The role of variable DNA tandem repeats in bacterial adaptation

Kai Zhou, Abram Aertsen & Chris W. Michiels

Department of Microbial and Molecular Systems (M²S), Faculty of Bioscience Engineering, Laboratory of Food Microbiology and Leuven Food Science and Nutrition Research Centre (LForCe), KU Leuven, Leuven, Belgium

Abstract

DNA tandem repeats (TRs), also designated as satellite DNA, are inter- or intragenic nucleotide sequences that are repeated two or more times in a head-to-tail manner. Because TR tracts are prone to strand-slippage replication and recombination events that cause the TR copy number to increase or decrease, loci containing TRs are hypermutable. An increasing number of examples illustrate that bacteria can exploit this instability of TRs to reversibly shut down or modulate the function of specific genes, allowing them to adapt to changing environments on short evolutionary time scales without an increased overall mutation rate. In this review, we discuss the prevalence and distribution of inter- and intragenic TRs in bacteria and the mechanisms of their instability. In addition, we review evidence demonstrating a role of TR variations in bacterial adaptation strategies, ranging from immune evasion and tissue tropism to the modulation of environmental stress tolerance. Nevertheless, while bioinformatic analysis reveals that most bacterial genomes contain a few up to several dozens of intra- and intergenic TRs, only a small fraction of these have been functionally studied to date.

Introduction

To cope with rapidly changing environmental conditions and ensure their survival, unicellular organisms have evolved a plethora of adaptation strategies (Aertsen & Michiels, 2004, 2005). Most of these strategies are based on transient alterations in gene expression in response to stressful conditions, and well-known examples include the SOS response (controlled by RecA and LexA), the general stress response (regulated by the sigma factor RpoS), the stringent response (mediated by pppGpp and ppGpp), and the heat shock response (mainly controlled by RpoH; Massey & Buckling, 2002; Foster, 2005; Saint-Ruf & Matic, 2006; Foster, 2007; Jolivet-Gougeon et al., 2011). In addition, adaptation can also stem from the acquisition of stochastic mutations that alter the genotype, which become positively selected and fixed in a population if they coincide with a beneficial phenotype (Rando & Verstrepen, 2007). However, an important drawback of the latter strategy is that random mutations are more often deleterious than beneficial.

Interestingly, neither the type nor the frequency of mutational events is randomly distributed over the genome, and some DNA sequences have evolved to be mutational hotspots that drive the variability of genes whose activity can impact the adaptive potential of their host. One type of such special sequences that is very abundant in prokaryotic and eukaryotic genomes is known as tandem repeats (TRs), a major class of direct DNA repeats. While at first TRs were considered to be junk DNA without any biological function, studies of the human genome have revealed some of these repetitive sequences to be hypermutable and the cause of diseases such as fragile X syndrome, spinobulbar muscular atrophy, and huntington disease (Hannan, 2010). In addition, accumulating evidence points out the potential role of TRs as engines of genetic variability and bacterial adaptation. In this review, we therefore focus on the
Identification and distribution of TRs in bacteria, the mechanisms behind their variability, and their biological significance in bacterial adaptation.

**Identification and distribution of TRs in bacterial genomes**

**Definition of TRs**

TRs are nucleotide sequences that are directly repeated in a head-to-tail manner. According to the conservation of the repeated sequence, TRs are classified as identical/perfect TRs or degenerated/imperfect TRs, respectively (Fig. 1). Furthermore, TRs are commonly classified into three categories according to their repeat unit size, although there is no consensus definition of these categories (Richard et al., 2008). Repeats with unit size varying from 1 to 9, 10 to 100, and > 100 bp are termed microsatellites, minisatellites, and macrosatellites, respectively (Lopes et al., 2006). The term 'satellite DNA' originally refers to the very large arrays of tandemly repeated non-coding DNA (often hundreds of copies) that are characteristic of large eukaryotic genomes, but, in the context of bacterial genomes, is also used to include small and intragenic TRs.

**In silico identification of TRs**

The increasing availability of genome sequences and specialized bioinformatics software greatly facilitates the search and identification of TR loci on a genomewide scale, which obviously is a prerequisite for understanding their distribution, predicting their function, and tracking their evolution. A variety of algorithms have been developed for detecting TRs, but it is important to be aware that these may differ in their ability to detect different types of TRs (Merkel & Gemmell, 2008; Treangen et al., 2009; Kajava, 2012). Hence, the choice of search tool should be determined by the TR type of interest, or the parallel use of several algorithms is advisable when a wide screen for TRs is performed. Furthermore, parameter settings (i.e. alignment weights, definition of repeats, and threshold scores) can also strongly affect outcome in terms of number and consensus motif of TRs (Lim et al., 2013). In particular, problems are still commonly encountered in detecting imperfect TRs (Leclercq et al., 2007; Schaper et al., 2012). Several algorithms are freely available online, such as Tandem Repeat Finder (Benson, 1999) and IMEx (Mudunuri & Nagarajaram, 2007). In addition, several databases of annotated TRs in prokaryotes have been established, such as TRs DB (http://minisatellites.u-psud.fr), PSSRDb (http://pssrdb.cdfd.org.in) and MICAS (http://180.149.48.108/micas/index.php). In the next section, we will review the major findings from some recent in silico studies of the genomic distribution of TRs in bacteria.

**The distribution of TRs in bacterial genomes**

The analysis of TRs in bacterial genomes so far has mainly focused on microsatellites with unit size 1–6 bp, also termed ‘simple sequence repeats (SSRs)’. A number of general observations regarding the distribution of SSRs can be stated. First of all, the abundance of SSRs in bacteria is lower than that in eukaryotes (Schlotterer et al., 2006). Nevertheless, the number of SSRs is orders of magnitude smaller than in eukaryotes. A variety of algorithms have been developed for detecting TRs, but it is important to be aware that these may differ in their ability to detect different types of TRs (Merkel & Gemmell, 2008; Treangen et al., 2009; Kajava, 2012). Hence, the choice of search tool should be determined by the TR type of interest, or the parallel use of several algorithms is advisable when a wide screen for TRs is performed. Furthermore, parameter settings (i.e. alignment weights, definition of repeats, and threshold scores) can also strongly affect outcome in terms of number and consensus motif of TRs (Lim et al., 2013). In particular, problems are still commonly encountered in detecting imperfect TRs (Leclercq et al., 2007; Schaper et al., 2012). Several algorithms are freely available online, such as Tandem Repeat Finder (Benson, 1999) and IMEx (Mudunuri & Nagarajaram, 2007). In addition, several databases of annotated TRs in prokaryotes have been established, such as TRs DB (http://minisatellites.u-psud.fr), PSSRDb (http://pssrdb.cdfd.org.in) and MICAS (http://180.149.48.108/micas/index.php). In the next section, we will review the major findings from some recent in silico studies of the genomic distribution of TRs in bacteria.

**Fig. 1.** Schematic representation of different types of TRs. (a) Different conservation of repeat unit sequence. (b) Different sizes of repeat unit. Space between repeat units has only been introduced to improve visual clarity of the figure.
magnitude higher than that of other repeat types (i.e. minisatellite and macrosatellite) in the genomes of most, if not all bacteria. Of course, this is not unexpected because this SSR count included even mononucleotide trimers (e.g. AAA), which account for about 70% of the total number of SSRs. While SSRs are generally believed to contribute to genome polymorphism and adaptation potential of bacteria (Kassai-Jäger et al., 2008), the contribution of these very small SSRs like mononucleotide trimers is probably limited. In fact, a rough threshold of minimum TR unit number (4–9) has been noted, below which a SSR is not likely to mutate or be variable (Lai & Sun, 2003; Dettman & Taylor, 2004; Kelkar et al., 2010).

Intriguingly, heptameric repeats were found to be overrepresented among these SSRs in most prokaryotes, and it was hypothesized that the seemingly preferred 7 bp length of a repeat unit might relate to the DNA segment size that interacts with the active site of the DNA polymerase, thus facilitating the occurrence of polymerase slippage (Mrážek et al., 2007).

