**Bacterial strain typing in the genomic era**

Wenjun Li, Didier Raoult & Pierre-Edouard Fournier

Unité des Rickettsies et pathogènes émergents, CNRS UMR 6236, IRD 198, Faculté de médecine, Université de la Méditerranée, Marseille, France

**Abstract**

Bacterial strain typing, or identifying bacteria at the strain level, is particularly important for diagnosis, treatment, and epidemiological surveillance of bacterial infections. This is especially the case for bacteria exhibiting high levels of antibiotic resistance or virulence, and those involved in nosocomial or pandemic infections. Strain typing also has applications in studying bacterial population dynamics. Over the last two decades, molecular methods have progressively replaced phenotypic assays to type bacterial strains. In this article, we review the current bacterial genotyping methods and classify them into three main categories: (1) DNA banding pattern-based methods, which classify bacteria according to the size of fragments generated by amplification and/or enzymatic digestion of genomic DNA, (2) DNA sequencing-based methods, which study the polymorphism of DNA sequences, and (3) DNA hybridization-based methods using nucleotideic probes. We described and compared the applications of genotyping methods to the study of bacterial strain diversity. We also discussed the selection of appropriate genotyping methods and the challenges of bacterial strain typing, described the current trends of genotyping methods, and investigated the progresses allowed by the availability of genomic sequences.

**Introduction**

Genetic diversity can ultimately explain most phenotypic variability in bacteria, such as geographic distribution, host specificity, pathogenicity, antibiotic resistance, and virulence. As bacterial strains pose ever greater challenges to human health, including increased virulence and transmissibility, resistance to multiple antibiotics, expanding host spectra, and the possibility of genetic manipulation for bioterrorism, identifying bacteria at the strain level is increasingly important in modern microbiology (Fournier et al., 2004).

The intraspecies diversity of bacteria mainly results from three genetic events: horizontal gene transfer, gene loss or acquisition, and recombination. The frequency of these three events makes the investigation of intraspecies diversity quite complicated (Fraser-Liggett, 2005). Although whole-genome sequencing, followed by genome comparison, is in theory an ideal way to elucidate the genetic variability within a bacterial species, it remains expensive and labor intensive (Field et al., 2004). Bacterial strain typing, characterizing a number of strains in detail and ascertaining whether they are derived from a single parental organism, is a way to identify bacteria at the strain level and to uncover the genetic diversity underlying important phenotypic characteristics.

Theoretically, there are two distinct typing systems: phenotyping and genotyping. Bacterial phenotypes, determined by the morphology of colonies on various culture media, biochemical tests, serology, killer toxin susceptibility, pathogenicity, and antibiotic susceptibility, are not variable enough for discriminating between closely related strains. In the genomic era, the scientific basis for the identification and subtyping of microorganisms has shifted to genetic methods. More recently, genotyping, which refers to the discrimination of bacterial strains based on their genetic content, has become widely used for bacterial strain typing due to its high resolution. The genetic profile of a given strain generated by a specific genotyping method can be as unique as a fingerprint. Thus, genotyping is also referred to as DNA fingerprinting.

Current bacterial strain typing methods may be classified into three main categories: DNA banding pattern-, DNA sequencing-, and DNA hybridization-based methods. DNA banding pattern-based genotyping methods discriminate the studied strains based on differences in the size of the DNA bands (fragments) generated by amplification of genomic DNA or by cleavage of DNA using restriction enzymes (REs). DNA sequencing-based genotyping methods generate the original sequence of nucleotides and discriminate among...
bacterial strains directly from polymorphisms in their DNA. DNA hybridization-based methods are mainly referred to as DNA macroarray and microarray studies. In this technique, bacterial strains are discriminated by analyzing the hybridization of their DNA to probes of known sequences. With the exception of genome sequencing, the discriminatory ability of genotyping methods is species dependent.

**Current methods for bacterial strain typing**

**DNA banding pattern-based methods (Fig. 1)**

DNA bands (fragments) can be generated by digestion of DNA with REs, DNA amplification, or by a combination of both. DNA amplification generates billions of copies of a genomic fragment and has many advantages, including sensitivity, speed, and applicability to a wide array of human specimens and environmental samples (Lo & Chan, 2006). REs precisely recognize and cut target DNA at a defined sequence, which makes enzymatic restriction an effective tool (Blakesley, 1987).

