2006) High fidelity of RecA-catalyzed recombination: a watchdog of genetic diversity. *Nucleic Acids Res* **34**: 5021–5031.
- Saha S, Bridges S, Magbanua ZV & Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res* **36**: 2284–2294.
- Santoyo G & Romero D (2005) Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol Rev* **29**: 169–183.
- Schaaper RM (1988) Mechanisms of mutagenesis in the *Escherichia coli* mutator mutD5: role of DNA mismatch repair. *P Natl Acad Sci USA* **85**: 8126–8130.
- Schmid-Appert M, Zoller K, Traber H, Vuilleumier S & Leisinger T (1997) Association of newly discovered IS elements with the dichloromethane utilization genes of methylotrophic bacteria. *Microbiology* **143**: 2557–2567.
- Schneider D, Duperchy E, Depuyrot J, Coursange E, Lenski R & Blot M (2002) Genomic comparisons among *Escherichia coli* strains B, K-12, and O157:H7 using IS elements as molecular markers. *BMC Microbiol* **2**: 18.
- Sedgwick SG & Smerdon SJ (1999) The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem Sci* **24**: 311–316.
- Sekine Y, Eisaki N & Ohtsubo E (1994) Translational control in production of transposase and in transposition of insertion sequence IS3. *J Mol Biol* **235**: 1406–1420.
- Serebrovsky A (1938) Genes scute and achaete in *Drosophila melanogaster* and a hypothesis of gene divergency. *CR Acad Sci URSS* **19**: 77–81.
- Shapiro JA (1999) Genome system architecture and natural genetic engineering in evolution. *Ann NY Acad Sci* **870**: 23–35.
- Sharples GJ & Lloyd RG (1990) A novel repeated DNA sequence located in the intergenic regions of bacterial genomes. *Nucleic Acids Res* **18**: 6503–6508.

- Sharples GJ, Ingleston SM & Lloyd RG (1999) Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. *J Bacteriol* **181**: 5543–5550.
- Shen P & Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112**: 441–457.
- Shen X, Gumulak J, Yu H, French CT, Zou N & Dybvig K (2000) Gene rearrangements in the *vsa* locus of *Mycoplasma pulmonis*. *J Bacteriol* **182**: 2900–2908.
- Shigenobu S, Watanabe H, Hattori M, Sakaki Y & Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**: 81–86.
- Shyamala V, Schneider E & Ames GF (1990) Tandem chromosomal duplications: role of REP sequences in the recombination event at the join-point. *EMBO J* **9**: 939–946.
- Siguier P, Filee J & Chandler M (2006a) Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* **9**: 526–531.
- Siguier P, Perochon J, Lestrade L, Mahillon J & Chandler M (2006b) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**: D32–D36.
- Simon DM, Clarke NA, McNeil BA *et al.* (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* **14**: 1704–1713.
- Singer BS & Westlye J (1988) Deletion formation in bacteriophage T4. *J Mol Biol* **202**: 233–243.
- Sirand-Pugnet P, Lartigue C, Marends M *et al.* (2007) Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. *PLoS Genet* **3**: e75.
- Slack A, Thornton PC, Magner DB, Rosenberg SM & Hastings PJ (2006) On the mechanism of gene amplification induced under stress in *Escherichia coli*. *PLoS Genet* **2**: e48.
- Slechts ES, Bunney KL, Kugelberg E, Kofoed E, Andersson DI & Roth JR (2003) Adaptive mutation: general mutagenesis is not a programmed response to stress but results from rare coamplification of *dinB* with *lac*. *P Natl Acad Sci USA* **100**: 12847–12852.
- Smith MD, Lennon E, McNeil LB & Minton KW (1988) Duplication insertion of drug resistance determinants in the radioresistant bacterium *Deinococcus radiodurans*. *J Bacteriol* **170**: 2126–2135.
- Snel B, Bork P & Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**: 17–25.
- Sniegowski PD, Gerrish PJ, Johnson T & Shaver A (2000) The evolution of mutation rates: separating causes from consequences. *Bioessays* **22**: 1057–1066.
- Sokol D, Benson G & Tojeira J (2007) Tandem repeats over the edit distance. *Bioinformatics* **23**: e30–e35.
- Sokurenko EV, Hasty DL & Dykhuizen DE (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol* **7**: 191–195.
- Sorek R, Kunin V & Hugenholtz P (2008) CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181–186.
- Sreenu VB, Alevoor V, Nagaraju J & Nagarajaram HA (2003) MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res* **31**: 106–108.
- Srividhya KV, Alaguraj V, Poornima G *et al.* (2007) Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS ONE* **2**: e1193.
- Stern MJ, Ames GF, Smith NH, Robinson EC & Higgins CF (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell* **37**: 1015–1026.
- Stern MJ, Prossnitz E & Ames GF (1988) Role of the intercistronic region in post-transcriptional control of gene expression in the histidine transport operon of *Salmonella typhimurium*: involvement of REP sequences. *Mol Microbiol* **2**: 141–152.
- Steward A, Adhya S & Clarke J (2002) Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J Mol Biol* **318**: 935–940.
- Stumpf JD, Poteete AR & Foster PL (2007) Amplification of *lac* cannot account for adaptive mutation to *Lac*<sup>+</sup> in *Escherichia coli*. *J Bacteriol* **189**: 2291–2299.