A remarkable feature of SSRs is their widely diverse distribution across species, even closely related ones, and this may indicate that they are subject to rapid evolutionary change (Yang et al., 2003; Mrážek, 2006; Kassai-Jäger et al., 2008). Analysis of more than 300 prokaryotic genomes showed that the distribution of SSRs varied with the bacterial species, genome size, and G + C content (Mrážek et al., 2007). More specifically, SSRs with small motif (1–4 bp) are more abundant in small genomes and particularly in host-adapted pathogens with reduced genomes (< 2 Mb) and low G + C content (< 40%), such as Mycoplasma and Haemophilus spp. (Moxon et al., 2006; Treangen et al., 2009). In contrast, SSRs with a larger motif (5–11 bp) are more frequent in nonpathogens and opportunistic pathogens with large genomes (> 4 Mb) and high G + C content (> 60%), such as Burkholderia and Anaerobaculum spp. Based on this observation, it was hypothesized that the differential representations of SSRs in bacteria may correlate with pathogenicity, but more work is needed to corroborate this. Another interesting observation is that some relatively large bacterial genomes (e.g. Pseudomonas aeruginosa, c. 5 Mb) have fewer SSRs than would be predicted based on their genome properties, but harbor comparatively more two-component sensor transducers. In contrast, some host-adapted pathogens with small genome size (i.e. Haemophilus influenzae, Neisseria meningitidis, and Helicobacter pylori) have comparatively more SSRs, but less two-component sensor transducers (Moxon et al., 2006). Thus, it seems that environmental adaptability in host-adapted pathogens depends primarily on SSR variations, while in opportunistic pathogens with a more versatile lifestyle it depends primarily on two-component sensor transducers.

Closer examination of the SSR distribution across the genome shows significant differences in coding and non-coding regions. Because bacterial genomes are more compact than those of eukaryotes, they have comparatively more intragenic than intergenic SSRs. For example, in Escherichia coli K-12, 79.5% of SSRs locate in coding regions (Gur-Arie et al., 2000), whereas in the genome of the Japanese pufferfish (Fugu rubripes), only 11.6% of SSRs are intragenic (Edwards et al., 1998). Generally, long mono- and dinucleotide SSRs are excluded from coding regions, probably because they have a higher probability to rearrange and cause frameshift mutations in genes (Conenye & Vandamme, 2005; Ackermann & Chao, 2006; Orsi et al., 2010; Lin & Kussell, 2012). In contrast, SSRs whose unit size is a multiple of three nucleotides (3, 6, 9 …) are overrepresented in open reading frames (ORFs) because their expansion or contraction does not disrupt the reading frame (Mrážek et al., 2007). However, exceptions have been reported. For example, tetranucleotide SSRs of H. influenzae are exclusively found in ORFs, which is consistent with their role in phase variation (Power et al., 2009). An interesting situation exists in the mycoplasmas, where long trinucleotide repeats are overrepresented in Mycoplasma genitalium, Mycoplasma gallisepticum, and Mycoplasma hyopneumoniae, but occur mainly in intergenic regions in the former two species, but in coding regions in the latter one (Mrážek, 2006). This difference in distribution is also reflected in different functional roles. In M. gallisepticum, the most prominent trinucleotide TRs are the GAA repeats in the 5′ untranslated region of the 42 up to 70 vlpA adhesin gene paralogs that exist in each strain, which regulate vlpA gene expression (Glew et al., 1998, 2000; Liu et al., 2000; Papazisi et al., 2003). In contrast, M. hyopneumoniae trinucleotide repeats are found mostly within hypothetical ORFs, but also in some adhesins, and their contraction or expansion results in variability of amino acid repeats that are believed to play a role in protein–protein interaction or adhesion (Mrážek, 2006).

A more detailed study on the occurrence of intragenic TRs in 44 bacteria and archaea revealed additional features (Lin & Kussell, 2012). Intragenic SSRs were found more frequently near the termini (5′ and 3′ ends) of the ORF rather than in the middle, which most likely stems from biophysical constraints of protein structure. In addition, SSR-induced frameshifts at the 3′ end are less harmful than at other parts, because most of the upstream coding region will not be affected. Nevertheless, an overrepresentation of SSRs was found in the 5′ end in ORFs of pathogens, probably because this allows SSR-induced frameshifts to function as an ON/OFF switch for these ORFs, which can be advantageous for pathogens because it facilitates rapid adaptation of populations. Similar observations had already been made earlier for...
some intragenic mononucleotide repeats at the 5′ end of genes (van Passel & Ochman, 2007; Janulczyk et al., 2010; Orsi et al., 2010). However, it remains unclear and often difficult to prove whether this type of distribution bias of intragenic SSRs is linked to selection pressures in bacteria. An argument in favor of such a link is that intragenic SSRs show a preference for certain categories of genes. In both Gram-negative (e.g. Haemophilus and Helicobacter) and Gram-positive (e.g. Streptococcus) pathogens, SSR-associated genes frequently encode virulence factors, cell surface components, and restriction–modification enzymes (van Belkum, 1999; Moxon et al., 2006; Guo & Mrázek, 2008; Power et al., 2009; Janulczyk et al., 2010). On the other hand, several intragenic SSRs with numerous repeat copies and a unit size that is not a multiple of three are also found in housekeeping genes whose products are essential for important cellular processes, such as cell division, energy production, and DNA replication and repair (Guo & Mrázek, 2008). Obviously, corresponding TR rearrangements leading to reading frame disruption are anticipated to be detrimental or even lethal for the cell, and it remains unclear why such TRs have been maintained during evolution.

Intergenic SSRs also show a nonrandom distribution, being found more frequently in the immediate vicinity of genes than at distant positions. For example, intergenic SSRs of E. coli K-12 concentrate in a region up to 200 bp from the start codon, which contains proximal regulators of gene expression (Gur-Arie et al., 2000). Another study showed that in most cases, the intergenic SSRs with numerous copies are located upstream of the first gene in prokaryotic operons (Guo & Mrázek, 2008). Together, both studies reflect the potential role of intergenic SSRs in the regulation of gene expression.

The variability of TRs

The variability of TRs is thought to be one of the drivers of genomic plasticity. The regions containing TRs are potentially hypermutable by contraction (deletion) or expansion (insertion) of TR units, and mutation frequencies up to 10⁻¹⁴ have been reported in bacteria (Rando & Verstrepen, 2007). Not surprisingly, the polymorphisms found in TR loci provide a foundation for DNA genotyping approaches such as variable number TR-based typing or multilocus variable repeat analysis, which are commonly applied for pathogen typing (Lindstedt, 2005, 2011; reviewed in Chiou, 2010).

The molecular mechanisms of TR variation

Based on extensive studies in both plasmid-based and chromosomal systems, two nonexclusive mechanisms, replication slippage and recombination, are currently widely accepted to explain TR variation (Pearson et al., 2005; Bichara et al., 2006; Gemayel et al., 2010). With regard to the slippage mechanism, several models have been proposed, and the most common one is strand-slippage mispairing (SSM), also called DNA slippage or polymerase slippage (Kornberg et al., 1964; Streisinger et al., 1966). This model proposes that TR rearrangements result from strand slippage during DNA replication. The process is initiated by the formation of a bulge structure of unpaired repeats either on the template or on the nascent strand. If the bulge is present on the template strand, it will result in a TR contraction (deletion) in the newly synthesized DNA. In contrast, TR expansion (insertion) will result when a bulge forms on the nascent strand (Fig. 2). Currently, substantial evidence supports the involvement of replication in TR instability in bacteria. In E. coli, for example, triplet nucleotide tracts in a plasmid or in the chromosome were dramatically destabilized in a dnaQ49 mutant. The dnaQ gene encodes the 3′–5′ exonuclease ε-subunit of DNA polymerase III, which is involved in proofreading, and it was therefore suggested that this mutant failed to correctly remove slipped structures in the TR tracts during replication (Iyer et al., 2000; Zahra et al., 2007). Moreover, mutations in the α subunit (encoded by dnaE) of DNA polymerase III holoenzyme, γ and τ subunits of the clamp-loading complex (encoded by dnaX), and β clamp (encoded by dnaN) have also been shown to increase instability of microsatellites and tandemly repeated DNA sequences (reviewed in Bichara et al., 2006). In general, the effect of DNA replication on TR rearrangements provides evidence for the replication slippage mechanism, because a prerequisite of this mechanism is that DNA replication is stalled by the secondary loop structures formed by TRs. Notably, the SSM mechanism is not only widely accepted to account for the majority of SSR variations, it is also invoked to explain the genesis of TRs from unrepeated DNA (Levinson & Gutman, 1987; Waite et al., 2003; Lindbäck et al., 2011).

Besides SSM, recombination is considered as another mechanism of TR instability that could explain phenomena that cannot be explained by the slippage mechanism. Generally, recombination is more important for the rearrangement of TRs with large unit size, whereas SSM is the dominant mechanism underlying variation of TRs with small unit size (Bi & Liu, 1996; Richard & Pâques, 2000; Bzymek & Lovett, 2001; Rocha, 2003; Gemayel et al., 2010). Both homologous (RecA-dependent) and illegitimate (RecA-independent) recombination can be involved. Several models have been proposed for the recombination mechanism, such as unequal crossover and intramolecular recombination (Fig. 3), and evidence for...
the involvement of recombination in TR variations is accumulating. For example, it has been suggested that double-strand DNA break (DSB) repair can induce TR instability via homologous recombination in prokaryotes and eukaryotes (Hebert & Wells, 2005; reviewed in Richard et al., 2008; Malkova & Haber, 2012), and mechanistic models explaining this process have been elaborated. One such model is the DSB repair slippage model, which combines the double Holliday junction intermediate pathway with the strand-slippage model (Richard et al., 2008). In an alternative explanation, the synthesis-dependent strand annealing pathway also contributes to the TR rearrangements mediated by DSB repair (Pâques et al., 1998, 2001; Richard et al., 1999; Richard & Pâques, 2000). This model is proposed for explaining TR rearrangements that are rarely associated with crossover events (Pâques et al., 1998, 2001; Richard et al., 1999).