**Enzymatic restriction**

**Pulsed-field gel electrophoresis (PFGE)**

PFGE is an electrophoretic technique used to separate large DNA molecules (10 kb–10 Mb). In a conventional constant electric field, DNA molecules > 20 kb show the same mobility, making it impossible to differentiate between these DNA molecules. By applying alternating electric fields at different angles, however, PFGE can separate large DNA molecules in a flat agarose gel (Schwartz & Cantor, 1984; Herschleb et al., 2007) (Fig. 1). REs with uncommon recognition motifs are used in PFGE to generate large DNA fragments (Table 1) and the banding patterns of PFGE in a group of strains reflect DNA polymorphism at the RE recognition sites. PFGE provides a high-resolution, macrorestriction analysis at the genome level, leading it to be considered as the ‘gold standard’ for subtyping many bacteria (Tenover et al., 1995; Gerner-Smidt et al., 2006).

The choice of RE is one of the most important factors in determining the PFGE banding pattern because the cleavage site of each RE is unique. REs with long, infrequently occurring recognition motifs may provide higher resolution in PFGE (Chen et al., 2005). In silico searches of completed bacterial genome sequences to detect recognition sites of rare-cutting REs proved to be an effective way to optimize the discriminatory power and reduce the cost of PFGE in the case of typing *Bordetella pertussis* (Lee et al., 2006). A website dedicated to in silico RE analysis of complete bacterial genomes was developed by Bikandi and colleagues. Their strategy allows the calculation of the theoretical number and size of restriction fragments generated by different REs and provides a graphically detailed restriction profile when the number of fragments is 50 or fewer (Bikandi et al., 2004) (Table 1). Nevertheless, PFGE results should be validated by comparison with epidemiological traits and other genotyping results. For example, genotyping *Bartonella henselae*...
Table 1. Comparison of DNA banding pattern-based genotyping methods used in bacterial strain typing

<table>
<thead>
<tr>
<th>Methods</th>
<th>Bacterial culture required</th>
<th>REs</th>
<th>Primers needed</th>
<th>DNA fragments</th>
<th>Size</th>
<th>Resolution</th>
<th>Main reagents</th>
<th>Main equipments*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFGE</td>
<td>Yes</td>
<td>Rare cutters</td>
<td>No</td>
<td></td>
<td>Large (&gt; 20 kb)</td>
<td>PFGE</td>
<td>REs and proteases</td>
<td>Gel box, switching unit, cooler, power supply</td>
</tr>
<tr>
<td>RFLP</td>
<td>Yes</td>
<td>Frequent cutters</td>
<td>No</td>
<td></td>
<td>Moderate (&lt; 10 kb)</td>
<td>GE followed by Southern blotting</td>
<td>REs and probes</td>
<td>Southern transfer</td>
</tr>
<tr>
<td>Ribotyping</td>
<td>Yes</td>
<td>Frequent cutters</td>
<td>No</td>
<td></td>
<td>Moderate (&lt; 10 kb)</td>
<td>GE followed by Southern blotting</td>
<td>REs and probes</td>
<td>Southern transfer</td>
</tr>
<tr>
<td>PCR-RFLP</td>
<td>No</td>
<td>Frequent cutters</td>
<td>Flanking target fragment</td>
<td>Small (&lt; 5 kb)</td>
<td>GE</td>
<td>Taq, primers and REs</td>
<td>Thermal cycler</td>
<td></td>
</tr>
<tr>
<td>AFLP</td>
<td>Yes</td>
<td>A pair of REs</td>
<td>Selective primers¹</td>
<td>Very small (&lt; 1 kb)</td>
<td>GE or CE</td>
<td>REs, adaptors, T4, ligase, Taq, primers</td>
<td>Thermal cycler</td>
<td></td>
</tr>
<tr>
<td>REP-PCR</td>
<td>Yes</td>
<td>No</td>
<td>Complementary to repetitive elements</td>
<td>Small (&lt; 5 kb)</td>
<td>GE or CE</td>
<td>Taq and primers</td>
<td>Thermal cycler</td>
<td></td>
</tr>
<tr>
<td>RAPD</td>
<td>Yes</td>
<td>No</td>
<td>Arbitrary primers</td>
<td>Small (&lt; 5 kb)</td>
<td>GE or CE</td>
<td>Taq and primers</td>
<td>Thermal cycler</td>
<td></td>
</tr>
<tr>
<td>MLVA</td>
<td>No</td>
<td>No</td>
<td>Flanking tandem repeats</td>
<td>Very small (&lt; 1 kb)</td>
<td>GE or CE</td>
<td>Taq and primers</td>
<td>Thermal cycler</td>
<td></td>
</tr>
<tr>
<td>DGE</td>
<td>No</td>
<td>No</td>
<td>Flanking target fragment</td>
<td>Small (&lt; 5 kb)</td>
<td>Denaturing gel</td>
<td>Taq and primers</td>
<td>Thermal cycler, specific electrophoresis equipment</td>
<td></td>
</tr>
<tr>
<td>HRM</td>
<td>No</td>
<td>No</td>
<td>Flanking target fragment</td>
<td>Small (&lt; 5 kb)</td>
<td>Melting curve analysis</td>
<td>Taq and primers</td>
<td>Real-time thermal cycler</td>
<td></td>
</tr>
</tbody>
</table>