- Swanson J, Bergstrom S, Robbins K, Barrera O, Corwin D & Koomey JM (1986) Gene conversion involving the pilin structural gene correlates with pilus<sup>+</sup> in equilibrium with pilus<sup>-</sup> changes in *Neisseria gonorrhoeae*. *Cell* **47**: 267–276.
- Szklarczyk R & Heringa J (2004) Tracking repeats using significance and transitivity. *Bioinformatics* **20** (suppl 1): i311–i317.
- Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH & Godelle B (1997) Role of mutator alleles in adaptive evolution. *Nature* **387**: 700–702.
- Taneda A (2004) Adplot: detection and visualization of repetitive patterns in complete genomes. *Bioinformatics* **20**: 701–708.
- Taylor JS & Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**: 615–643.
- Teichmann SA, Park J & Chothia C (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *P Natl Acad Sci USA* **95**: 14658–14663.
- Temin HM (1989) Reverse transcriptases. Retrons in bacteria. *Nature* **339**: 254–255.
- Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S & McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* **11**: 1441–1452.
- Tenaillon O, Taddei F, Radman M & Matic I (2001) Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res Microbiol* **152**: 11–16.
- Tobes R & Pareja E (2006) Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics* **7**: 62.
- Touchon M & Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* **24**: 969–981.



- Tran-Nguyen LT, Kube M, Schneider B, Reinhardt R & Gibb KS (2008) Comparative genome analysis of "Candidatus *Phytoplasma australiense*" (subgroup tuf-Australia I; rp-A) and "Ca. *Phytoplasma asteris*" Strains OY-M and AY-WB. *J Bacteriol* **190**: 3979–3991.
- Treangen TJ, Darling AE, Achaz G, Ragan MA, Messegueur X & Rocha EPC (2009) A novel heuristic for local multiple alignment of interspersed DNA repeats. *IEEE/ACM Trans Comput Biol BioInf*, in press.
- Tripp KW & Barrick D (2004) The tolerance of a modular protein to duplication and deletion of internal repeats. *J Mol Biol* **344**: 169–178.
- Turlan C & Chandler M (1995) IS1-mediated intramolecular rearrangements: formation of excised transposon circles and replicative deletions. *EMBO J* **14**: 5410–5421.
- Valens M, Penaud S, Rossignol M, Cornet F & Boccard F (2004) Macrodome organization of the *Escherichia coli* chromosome. *EMBO J* **23**: 4330–4341.
- van Belkum A, Scherer S, van Alphen L & Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol R* **62**: 275–293.
- Van Valen L (1973) A new evolutionary law. *Evol Theor* **1**: 1–30.
- Venditti R, De Gregorio E, Silvestro G, Bertocco T, Salza MF, Zarrilli R & Di Nocera PP (2007) A novel class of small repetitive DNA sequences in *Enterococcus faecalis*. *FEMS Microbiol Lett* **271**: 193–201.
- Vogel C, Bashton M, Kerrison ND, Chothia C & Teichmann SA (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struc Biol* **14**: 208–216.
- Volff JN & Altenbuchner J (1998) Genetic instability of the *Streptomyces* chromosome. *Mol Microbiol* **27**: 239–246.
- Volfovsky N, Haas BJ & Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biol* **2**: 0027.0021–0027.0011.
- Vulic M, Dionisio F, Taddei F & Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *P Natl Acad Sci USA* **94**: 9763–9767.
- Wagner A (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol Biol Evol* **23**: 723–733.
- Wagner A, Lewis C & Bichsel M (2007) A survey of bacterial insertion sequences using IScan. *Nucleic Acids Res* **35**: 5284–5293.
- Waterman MS & Vingron M (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *P Natl Acad Sci USA* **91**: 4625–4628.
- Wildschutte H, Wolfe DM, Tamewitz A & Lawrence JG (2004) Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonella*. *P Natl Acad Sci USA* **101**: 10644–10649.
- Wilson TE, Topper LM & Palmbo PL (2003) Non-homologous end-joining: bacteria join the chromosome breakdance. *Trends Biochem Sci* **28**: 62–66.
- Wren BW (1991) A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. *Mol Microbiol* **5**: 797–803.
- Wright CE, Teichmann SA, Clarke J & Dobson CM (2005) The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* **438**: 878–881.
- Wu TH & Marinus MG (1999) Deletion mutation analysis of the mutS gene in *Escherichia coli*. *J Biol Chem* **274**: 5948–5952.
- Xu J, Bjursell MK, Himrod J *et al.* (2003) A genomic view of the human–bacteroides thetaiotaomicron symbiosis. *Science* **299**: 2074–2076.
- Yang F, Yang J, Zhang X *et al.* (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res* **33**: 6445–6458.
- Yang Y & Ames GF (1988) DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. *P Natl Acad Sci USA* **85**: 8850–8854.
- Zambrano MM, Siegle DA, Almiron M, Tormo A & Kolter R (1993) Microbial competition: *Escherichia coli* mutants that take over stationary phase cultures. *Science* **259**: 1757–1760.
- Zhang JR, Hardham JM, Barbour AG & Norris SJ (1997) Antigenic variation in Lyme disease borreliae by promiscuous recombination of VMP-like sequence cassettes. *Cell* **89**: 275–285.
- Zhang Y & Waterman MS (2003) An Eulerian path approach to global multiple alignment for DNA sequences. *J Comput Biol* **10**: 803–819.