**Factors influencing TR rearrangement frequencies**

While TRs generally are intrinsically prone to incur contraction or expansion, the actual frequency of these events can vary widely depending on both intrinsic (structural) and extrinsic (environmental) factors. Regarding intrinsic TR features, a positive correlation has been established between TR copy number and rearrangement frequency in several studies. Through in silico genome analysis of multiple strains of 42 fully sequenced prokaryotic species, Lin & Kussell (2012) observed that the variability of three types of SSRs (monomeric, dimeric, and trimeric repeats) increased dramatically with the number of repeat units. In another study, an exponential relation between the number of repeat units and rearrangement frequency was observed in a comparison of 30 artificially constructed TRs (unit lengths of 2, 10, and 20 nucleotides; number of units between 2 and 50; sequence conservation between 62.5% and 100%; Legendre et al., 2007). This also accounts for the fact that long TRs are relatively uncommon (Lai & Sun, 2003). Similar findings have been reported in several other studies (Goldstein & Clark, 1995; Brinkmann et al., 1998; Lai & Sun, 2003; Vogler et al., 2006). Likewise, a positive relationship is generally found between TR mutation frequency and the size of the repeat unit (Sia et al., 1997; Schug et al., 1998; Eckert & Hile, 2009; Bayliss et al., 2012) as well as the degree of
conservation between repeats (Legendre et al., 2007). In addition, the GC content of the TR is also an important factor determining stability, because repeated sequences are prone to form diverse non-B DNA structures (i.e. hairpins and triplexes), which may cause pausing of the DNA polymerase and replication fork collapse, and in turn necessitate intervention of the repair and recombination machinery to reinitiate replication (Wells et al., 2005; Choudhary & Trivedi, 2010). Besides, the orientation of the repeats with respect to the direction of replication can affect the mutation ratio as well, because TRs are more prone to form secondary structures in one orientation than in the other (Hebert et al., 2004), and replication fidelity is not equal in the leading strand compared with the lagging strand (Gawel et al., 2002). Intriguingly, based on a theoretical model, Lai & Sun (2003) suggested that expansion occurs more frequently for short microsatellites, while contraction occurs more frequently for long ones, suggesting the rearrangement pattern might be dependent on the repeat type. However, the lack of experimental evidence so far cannot validate this observation.

Aside from intrinsic factors, extrinsic environmental conditions may also affect TR rearrangement frequencies. For instance, it was shown that several TRs used in a multilocus TR typing scheme of E. coli showed enhanced variation at increased growth temperature and upon starvation in E. coli O157:H7, but not upon irradiation (Cooley et al., 2010). Likewise, in sclB of Streptococcus pyogenes, encoding a collagen-like surface protein, TR variations occurred during growth in fresh human blood, but not in medium (Rasmussen & Björk, 2001). However, the underlying mechanisms by which environmental stresses affect the TR mutation frequency are poorly understood.

The phenotypical impact of intergenic TR variations in bacteria

Accumulating evidence indicates that rearrangements of intergenic TRs can confer transcriptional evolvability (Jansen et al., 2012). More specifically, SSRs positioned as cis-regulatory elements around the promoter region can induce phase variation (i.e. stochastic, high frequency, reversible switching of genotype, and/or phenotype) by modulating the transcription of the corresponding genes. Most studies indicate that intergenic SSRs, except monomeric SSRs, involved in phase variation tend to be A/T rich, which makes them prone to melting and SSM. In this section, we review examples of this mechanism of phase variation according to the different positions of intergenic TRs relative to the transcriptional start site (Fig. 4; Table 1).
Variable tandem repeats and bacterial adaptation

TRs upstream of −35 site of promotor

When TRs are present upstream of the −35 site, the repeat copy number is expected to affect the binding of transcription factors and thus modulate gene expression. This has been reported in the pathogen *N. meningitidis*, where expression of the *nadA* gene (encoding a protein that functions as invasin and adhesin) is regulated by the number of tetrameric repeats (TAAA) within the upstream region of the RNA polymerase binding site, resulting in three significantly distinct transcriptional levels (high, intermediate, and low). The frequency of phase variation between these levels has been estimated at $c. 4.4 \times 10^{-4}$. Remarkably, the levels of *nadA* transcription show a periodic rather than a monotonous relation to the number of repeat units. As such, the transcription level varies in the order low–high–intermediate–low–high for repeat copy numbers of 4–5–6–7–8 and again for 9–10–11–12–13 repeats. Further mechanistic studies indicated that variation of the tetrameric repeats affects the binding of the integration host factor transcriptional regulator protein to the *nadA* promoter (Martin et al., 2003, 2005). More recently, Metruccio et al. (2009) discovered that depending on the number of TAAA repeats, a novel repressor (NadR) prevents transcription of *nadA* through binding of two operators flanking the variable tetrameric repeat tract on both sides. As a result, it was proposed that alteration of the spacing between these two operators by variation of the number of TAAA repeats may affect the ability of NadR to repress *nadA* expression (Metruccio et al., 2009).

TRs between −35 and −10 sites of promotor

The spacing between the −35 and −10 sites of a standard promoter is critical for the binding efficiency of RNA polymerase (Fig. 4), and the optimal distance is around 17 bp in bacteria. Consequently, when TRs locate in that region, copy number changes are expected to modulate the transcription level of the genes in the transcriptional unit. An example of this is found in the host-adapted pathogen *H. influenzae*, which adheres to human epithelial cells with the help of LKP fimbriae (also called long, thick pili). These fimbriae are encoded by the *hif* gene, which is important for *H. influenzae* infection at different stages (van Ham et al., 1993). Phase variation in the expression of these fimbriae is mediated by a string of dinucleotide repeats between the −35 and −10 sites within the overlapping, but divergent promoter regions of *hifA* and *hifB* genes, with TR copy numbers of 9, 10, or 11, respectively, resulting in no, high, and low expression of both genes.

Another intriguing example is the FetA protein of *Neisseria gonorrhoeae*, an iron-repressible protein functioning as ferric enterobactin receptor. The expression of FetA exhibits extremely rapid phase variation (switching frequency up to 1.3% per generation) correlating with polymorphism of a poly-C tract between the −10 and −35 regions of the *fetA* promoter. The various lengths of the poly-C tract result in either high or low expression.

**Table 1.** Overview of mechanisms by which intergenic TRs can modulate gene expression

<table>
<thead>
<tr>
<th>Location of TRs</th>
<th>Mechanisms</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upstream of −35 site</td>
<td>Affects transcription initiation by modifying binding affinity of regulatory proteins</td>
<td>Miller et al. (1987), Martin et al. (2003, 2005), Metruccio et al. (2009)</td>
</tr>
<tr>
<td>Between −35 and −10 sites</td>
<td>Affects transcription initiation by altering the distance of promoter elements</td>
<td>Willems et al. (1990), Yogev et al. (1991), van Ham et al. (1993), Sarkari et al. (1994), Carson et al. (2000), van der Ende et al. (2009)</td>
</tr>
<tr>
<td>Between transcriptional start and ORF</td>
<td>May modify binding affinity of regulatory proteins or mRNA stability</td>
<td>Lafontaine et al. (2001), Attia &amp; Hansen (2006)</td>
</tr>
<tr>
<td>Between two separate transcription start sites</td>
<td>Unknown</td>
<td>Dawid et al. (1999)</td>
</tr>
</tbody>
</table>

ORF, open reading frame; TR, tandem repeat.
It was suggested that phase variation of FetA reflects a balance between the advantages of iron scavenging, on the one hand, and evasion of the host immune response (based on FetA immunogenicity), on the other (Carson et al., 2000).

A more complex situation is found in the PorA outer membrane protein in N. meningitidis. Not only is expression of the porA gene stochastically modulated by two variable homopolymeric tracts (poly-G and poly-T) between the −35 and −10 sites of the promoter region, there is also a variable poly-A tract within the porA coding region. Both sites are believed to serve evasion of the host immune response to this protein and explain the poor efficacy observed for PorA-based vaccines (van der Ende et al., 2000).