*Conventional gel electrophoresis, photography, and digitalizing equipment are not included.

¹Adapter sequences at 5’ end, extending with a variable number of 3’ nucleotides (usually one to three).

GE, gel electrophoresis; CE, capillary electrophoresis.

Using NotI-based PFGE proved to be much more discriminatory than using four other REs: Smal, Apal, Eco52I, and XmaI. However, NotI-based PFGE types did neither match other PFGE types nor genotypes obtained using multilocus sequence typing, indicating that NotI is not an appropriate RE for PFGE typing of B. henselae (Arvand & Viezens, 2007). Therefore, the choice of a suitable RE is crucial to generate reliable PFGE types.

PFGE is now widely used in epidemiology, microbiology, and evolutionary biology. The rapid accumulation of PFGE patterns necessitates the standardization of protocols and criteria for bacterial typing, as well as the development of online databases for global comparison of PFGE patterns (Ransom & Kaplan, 1998; Herschleb et al., 2007). The largest PFGE database, PulseNet, currently tracks four foodborne bacteria: Escherichia coli O157:H7, nontyphoidal Salmonella, Shigella, and Listeria monocytogenes (Ransom & Kaplan, 1998; Ribot et al., 2006; Swaminathan et al., 2006) (Table 1). As proposed by Tenover and colleagues, bacterial isolates yielding the same PFGE profile are considered as belonging to the same strain. Isolates that differ by a single genetic event, which is reflected as a difference of one to three bands, are considered as being ‘closely related,’ and isolates differing by four to six bands, likely representing two independent genetic events, are considered as being ‘possibly related.’ Bacterial isolates containing six or more band differences (representative of three or more genetic changes) are considered as being ‘unrelated’ (Tenover et al., 1995).

However, although widely used, PFGE has several limitations. This method is time and labor consuming and lacks reproducibility and interlaboratory comparability. In addition, it requires high-quality DNA, is poorly applicable to human or environmental samples, and may lack resolution power to distinguish bands of nearly identical size (Davis et al., 2003). Other drawbacks include the risk of laboratory-acquired infection due to prolonged handling of bacterial strains before treating with proteases and REs, and many other factors such as concentration of DNA in the agarose plugs, amount of agarose in the gel, electrophoresis voltage, gel temperature, and buffer strength, which may also influence patterns (Chung et al., 2000).

Restriction fragment length polymorphism (RFLP)

RFLP analysis is one of the first techniques to be widely used for detection of variations in DNA sequence (Thibodeau, 1987; Todd et al., 2001). As its name implies, RFLP measures the size of restriction fragments separated by conventional agarose gel electrophoresis. The general procedure for RFLP analysis is shown in Fig. 1. Digestion of genomic DNA with frequently cutting REs may produce hundreds of short restriction fragments, making clear separation of fragments...
difficult using agarose gel electrophoresis (Table 1). However, RFLP analysis can be greatly simplified by subjecting the partial restriction fragments to Southern blotting with labeled probes (Southern, 1975).

In bacteriology, if two strains differ in the distance between cleavage sites for a particular RE, the length of the restriction fragments is different between the strains. The similarity of the generated patterns of restriction fragments can be used to differentiate strains and to analyze the genetic relatedness (Busse et al., 1996). RFLP patterns are mainly determined through the specific combination of REs and nucleic acid probes. RFLP analysis using probes derived from the insertion element, IS6110, has been considered as the ‘gold standard’ method for typing Mycobacterium tuberculosis complex (MTC) bacteria (Otal et al., 1991; Mostrom et al., 2002; Kanduma et al., 2003). These bacteria were also subtyped using RFLP analysis based on polymorphic GC-rich repeat sequences (Poulet & Cole, 1995; Mostrom et al., 2002; Kanduma et al., 2003).