Further examples include the genes encoding the lipoprotein Vbps of Mycoplasma hyorhinis (Yoge et al., 1991), the fimbrial subunits of Bordetella pertussis (Willemms et al., 1990), and the outer membrane protein Opc of N. meningitidis (Sarkari et al., 1994), and it can be concluded that variable SSRs within the −35 and −10 region represent a widespread strategy of phase variation by modulated gene expression in pathogens.

TRs between transcriptional start site and ORF

While intergenic TR-dependent phase variations in bacteria mostly belong to the two classes described above, some cases are mediated by TRs located between the transcriptional start site and the ORF. In the Gram-negative pathogen Moraxella catarrhalis, the UspA1 protein functions as an adhesin to mediate binding to human epithelial cells (Lafontaine et al., 2000). Its expression was shown to be phase variable and correlated with adherence capacity. Sequence analysis revealed a variable homopolymeric poly-G tract between the transcriptional start site and the start codon to be responsible for the variability in uspA1 expression. Stable mRNA and strong expression of UspA1 was detected with a 10-bp G repeat tract, while truncated mRNA and weak expression occurred with a 9-bp G tract. Based on these observations, it was proposed that alterations in the poly-G tract affect the binding efficiency of transcriptional regulators and/or the stability of the uspA1 mRNA (Lafontaine et al., 2001).

Interestingly, a similar case was recently uncovered in another outer membrane protein of M. catarrhalis, UspA2, which is involved in serum resistance and vitronectin binding (Attia & Hansen, 2006). It was observed that a tetrameric (AGAT) repeat tract, located between the uspA2 transcription start site and the start codon, is highly variable in M. catarrhalis isolates. Moreover, the expression level of UspA2 and, as a result, the serum resistance and vitronectin binding capacity of the cells were increased as the repeat copy number increased, possibly because the variable number of AGAT repeats may affect the secondary structure of the UspA2 mRNA transcript.

TRs between two separated transcription start sites

This special example of phase switching induced by variable intergenic TRs has so far only been described in H. influenzae (Dawid et al., 1999). HMW1 and HMW2 are two adhesins of H. influenzae that exhibit different cellular binding specificities and are encoded by two separate chromosomal loci hmw1AC and hmw2AC, respectively. Both genes show 80% similarity, suggesting that they can be considered as alleles (Barenkamp & Leininger, 1992; Eccevit et al., 2004). Promoter analysis revealed two transcription start sites (P1 and P2) within the upstream region of each gene and a heptameric (ATCTTTC) TR array between P1 and P2 of both genes. The occurrence of variations in the number of TR copies has been confirmed in H. influenzae isolates from patients with chronic obstructive pulmonary disease (Dawid et al., 1999; Cholon et al., 2008). These variations can affect mRNA synthesis and thereby influence the expression of corresponding proteins. For example, an increasing number of TR units resulted in a gradual decrease in specific mRNA synthesis and protein expression and vice versa. However, the underlying molecular mechanism remains obscure.

The phenotypical impact of intragenic TRs in bacteria

There is abundant evidence that besides intergenic TRs, also intragenic TRs can trigger phase variation (Table 2). However, the underlying mechanisms are dependent on the nature of the TR. More specifically, if the TR unit size is not a multiple of three, rearrangements are able to induce frameshift mutations as the cause of ON–OFF phase variation. In comparison, phase variation induced by TRs whose unit size is a multiple of three is more complex and is probably related to specific structural and functional alterations of the corresponding proteins (Gemayel et al., 2010). In this section, we review studies on the phenotypical impact of intragenic TRs, grouped according to their location in different functional classes of genes.

Cell surface structural genes with TRs

It has been noted that TRs are most abundant in genes whose products are either exposed on the cell surface or involved in the biogenesis of cellular surface structures,
such as lipopolysaccharides (LPS), adhesins, pili, fimbriae, and capsules (Moxon et al., 1994; Jordan et al., 2003; Verstrepen et al., 2004; Gibbons & Rokas, 2009; Janulczyk et al., 2010; Jerome et al., 2011). Extensive studies in different organisms and with different cell surface genes support the notion that stochastic TR-based switching

<table>
<thead>
<tr>
<th>Affected moiety</th>
<th>Bacterial species</th>
<th>Gene(s) or operon</th>
<th>Repeat motif (5’-3’)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adhesin</td>
<td><em>Haemophilus influenzae</em></td>
<td>cha</td>
<td>56 bp</td>
<td>Sheets &amp; St Geme (2011)</td>
</tr>
<tr>
<td></td>
<td><em>Helicobacter pylori</em></td>
<td>sabA</td>
<td>CT</td>
<td>Goodwin et al. (2008)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria gonorrhoeae</em></td>
<td>opa</td>
<td>CTTT</td>
<td>Murphy et al. (1989)</td>
</tr>
<tr>
<td></td>
<td><em>Mycoplasma hominis</em></td>
<td>vaa</td>
<td>A</td>
<td>Zhang &amp; Wise (1997)</td>
</tr>
<tr>
<td></td>
<td><em>Legionella pneumophila</em></td>
<td>lcl</td>
<td>45 bp</td>
<td>Vandersmissen et al. (2010)</td>
</tr>
<tr>
<td>Capsule</td>
<td><em>Streptococcus pneumoniae</em></td>
<td>cps15bM</td>
<td>TA</td>
<td>van Selm et al. (2003)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria meningitidis</em></td>
<td>siaD</td>
<td>C</td>
<td>Hammerschmidt et al. (1996b)</td>
</tr>
<tr>
<td></td>
<td><em>Escherichia coli</em></td>
<td>neuO</td>
<td>AAAGCTC</td>
<td>Deszo et al. (2005)</td>
</tr>
<tr>
<td>Effector</td>
<td><em>Xanthomonas campestris</em></td>
<td>avrBs3</td>
<td>102 bp</td>
<td>Herbers et al. (1992), Kay et al. (2007)</td>
</tr>
<tr>
<td>Fe binding</td>
<td><em>Haemophilus influenzae</em></td>
<td>hpo</td>
<td>CCAA</td>
<td>Jin et al. (1999), Ren et al. (1999)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria meningitidis</em></td>
<td>hpuA, hmbR</td>
<td>G</td>
<td>Lewis et al. (1999), Richardson &amp; Stojilkovic (1999)</td>
</tr>
<tr>
<td>Flagellin</td>
<td><em>Campylobacter coli</em></td>
<td>flaA</td>
<td>T</td>
<td>Park et al. (2000)</td>
</tr>
<tr>
<td>Lipoprotein</td>
<td><em>Mycoplasma hyorhinis</em></td>
<td>vlp</td>
<td>24/26 bp</td>
<td>Rosengarten &amp; Wise (1991)</td>
</tr>
<tr>
<td></td>
<td><em>Mycoplasma bovis</em></td>
<td>vspA</td>
<td>18/24 bp</td>
<td>Lysnyansky et al. (1996)</td>
</tr>
<tr>
<td></td>
<td><em>Mycoplasma pulmonis</em></td>
<td>vsa</td>
<td>34 bp</td>
<td>Bhugra et al. (1995)</td>
</tr>
<tr>
<td>Lipopolysaccharides</td>
<td><em>Haemophilus influenzae</em></td>
<td>losA</td>
<td>CGAGCATA</td>
<td>Erwin et al. (2006)</td>
</tr>
<tr>
<td></td>
<td><em>Haemophilus influenzae</em></td>
<td>lic1, lic2 A, lic3 A</td>
<td>CAAT</td>
<td>Hosking et al. (1999), Dixon et al. (2007)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria meningitidis</em></td>
<td>lex2, oafA</td>
<td>GCAA</td>
<td>Griffin et al. (2003), Fox et al. (2005)</td>
</tr>
<tr>
<td></td>
<td><em>Haemophilus influenzae</em></td>
<td>lic3</td>
<td>15-18 bp</td>
<td>Yang &amp; Gotschlich (1996), Jennings et al. (1999)</td>
</tr>
<tr>
<td></td>
<td><em>Escherichia coli</em></td>
<td>tolA</td>
<td>C &amp; 21 bp</td>
<td>Appelmelk et al. (1999), Nilsson et al. (2008)</td>
</tr>
<tr>
<td>Metabolism</td>
<td><em>Escherichia coli</em></td>
<td>xylB</td>
<td>G</td>
<td>Funchain et al. (2000)</td>
</tr>
<tr>
<td>Mismatch repair</td>
<td><em>Escherichia coli</em></td>
<td>mutL</td>
<td>CTGGGC</td>
<td>Shaver &amp; Sniegowski (2003)</td>
</tr>
<tr>
<td></td>
<td><em>Salmonella Typhimurium</em></td>
<td>mutL</td>
<td>GCTGGC</td>
<td>Chen et al. (2010)</td>
</tr>
<tr>
<td>Outer membrane protein</td>
<td><em>Neisseria meningitidis</em></td>
<td>porA</td>
<td>G</td>
<td>van der Ende et al. (2000)</td>
</tr>
<tr>
<td></td>
<td>Group B streptococci</td>
<td>bca</td>
<td>246 bp</td>
<td>Gravekamp et al. (1996), Jennings et al. (1999)</td>
</tr>
<tr>
<td></td>
<td><em>Mycoplasma fermentans</em></td>
<td>p78</td>
<td>A</td>
<td>Theiss &amp; Wise (1997)</td>
</tr>
<tr>
<td>Inner membrane protein</td>
<td><em>Escherichia coli</em></td>
<td>tolA</td>
<td>15-18 bp</td>
<td>Zhou et al. (2012a, b)</td>
</tr>
<tr>
<td>Peroxiredoxin</td>
<td><em>Escherichia coli</em></td>
<td>ahpC</td>
<td>TCT</td>
<td>Ritz et al. (2001)</td>
</tr>
<tr>
<td>Pili</td>
<td><em>Neisseria gonorrhoeae</em></td>
<td>pilC</td>
<td>G</td>
<td>Jonsson et al. (1991, 1992)</td>
</tr>
<tr>
<td></td>
<td><em>Legionella pneumophila</em></td>
<td>fimV</td>
<td>18 bp</td>
<td>Coil &amp; Anne (2010)</td>
</tr>
<tr>
<td>Regulator</td>
<td><em>Listeria monocytogenes</em></td>
<td>ctsR</td>
<td>GGT</td>
<td>Karatzas et al. (2003, 2005)</td>
</tr>
<tr>
<td></td>
<td>prfA</td>
<td>CAGGAGT</td>
<td></td>
<td>Lindbäck et al. (2011)</td>
</tr>
<tr>
<td>R-M&lt;sup&gt;Ⅰ&lt;/sup&gt; system I</td>
<td><em>Haemophilus influenzae</em></td>
<td>hsdM</td>
<td>GACGA</td>
<td>Zaleski et al. (2005)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria gonorrhoeae</em></td>
<td>hsdS</td>
<td>G</td>
<td>Adamczyk-Poplawska et al. (2011)</td>
</tr>
<tr>
<td>R-M&lt;sup&gt;Ⅰ&lt;/sup&gt; system III</td>
<td><em>Haemophilus influenzae</em></td>
<td>modA</td>
<td>AGTC</td>
<td>Singhanta et al. (2005)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria spp.</em></td>
<td>modA</td>
<td>AGCC</td>
<td>Singhanta et al. (2009)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria spp.</em></td>
<td>modB</td>
<td>CCCC</td>
<td>Singhanta et al. (2009)</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria meningitidis</em></td>
<td>modD</td>
<td>ACCG</td>
<td>Seib et al. (2011)</td>
</tr>
<tr>
<td></td>
<td><em>Helicobacter pylori</em></td>
<td>res</td>
<td>C</td>
<td>de Vries et al. (2002)</td>
</tr>
<tr>
<td></td>
<td><em>Helicobacter pylori</em></td>
<td>modH</td>
<td>G</td>
<td>Singhanta et al. (2011)</td>
</tr>
<tr>
<td></td>
<td><em>Pasteurella haemolytica</em></td>
<td>mod</td>
<td>CACAG</td>
<td>Ryan &amp; Lo (1999)</td>
</tr>
<tr>
<td>R-M&lt;sup&gt;Ⅳ&lt;/sup&gt; system</td>
<td><em>Escherichia coli</em></td>
<td>mrr</td>
<td>G</td>
<td>Testaguz Mebrhatu et al. (2011)</td>
</tr>
</tbody>
</table>

TR, tandem repeat.
*The motif sequences of microsatellites (≤ 9 bp) are listed; for longer TRs, only the length is given.
<sup>†</sup>Different repeat unit lengths or sequences have been reported for this TR locus.
<sup>‡</sup>Two TR loci exist at different positions in the same gene.
<sup>§</sup>Restriction–modification system.
contributes to the rapid generation of diversity in surface structures, which in pathogens can serve as a mechanism for escaping the immune response and/or for determining tissue tropism. Some examples are reviewed more elaborately below.

**TRs within LPS biosynthesis genes**

LPS is a complex macromolecule in Gram-negative bacteria that is composed of three distinct parts (i.e. lipid A, core sugar, and O-antigen side chains), and a large number of genes are involved in its synthesis and export of LPS. As one of the major cell surface antigens, LPS is often implicated in cell adhesion and virulence. Furthermore, because LPS molecules are an essential structural component of the outer membrane that forms the outer shell of the Gram-negative cell, it also determines bacterial resistance to a variety of toxic chemicals, including some antibiotics and xenobiotics. A notable feature of LPS of some pathogens is the extensive intra- and inter-strain heterogeneity of the glycoform structure (i.e. the moieties comprising the core and O side chain sugars), which is mainly due to incomplete biosynthesis during the stepwise assembly of the sugar residues resulting from the phase-variable expression of the corresponding biosynthesis genes (Schweda et al., 2007). Typically, the phase variability of these genes derives from the fact that they contain nontrimeric TR tracts that exhibit stochastic variation. In some pathogens, this type of stochastic variation occurs in more than one gene for LPS synthesis, turning phase switching into a combinatorial process.

Unlike that of most Gram-negative bacteria, the core LPS of *H. influenzae* lacks the homopolymeric sugar units comprising the O-antigen side chains. Therefore, phase variation in LPS biosynthesis of *H. influenzae* occurs mainly through reversible expression switching of a subset of TR-containing genes, involved in the addition of core sugars (i.e. glucose and sialic acid) to the conserved tri-heptose backbone and in the addition of phospho-rycholine or acyl groups to these core sugars (Moxon et al., 2006; Fig. 5a). These genes include lic1A, lic1B, lic1C, and lic1D (collectively designated as lic1 locus), and lic2A, lic3A, lgtC, lex2, and oafA, each of which contains a variable tetrameric repeat tract (reviewed in Schweda et al., 2007). Rearrangements in the tetrameric repeat tracts of these genes can individually turn their expression on or off, resulting in a modified LPS molecule at the single cell level, and a repertoire of different LPS epitopes throughout the population (Fig. 5b; Moxon et al., 1994; Bayliss et al., 2001; Moxon et al., 2006). Remarkably, all the phase-variable genes involved in LPS biosynthesis of *H. influenzae* account for structural elements directly relevant to virulence as well (Schweda et al., 2007).

Interestingly, searching for genes harboring tetrameric TRs has proven to be a successful strategy to identify novel LPS biosynthesis genes in *H. influenzae* genomes (Hood et al., 1996; Fox et al., 2005). Similar mechanisms generating LPS diversity have also been described in *N. meningitidis*, where the ON/OFF variation in expression of genes of the *lgt* gene family (encoding LPS biosynthesis functions) is mediated by stochastic expansions or contractions in their intragenic homopolymeric tracts and where the corresponding LPS structures have so far been classified into 12 immunotypes (Jennings et al., 1999; Berrington et al., 2002; Bayliss et al., 2008).

The O-chain of *H. pylori* LPS can be fucosylated, thereby generating Lewis antigens that mimic human blood group antigens and mediate immune evasion (reviewed in Kusters et al., 2006). Lewis antigens of *H. pylori* are subject to reversible, high-frequency phase variation, and one of the mechanisms is the slipped-strand mispairing of poly-C tracts within three fucosyltransferase genes (*futA*, *futB*, and *futC*; Appelmelk et al., 1999; Wang et al., 1999). In addition, a unique TR region with a 21-bp unit size was recently uncovered in the 3′ end of *futA* and *futB*, but not in *futC*. Strikingly, although copy number variations of the 21-bp TR tract in both *futA* and *futB* did not alter the reading frame, they could affect the fucosyltransferase activity. In fact, a correlation was observed between the copy number of the 21-bp TRs and the number of O-antigen units being fucosylated, and the addition of one repeat unit led to the addition of an N-acetyl-β-D-lactosamine (LaCNA) unit in the O-antigen polysaccharide (Nilsson et al., 2006, 2008). These studies show that the variability of TRs in *futA*, *futB*, and *futC* increases the antigen diversity and population heterogeneity and thereby supports adaptability of *H. pylori* to fluctuating conditions in the gastric mucosa.