Ribotyping, a variation of RFLP analysis, uses rRNA probing. Bacterial rRNA operons comprise a family of highly conserved genes, each of which is flanked by more variable DNA regions. Sequence variations in the flanking restriction sites result in different RFLP banding patterns when probed with conserved domains of the 16S or 23S rRNA genes (Bingen et al., 1994; Harvey & Minter, 2005). The banding patterns of DNA fragments corresponding to the relevant rRNA are named ribotypes. An advantage of ribotyping is that it allows analysis without prior knowledge of genomic DNA sequence, because rRNA operons are universal. In addition, the results of ribotyping are easier to interpret because fewer fragments are produced (Bingen et al., 1994; Harvey & Minter, 2005). An automated apparatus for ribotyping is commercially available, with the Riboprinter™ Microbial Characterization System (Qualicon, Wilmington, DE). This system provides highly reproducible and standardized ribotyping data. Ribobank is a ribotyping database comprising numerous bacterial ribotypes generated by Qualicon’s Riboprinter approach (Table 2).

Because it does not require expensive equipment, RFLP analysis is cost effective except for automatic ribotyping that requires a relatively expensive kit. Among limitations, RFLP analysis requires large amounts of high-quality genomic DNA, which can limit its application in many cases. In addition, it is time and labor consuming, especially when coupled with Southern blotting, and involves detection systems that use either radioisotopes or complex biochemistry.

**DNA amplification**

**Arbitrarily primed PCR (AP-PCR)**

AP-PCR is a variation of classic PCR for random amplification of unknown genomic regions using arbitrary primers. Therefore, it is also termed random amplification of polymorphic DNA (RAPD) (Welsh & McClelland, 1990; Williams et al., 1990). Using short single primers (usually < 10 bp) combined with low annealing temperatures, AP-PCR enables genomic DNA to be amplified at multiple loci, which generates different-sized amplicons (Table 1, Fig. 1). Unlike traditional PCR analysis, AP-PCR does not require prior knowledge of the target DNA: the identical 10-mer primers may or may not amplify a segment of DNA, depending on the positions that are complementary to the primer sequences (Power, 1996) (Fig. 1). For example, no fragment is produced if primers anneal too far apart or if the 3’ ends of the primers fail to face each other. If a mutation has occurred in the template DNA at a site that was previously complementary to the primer, a PCR product may not be produced, and a different pattern of amplified DNA fragments is observed on the gel.

Based on the observation that the number of fragments generated by AP-PCR correlates with the number of arbitrary primer annealing sites in the target genome sequence, Li et al. (2006a) proposed a statistical method and developed a computer program named *RAPD* to determine the number of annealing sites and amplicons for RAPD typing (Table 2). Some arbitrary primers, namely the universal rice primers extracted from a repetitive sequence in the rice genome (Kang et al., 2002; Jin et al., 2006) and the highly variable penta-GTG oligonucleotide [(GTG)5] (Matshika et al., 2006), have frequently been used for RAPD analysis of *Salmonella enterica* and *Campylobacter concisus*, respectively.

‘RAPD-GENERATOR’ a Java-based program is also useful for generating arbitrary primers (Table 2).

As a PCR-based genotyping method, AP-PCR is inexpensive, fast, and sensitive. It has widely been used to study the genetic variability of many bacterial species, including important human pathogens (Power, 1996; Jin et al., 2006). However, the reproducibility of this method remains a challenge, which hinders comparison of AP-PCR patterns between and within laboratories (Power, 1996; Tyler et al., 1997). In addition, several factors may influence AP-PCR results, such as sequence and annealing temperature of arbitrary primers, DNA template purity and concentration, and PCR equipment and reagents (Tyler et al., 1997). Optimization and standardization of the procedure may create sufficiently reliable conditions for using AP-PCR in epidemiological typing of an expanding range of bacteria (Mortimer & Arnold, 2001; Mostrom et al., 2002). The Ready-To-Go RAPD analysis beads (Amersham Pharmacia Biotech) provide a strictly consistent and uniform source of Taq polymerase, dNTPs, and buffer conditions along with a consensus protocol for RAPD analysis. This approach may prove useful for standardizing and optimizing this genotyping method (Hyttia et al., 1999; Vogel et al., 1999; Chansiri-pornchai et al., 2000).
Repetitive sequencing-based PCR (REP-PCR)