**TRs within capsule biosynthesis genes**

As one of the most external structures on the bacterial surface, capsules may completely conceal other antigenic surface molecules or may be co-exposed with other antigens, which are thought to be important for pathogenicity of bacteria. Production of a capsule provides some pathogenic bacteria with resistance to phagocytic and complement-mediated killing and, at the same time, affects bacterial attachment to host cells. Consequently, the ability to regulate capsule expression might confer a selective advantage for pathogens to cope with host immune responses during different stages of the infection process. For example, it was found that acapsulate variants of *N. meningitidis* serogroup B show much higher adherence and invasion of epithelial cells than their capsulated progenitors. Such variants can be generated at high frequency due to SSM of a poly-C tract in the polysialyltransferase
On the other hand, meningococcus isolates from the blood of meningitis patients are almost always capsulated, while both capsulated and noncapsulated strains typically coexist in the nose or throat of healthy carriers. Because phase variation of \( \text{siaD} \) is reversible, it was therefore proposed that infection is initiated by acapsulate strains, but that capsule biosynthesis is reactivated at a later stage during infection, allowing \( N. \text{meningitidis} \) to resist the host immune system and to cause disease (i.e., sepsis and meningitis). However, other findings indicate that the presence of a capsule does not necessarily preclude invasiveness, and the role of the capsule and capsule phase switching in different stages of meningococcal infection may therefore be more subtle and strain dependent (Spinosa et al., 2007; Bartley et al., 2013).

Besides binary ON/OFF switching, some bacteria can also modulate the composition of their capsule by a mechanism of TR-dependent phase switching. This is exemplified by the polysialic acid capsule (K1 antigen) of \( E. \text{coli} \) K1, which can be modified through phase-variable expression of the sialyl \( O \)-acytating activity, resulting in an altered immunogenicity and susceptibility to glycosidases. The phase-variable acetylation is driven by a heptanucleotide TR tract (AAGACTC; copy number typically 14–39) within the \( \text{neuO} \) gene. Loss or gain of a number of repeat units that is not a multiple of three results in a disruption of the \( \text{neuO} \) reading frame and subsequent NeuO expression (Deszo et al., 2005). Functional analysis furthermore revealed enhanced desiccation resistance, but reduced biofilm formation in \( E. \text{coli} \) K1 with active NeuO, suggesting not only a role in host interaction, but also a more subtle ecological impact of phase-variable \( \text{neuO} \) expression (Mordhorst et al., 2009). Interestingly, each set of three repeat units encodes a protein structure designated the poly(\( \psi \)) motif, and NeuO enzymatic activity was found to increase with the number of poly(\( \psi \)) motifs, supporting maintenance of high repeat copy numbers in the population (Bergfeld et al., 2007).
Another intriguing aspect of this capsular acetylation is that the *neuO* gene resides in a lambdoid prophage termed ‘CUS-3’, and mitomycin treatment of *E. coli* K1 can induce release of this phage, which specifically infects K1 antigen-expressing bacteria (Deszo *et al.*, 2005; King *et al.*, 2007). This suggests that variant *neuO* alleles can be redistributed among K1-encapsulated bacteria via horizontal transfer. Indeed, CUS-3/*neuO* has been found in several serotypes, such as O18 and O45 (King *et al.*, 2007). On the other hand, although the receptor for CUS-3 is polysialic acid, superinfection was not prevented by NeuO-mediated acetylation (Vimr & Steenbergen, 2006; King *et al.*, 2007). Notably, sialic acid is also known as a modification of LPS in numerous pathogens, underscoring the vital role of O-acetylation in causing structure variation of polysaccharide epitopes (also see TRs within LPS biosynthesis genes).

Also the mini- and macrosatellites within capsule biosynthesis genes of Gram-positive pathogens can mediate capsule diversity. As such, noncapsulated serotype 3 *Streptococcus pneumoniae* strains were shown to carry an out-of-frame perfect tandem duplication in one of the capsule biosynthesis genes (i.e. *cap3A*). Interestingly, based on the sequence and length of the duplication, at least seven different *cap3A* alleles were found in different noncapsulated strains. Analysis of the phase reversion frequency (OFF to ON) induced by TR contractions revealed a positive correlation between the frequency of reversion and the length of the duplication (Waite *et al.*, 2001). A similar mechanism of capsular phase variation correlating with a 223-bp and a 22-bp perfect tandem duplication in *cap8E* and *tts* was also demonstrated in serotype 8 and serotype 37, respectively, of *S. pneumoniae* (Waite *et al.*, 2003).

TRs within adhesin-associated genes

Adhesins mediate bacterial attachment to and further colonization of host tissues, but they also act as surface antigens (Bayliss *et al.*, 2001). The opacity proteins Opa of *Neisseria* spp. constitute a family of closely related, but size-variable outer membrane proteins, which enhance the adherence to epithelial, leukocyte, and phagocytic cells (reviewed in Sadarangani *et al.*, 2011). They do not only determine host and tissue specificity, but also facilitate efficient cellular invasion (Carbonnelle *et al.*, 2009). *Neisseria* spp. strains have 3–11 *opa* genes whose phase-variable expression is modulated by a pentanucleotide TR tract (CTCTT) in their coding regions or by intergenic recombination. As a result, a vast array of Opa variants can be generated to confer differential molecular specificities, allowing *Neisseria* both to alter its tissue tropism and to escape the host immune system. In fact, it has been argued that the occurrence of different arrays of Opa variants in clinical (disease) and carriage (nondisease) isolates of *N. meningitidis* may be the result of immune selection pressure (Callaghan *et al.*, 2006, 2008; Sadarangani *et al.*, 2011).

In the same vein, adhesins of *H. pylori* are categorized into two main subgroups, Hop (*Helicobacter* outer membrane porins) and Hor (Hop-related). For some, but not all of these adhesins, the corresponding host cell receptors have been identified (reviewed in Backert *et al.*, 2011). For example, BabA binds to the Lewis B (Le^B^) antigen expressed on the gastric mucosa (Iлер *et al.*, 1998), and SabA binds to glycosphingolipids of host cells that display a sialyl-dimeric Lewis X (*sialyl-Le^X^*) antigen (Mahdavi *et al.*, 2002). Interestingly, expression of some of the adhesins (i.e. BabB, HopZ, SabA, OipA) is subject to ON/OFF phase variation due to the presence of variable dinucleotide (CT) tracts within the corresponding genes. As a consequence, the expression patterns of adhesins generated by stochastic phase variation can not only affect the adhesion efficiency, but also alter the tissue tropism (Yamaoka *et al.*, 2002; Backert *et al.*, 2011). Additionally, the resulting repertoire of phase-variable adhesins can be advantageous for *H. pylori* to escape the host immune system.

A final example in this category is the Eap protein of the Gram-positive pathogen *Staphylococcus aureus*. Eap is a multifunctional adhesin with a variable TR tract, in which the size of each repeat unit is not identical (93–110 aa). A minimum of two TRs in the *eap* gene are required for Eap to cause agglutination, adherence and cellular invasion by *S. aureus*. Furthermore, as the repeat copy number increases from 2 to 5, those capacities are significantly enhanced, suggesting that TR copy number expansion in *eap* supports host adaptation of *S. aureus* (Hussain *et al.*, 2008).

TRs within iron (heme) acquisition genes

Free iron is typically limited in the host and is usually sequestered by iron-binding proteins (e.g. hemoglobin, transferrin, and lactoferrin). Iron acquisition mechanisms are therefore considered indispensable for bacterial pathogenicity, and pathogens have evolved many different strategies for adaptation to fluctuating iron concentrations. As such, pathogens are able to extract the iron from heme groups of iron-binding proteins via surface receptors. In *H. influenzae*, a family of hemoglobin-binding and hemoglobin–haptoglobin-binding proteins (Hgp) is known to mediate heme scavenging (Ren *et al.*, 1998; Jin *et al.*, 1999; Morton *et al.*, 1999). Individual strains of *H. influenzae* have 1–4 *hgp* genes, of which knockout analysis has confirmed that they are indispensable.
Variable tandem repeats and bacterial adaptation

virulence determinants in invasive disease (Seale et al., 2006). Interestingly, a CCAA tetranucleotide repeat tract exists in each hgp gene, which is reminiscent of the LPS biosynthesis genes of H. influenzae harboring a CAAT repeat tract (Moxon et al., 2006; see TRs within LPS biosynthesis genes). Phase variation induced by repeat rearrangements within the hgp genes has been observed; however, its biological significance is still obscure.

A similar case exists in N. meningitidis where iron acquisition is modulated by the variable poly-G tract in the hpuA and hmbR genes that are involved in the biosynthesis of two hemoglobin receptors (Lewis et al., 1999; Richardson & Stojiljkovic, 1999). More recently, it was revealed that 91% of N. meningitidis pathogenic isolates, but only 71% of commensal isolates have at least one receptor in an ON state, suggesting expression of hemoglobin(s) to be important for the systemic spread of meningococci (Tauseef et al., 2011).

**TRs within genes involved in restriction–modification systems**

In addition to their role as drivers of genetic variation as reviewed above, TRs can also drive epigenetic variation when they affect restriction–modification (R-M) systems. Currently, TR-dependent phase variations have been documented only in type I and III R-M systems. Type I R-M systems are generally comprised of three subunits (S, M, and R), which together form a holoenzyme that has both methylation and restriction activity. In H. influenzae, the type I R-M system HindI functions as the main defense system against the entry of foreign DNA (Glover & Piekarowicz, 1974). Phase variation of HindI is driven by a GACGA pentanucleotide repeat tract in the methyltransferase subunit gene (hsdM), as supported by the finding that an hsdM allele with four repeat units encoding an HsdM protein of normal length was associated with resistance to phage HP1c1 infection, while gain or loss of one repeat unit resulted in phage sensitivity (Zaleski et al., 2005). Interestingly, phase switching of phage susceptibility could also independently be conferred by LPS alterations induced by the variable tetranucleotide repeat tract of the lic2A gene, which confirmed the role of Lic2A-modified LPS as the receptor of HP1c1 phage (Zaleski et al., 2005; see also TRs within LPS biosynthesis genes). Moreover, the frequencies of the phase variations of phage susceptibility described above are affected by Dam (deoxyadenosine methyltransferase) activity, although the mechanism is not clear yet (Zaleski et al., 2005).

Another example was found in the NgoAV type I R-M system of N. gonorrhoeae, which is encoded by four genes: hsdM<sub>NgoAV</sub>, hsdR<sub>NgoAV</sub>, hsdS<sub>NgoAV1</sub>, and hsdS<sub>NgoAV2</sub>. The product of the hsdS<sub>NgoAV1</sub> is responsible for the specific recognition of the target site, whereas hsdS<sub>NgoAV2</sub> is nonfunctional. It was postulated that hsdS<sub>NgoAV1</sub> and hsdS<sub>NgoAV2</sub> are actually truncated proteins derived from an integral hsdS locus that has become interrupted by a frameshift mutation resulting in the formation of a stop codon between hsdS<sub>NgoAV1</sub> and hsdS<sub>NgoAV2</sub> (Piekarowicz et al., 2001). Interestingly, a recent study uncovered a variable poly-G tract within the 3' end of hsdS<sub>NgoAV1</sub> in N. gonorrhoeae, and loss of a guanine in that tract restores the fusion of the hsdS<sub>NgoAV1</sub> and hsdS<sub>NgoAV2</sub> genes, resulting in the generation of a new HsdS<sub>NgoAV</sub> protein, responsible for a novel NgoAV R-M system, termed ‘NgoAV’. The NgoAV system has a modified DNA recognition specificity, thereby conferring an altered susceptibility to various phages (Adamczyk-Poplawska et al., 2011).

Type III R-M systems only consist of two subunits, the methyltransferase (encoded by a mod gene) and restriction endonuclease (encoded by a res gene). In host-adapted pathogens, ON/OFF phase variation of this system has been reported due to variable TRs (with repeat unit not being a multiple of three bases) within either the mod or the res gene (Ryan & Lo, 1999; De Bolle et al., 2000; de Vries et al., 2002; Srikhanta et al., 2005, 2009, 2011; also see review Srikhanta et al., 2010). Interestingly, TR-induced phase variation in the mod gene has been shown to affect the expression of a number of genes, referred to as a phase-variable regalon or phasevarion (Srikhanta et al., 2005, 2010). In H. influenzae strain Rd, when mod expression was switched off by a TR rearrangement in a tetrameric repeat tract, nine other genes were down-regulated, and seven, up-regulated (Srikhanta et al., 2005). In the same vein, N. gonorrhoeae formed significantly thicker biofilms and thus may indirectly benefit from an increased resistance to external stresses, when its mod was in the OFF state. Additionally, a mod-ON phenotype resulted in an increased ability to associate with human cervical epithelial (pex) cells, whereas the mod-OFF configuration enhanced the ability to invade and survive within pex cells following invasion (Srikhanta et al., 2009). Further studies will be needed to clarify whether these phenotypes directly relate to mod deficiency itself or to altered expression of one or more members of the mod phasevarion.

Interestingly, most N. meningitidis and N. gonorrhoeae strains have a second phase-variable methyltransferase gene (Srikhanta et al., 2009), and some, even a third (Seib et al., 2011), and it has been anticipated that the combinatorial use of different phasevarions may contribute to further phenotypic variability (Seib et al., 2011). Furthermore, Fox et al. (2007) suggested that the apparent evolution of this type III R-M system into an
An epigenetic mechanism for controlling gene expression has resulted in loss of the DNA restriction function in some strains. In the latter context, it is also interesting to note the existence of solitary type IV restriction endonucleases that have no cognate methyltransferase. Because these peculiar enzymes display a specificity for methylated DNA, it has been anticipated that they might function to ward off the lateral acquisition of methyltransferases that might affect the cell’s epigenetic regulation (Fukuda et al., 2008; Tesfazgi Mebrhatu et al., 2011). As such, the phase variability of methyltransferase functions might serve to prevent detection of their activity during their establishment in a new host (Tesfazgi Mebrhatu et al., 2011).

**TRs within transcription activator-like effectors**

Gram-negative plant-pathogenic bacteria of the genus *Xanthomonas* can infect a broad spectrum of plants. A common feature of their infection process is the injection of virulence proteins (termed effectors) into host cells, mostly by means of a type III secretion system. The type III effectors are currently classified into nearly 40 groups based on sequence similarity and biochemical activity, and the largest group is the AvrBs3/PthA family, also known as transcription activator-like effector (TALE) family (reviewed in White et al., 2009). The most conspicuous feature of TALE genes is the variation of their central domain, mostly consisting of 15.5–19.5 nearly identical TRs with a 102-bp unit (Gürelbeck et al., 2006; Mak et al., 2013; Fig. 6).

A recent study revealed that the AvrBs3 can determine bacterial fitness in plants during infection (Kay et al., 2007). A set of *upa* (up-regulated by AvrBs3) genes among which *upa20*, the key regulator of the plant cell hypertrophy phenotype, has been identified as AvrBs3 targets in pepper plants. AvrBs3 acts as a transcription factor by binding to a conserved promoter element (UPA box) of *upa20*, resulting in up-regulation. Binding was shown to be mediated by the TR region of AvrBs3, suggesting the repeats to act as a DNA-binding motif (Kay et al., 2007). More specifically, it was found that the number of base pairs of the UPA box closely matches the TR copy number

---

**Fig. 6.** Domain organization and molecular function of *Xanthomonas* TALE AvrBs3. (a) The AvrBs3 functional domains shown include a type three secretion system (TTSS) signal sequence (dark blue), a central DNA-binding TR domain, a nuclear localization signal (purple), and a transcriptional activation domain (olive green). The TR domain in this case comprises 17.5 imperfect repeat units, each of which consists of 34 amino acids (aa). The unit numbered zero (red, dashed rectangle) is not a true repeat because it has a different aa sequence, but it also contributes to DNA binding. Each repeat binds to a base in the target sequence, and the binding specificity of the repeats is determined by aa 12 and 13 (known as RVD) and displayed with repeat-specific colors. The complete aa sequence of one repeat (no. 9) is shown with its RVD highlighted in green. (b) After injection into the plant cell by the bacterial type three secretion system, AvrBs3 is targeted to the nucleus and will bind with its TR domain to a specific DNA sequence known as UPA box. The consensus UPA box matches closely the binding specificity of the TR region of AvrBs3 as determined by the different TR units and their RVD (aa 12 and 13). AvrBs3 binding activates transcription of several UPA genes (adapted from Mak et al., 2013).
in AvrBs3. An elegant study demonstrated that each repeat unit of AvrBs3 binds to one base pair of the UPA box and that the base recognition specificity of each repeat is determined by two hypervariable amino acids (the 12th and 13th amino acids of each repeat unit), known as repeat variable di-residues (RVDs; Boch et al., 2009). Moreover, a minimum of 6.5 TRs in AvrBs3 are necessary for activating upa gene expression (Boch et al., 2009; Moscou & Bogdanove, 2009; Scholze & Boch, 2011; Fig. 6). Most recently, the structural basis for this kind of sequence-specific recognition has been uncovered by crystallographic analysis of an artificially engineered TAL effector hybridized with its target DNA (Deng et al., 2012). AvrBs3-dependent modulation of plant signaling pathways causes enlargement of mesophyll cells and hypertrophy of the infected tissue, which might help the bacteria to proliferate and escape from infection sites to facilitate bacterial spreading (Kay et al., 2007; Boch & Bonas, 2010). Consequently, TR variations in the avrBs3 gene preclude activation of the UPA box causing a failure in inducing the hypersensitive response in plants (Kay et al., 2007).