A series of naturally occurring repetitive DNA sequences are dispersed in multiple copies throughout bacterial genomes (Gilson et al., 1984; Stern et al., 1984; Hulton et al., 1991; Lupski & Weinstock, 1992; Koeth et al., 1995). Although the functions of these interspersed repetitive DNA elements remain unknown, their presence is useful for DNA fingerprinting of bacteria. In REP-PCR, primers complementary to those interspersed repetitive consensus sequences are used to amplify DNA fragments between repetitive elements (Versalovic et al., 1991; de Bruijn, 1992) (Fig. 1). Three families of repetitive sequences: the 35–40-bp repetitive extragenic palindromic (REP) sequence (Gilson et al., 1984; Stern et al., 1984), the 124–127-bp enterobacterial repetitive intergenic consensus (ERIC) sequence (Hulton et al., 1991; Lupski & Weinstock, 1992), and the 154-bp BOX element sequences (Koeth et al., 1995), have frequently been used in REP-PCR assays. The corresponding protocols are referred to as REP-PCR, ERIC-PCR, and BOX-PCR genomic fingerprinting, respectively.
Detection of DNA fragments is obtained by agarose gel electrophoresis or capillary electrophoresis. The optimization of electrophoresis conditions (electrokinetic injection and applied voltage) can dramatically increase the resolution of amplified DNA fragments up to 1000 bp (Sciaccitano, 1998).

Because of the low cost of the materials and the rapidity, ease of use, and low labor intensity, REP-PCR may be a valuable method for bacterial strain typing. In contrast to AP-PCR, REP-PCR is more reproducible because specific primers are used for amplification. REP-PCR has been shown to be beneficial for typing eukaryotic and prokaryotic organisms, including bacteria of epidemiological significance, i.e. Acinetobacter spp. (Snelling et al., 1996), Streptococcus pneumoniae (Versalovic et al., 1993), M. tuberculosis (Cangelosi et al., 2004), Clostridium difficile (Northe et al., 2005), and L. monocytogenes (Chou & Wang, 2006). A commercial kit, the REP-PCR DiversiLab microbial typing system (Spectral Genomics Inc., Houston, TX), is reported to be more convenient than manual methods (Healy et al., 2005; Shutt et al., 2005). The online DIVERSILAB software not only provides standardized comparisons among isolates almost instantaneously, but also generates user-friendly customized reports and provides a user-specific data storage and retrieval system (Healy et al., 2005). In addition, REP-PCR genomic fingerprinting protocols can be performed on whole cells of some species, obviating the need for DNA extraction (Woods et al., 1993; Snelling et al., 1996). However, REP-PCR has disadvantages similar to other PCR-based assays, including the potential for contamination, artifacts, and the need for multiple controls.

Multiple-locus variable number tandem repeat analysis (MLVA)

Variable number tandem repeats (VNTR) are 'head-to-tail' (tandemly) repeated DNA sequences that vary in copy number and are dispersed widely in human and bacterial genomes (Lupski & Weinstock, 1992; van et al., 1998; Vergnaud & Denoeud, 2000). In bacterial genomes, VNTR loci were found in noncoding regions as well as in genes. These VNTRs can be an important source of genetic polymorphism for strain typing because of their rapid evolution (van et al., 1998, 2001; Ogata et al., 2000; Fournier et al., 2004). The number of tandem repeats per locus may vary dramatically between strains within a given species. MLVA is a PCR-based genotyping method based on the polymorphic analysis of multiple VNTR loci on the chromosome (Lindstedt, 2005; van, 2007). For each VNTR locus, the number of repeats can be determined by PCR amplification using primers complementary to the well-conserved sequences flanking the tandem repeats. The fragment size varies with regard to the size and number of repeat units, and the banding patterns can be analyzed using a software to reveal the genotype and infer the phylogenetic relationships (Keim et al., 2000; Tenover et al., 2007) (Fig. 1). By fluorescently labeling PCR amplicons with different dye colors and separating them by capillary electrophoresis in an automated sequencer, MLVA can be performed in a single PCR tube (multiplex PCR), which dramatically improves its efficiency (Lindstedt, 2005; Lista et al., 2006).