While the above studies highlight the biological importance of TALEs as major virulence determinants, TALEs also provide interesting avenues for biotechnological applications. As such, it will be possible to engineer broader pathogen resistance in crops by combining several UPA boxes into the promoter of plant resistance genes like Bs3, which will render these transgenic plants resistant to infection by bacteria delivering matching TAL effectors (Boch & Bonas, 2010; Scholze & Boch, 2011). Alternatively, the TR region of TAL effectors can be tailored as to recognize and bind to a pre-defined DNA sequence, resulting in activating the expression of downstream target genes (Morbitzer et al., 2010). Moreover, engineered TAL effectors can be fused with endonuclease domains to generate TAL effector nucleases that can introduce cuts or double-strand breaks in or near specific sequences on the chromosomes, targeting these loci for mutagenesis or recombinational repair and gene therapy (Bogdanove & Voytas, 2011; Muñoz Bodnar et al., 2013).

**TRs within stress response genes**

Regulation of stress response genes is one of the most common strategies employed by bacteria to cope with stresses. TRs have been identified in a number of stress response genes (Rocha et al., 2002), and some studies have addressed the role of these repeats in the modulation of a stress response. For example, the gene encoding the CtsR regulator (class III stress gene repressor) of *Listeria monocytogenes* carries a triplet repeat (GGT) tract with three copies. Stochastic deletion of one triplet in the *ctsR* gene results in an inactive CtsR repressor, leading to expression of the *clp* genes (Karatzas et al., 2003, 2005). This alteration confers increased resistance to high hydrostatic pressure, heat, acid, and *H₂O₂*, but attenuates virulence in *L. monocytogenes* (Karatzas & Bennik, 2002).

Another example is *mutL*, which is involved in mismatch repair in *Salmonella Typhimurium* and *E. coli*. The functional allele of *mutL* carries a trimeric hexanucleotide repeat in the region encoding the ATP-binding pocket of the protein. Spontaneous loss of one repeat unit resulting in a mutator phenotype due to MutL deficiency has been observed in long-term cultures of both *E. coli* and *S. Typhimurium* (Shaver & Sniegowski, 2003; Chen et al., 2010). Expansion of this TR region from 3 to 4 units and from 2 to 3 units has also been observed and, in the latter case, caused reversion of the mutator phenotype (Shaver & Sniegowski, 2003; Chen et al., 2010; Le Bars et al., 2013). This genetic switching of MutL may serve as a strategy to control the balance between genetic stability and mutability and thus serve as an element controlling bacterial evolution. Interestingly, not only *mutL*, but also many other genes (i.e. *mutT*, *mutY*, *mutS*, *dinI*, and *ruvC*) involved in DNA repair of *E. coli* harbor SSRs, further underscoring the putative regulatory role of TRs in stress response (Rocha et al., 2002).

Some membrane proteins are essential for membrane integrity and as such for tolerance to a variety of toxic chemicals and other stresses. One such protein, that is present in many Gram-negative bacteria, is TolA, which harbors a variable TR tract composed of 8–16 imperfect copies of a 15- to 18-bp repeat unit in *E. coli* (Levengood et al., 1991; Zhou et al., 2012b). TR variations in TolA occurred at a frequency of at least 6.9 × 10⁻⁵ in a clonal wild-type population of *E. coli* MG1655 and were shown to modulate stress tolerance, with the most outspoken TR-dependent phenotype being deoxycholic acid tolerance (Zhou et al., 2012a). However, the precise molecular mechanism underlying this phenotypic variation remains unclear.

A peculiar case is the triplet repeat (TCT) in *E. coli* *ahpC*, where expansion of the TR tract with 1 unit converts the AhpC protein from a peroxidase into a disulfide reductase, as demonstrated by the ability of this newly acquired enzyme activity to restore normal growth of a mutant lacking thioredoxin and glutathione reductase (Ritz et al., 2001). To our knowledge, this is the only example of an intragenic TR variation generating a truly novel function in bacteria.

**Conclusions and outlook: TRs and bacterial evolution**

The generation of mutations is the basis of evolution in bacteria and all other living organisms. Under changing environmental conditions, the evolution of better adapted
descendants may be essential for survival of the species, and an increased mutation rate obviously could confer higher survival chances. On the other hand, the generation of random mutations all over the genome also causes an important burden because most mutations are deleterious. Therefore, to be successful, bacteria will have to maintain a balance between genome plasticity and stability. One way to do this is to control the spatial distribution of mutations over the genome by evolving highly mutable sequences that are associated with loci (also known as contingency loci) that are needed for a flexible response to environmental conditions or stresses, especially those that cannot be detected by conventional bacterial sensors (e.g. phages or host receptors; Fonville et al., 2011). TRs and their variability provide a paradigm for this regulatory strategy of adaptability.

The formation of TRs is suggested to be a random process based on replication slippage (Levinson & Gutman, 1987); however, only some of the formed TRs are believed to be maintained during evolution. Generally, variable TRs tend to localize in flexible genes involved in the biosynthesis of surface structures and with a function in adhesion and (for pathogens) invasion, although they are occasionally also found in genes encoding critical cellular functions like DNA replication (Moxon et al., 2006; Guo & Mrázek, 2008). The association between TRs and cell surface structures is suggested to allow populations to anticipate changes in the environment in order to enhance their survival rate (Moxon et al., 1994). This is particularly common and critical for pathogens such as H. influenzae and H. pylori with limited genetic information (i.e. reduced genome) to cope with complex environments (Razin et al., 1998).

As such, TR-dependent phase variation can be regarded as a strategy for bacterial adaptation that is complementary to conventional mutations (i.e. single nucleotide polymorphism), but has some distinct features. First, the hypermutability (typical mutation rates of $10^{-2}$–$10^{-5}$ per generation) can be advantageous for adaption on a short time scale, and it has been demonstrated using a theoretical model that TRs mediating stochastic switching can evolve and be maintained under a wide range of alternating selection regimens (Palmer et al., 2013). Second, TR rearrangements, both local and combinatorial, can effectuate alterations at both the transcriptional and the translational levels, resulting in either binary switching (‘ON’ and ‘OFF’) or gradual control, and this facilitates subtle adaptation of bacterial fitness under stress. Further, TR-dependent mechanisms operate not only in classic genetic, but also in epigenetic pathways and from the single locus to the more global level of phasevariations, suggesting the power of TR intermediate regulation in bacterial adaptation.

Although rapid TR-dependent phase switching facilitates bacterial adaption on a short time scale, it is probably not necessary for bacteria to maintain hypermutable regions in the absence of selection pressures, because of the associated cost of DNA replication fidelity and metabolic energy. An interesting, but mostly unanswered question is how bacteria optimize the mutation rate of TRs in phase-variable genes in a fluctuating environment. Some intrinsic features of the TR sequence (i.e. conservation and copy number of repeat unit) and cellular function (i.e. DNA replication, recombination, and repair) are known to affect the frequency of TR-dependent phase variation and may be used to this purpose (Bayliss, 2009). On the other hand, some studies have shown a modulation of the TR rearrangement frequency by environmental stress (Kanbashi et al., 1997; Jackson & Loeb, 2000; Rasmussen & Björk, 2001; Srihanta et al., 2009; Cooley et al., 2010). However, in general, it remains unclear how environmental signals are transduced to modulate the switching rate of TR-containing genes.

In conclusion, TRs confer local sequence flexibility in bacterial genomes, thereby allowing targeted mutation and evolution. The genotypic and phenotypic variations modulated by TRs are rapid and coordinative and support the generation of substantial biological diversity on a short time scale. In spite of a certain metabolic cost, ‘prepared genomes’ (i.e. with TR-based contingency loci; Caporale, 1999) have a higher adaptability and thus a fitness advantage in frequently fluctuating environments.

**Acknowledgements**

This work was supported by the Research Foundation – Flanders (FWO – Vlaanderen; Research project G.0289.06) and by the KU Leuven Research Fund (project METH/07/03).

**References**


Appelmelk BJ, Martin SL, Monteiro MA et al. (1999) Phase variation in *Helicobacter pylori* lipopolysaccharide due to...
changes in the lengths of poly(C) tracts in alpha3-fucosyltransferase genes. Infect Immun 67: 5361–5366.


Bartley SN, Tseng YL, Heel K et al. (2013) Attachment and invasion of Neisseria meningitidis to host cells is related to surface hydrophobicity, bacterial cell size and capsule. PLoS ONE 8: e55798.


Gravekamp C, Kasper DL, Michel JL, Kling DE, Carey V & Madoff LC (1997) Immunogenicity and protective efficacy of...
of the alpha C protein of group B streptococci are inversely related to the number of repeats. *Infect Immun* 65: 5216–5221.


Guo X & Mann (2010) TRPing up the genome: tandem repeat


Ren Z, Jin H, Whitby PW, Morton DJ & Stull TL (1999) Role of CCAA nucleotide repeats in regulation of hemoglobin and hemoglobin-haptoglobin binding protein...