Scanning the bacterial genome and extracting VNTR loci are the first steps in MLVA. Tandem repeat–finding programs, such as TANDEM REPEAT FINDER (Benson, 1999), the Tandem Repeat database, and the Microsatellite Repeats database provide a series of bioinformatic tools for searching bacterial genomes for potential VNTR loci (Table 2). Databases and web-based tools for VNTR searching, primer picking, and phylogenetic analysis facilitate the implementation of MLVA (Table 2) (Denoeud & Vergnaud, 2004; Grissa et al., 2007).

Since its first use to type bacteria in 2000, MLVA has proven to be a high-resolution method for discrimination of many bacteria. It is now regarded as a reference typing method for many bacterial species, such as Francisella tularensis (Farlow et al., 2001; Johansson et al., 2004), Bacillus anthracis (Keim et al., 2000; Hoffmaster et al., 2002; Lista et al., 2006), Yersinia pestis (Klevetska et al., 2001), and M. tuberculosis (Le et al., 2002). MLVA has also been applied to other important human pathogens such as methicillin-resistant Staphylococcus aureus strains (Tenover et al., 2007), Burkholderia pseudomallei (U’Ren et al., 2007), and C. difficile (van den Berg et al., 2007). Because of the rapid evolution of VNTR loci in bacterial genomes, MLVA has the same or more discriminatory power than most other typing methods (Johansson et al., 2004; Lindstedt, 2005; Liao et al., 2006; Tenover et al., 2007).

The stability of VNTRs is mainly associated with the length of repetitive unit, copy number of repetitive unit, and the purity (Legendre et al., 2007). A program, SERV, is now available for the selection and comparison of suitable VNTR for genotyping based on scoring the length of repetitive unit, copy number of repetitive unit, and the purity (Legendre et al., 2007) (http://hulsweb1.cgr.harvard.edu/SERV/).

Although MLVA is a rapid, easy to perform, inexpensive, and reproducible genotyping method with high resolution, VNTR loci may evolve too quickly to provide reliable phylogenetic relationships among closely related strains. In Mycobacterium leprae, variation in the VNTR pattern was observed not only between isolates of M. leprae but also between biopsies from the same patient (Monot et al., 2008). Therefore, MLVA is unsuitable for long-term epidemiological surveillance, though it may be useful for tracking outbreaks of bacterial infections (Lindstedt, 2005). In addition, MLVA showed less discriminatory power than multilocus sequence typing (MLST) and PFGE for Enterococcus faecium (Werner et al., 2007). In addition, VNTR loci are not always common in bacterial genomes, such as in Mycoplasma
Denaturing gel electrophoresis (DGE)

In DGE, PCR amplicons with the same length are separated by electrophoresis in polyacrylamide gels in a sequence-dependent manner. The increasing gradient of denaturing components along the polyacrylamide gel opens double-stranded amplicons into single-stranded DNA through melting domains. The melting temperature of these domains is sequence-specific and decreases their mobility. Thus, different sequences will result in different origins of melting domains and consequently in different positions in the gel (Muyzer et al., 1993). Such a gradient is obtained using either denaturing chemicals for denaturing gradient gel electrophoresis (DGGE), or heat for temperature gradient gel electrophoresis (TGGE) and temporal temperature gradient electrophoresis (TTGE) (Fromin et al., 2002). 16S rRNA gene is the most frequent target gene for DGE analysis because it exists in all bacteria and can easily be amplified without prior knowledge of studied strains. Applying statistical method makes DGE fingerprinting technique a promising tool (Fromin et al., 2002). Numerous samples can be analyzed simultaneously, allowing the monitoring of microbial communities.

Constant denaturant capillary electrophoresis (CDCE) is similar to DGGE but has better resolution and can be quantitative (Lim et al., 2001). This method consists of competitive quantitative PCR with group-specific primers targeting a variable region of 16S rRNA gene, followed by separation and quantification of the amplicons in the capillary containing high-molecular-weight linear polyacrylamide gel. CDCE can resolve sequences that differ by as little as a single base pair, and quantification of sequences is extremely sensitive due to the use of a laser-based detection system (Lim et al., 2001; Thompson et al., 2004).

DGE has been widely used in exploring genetic diversity of bacteria since its first application in 1993 (Muyzer et al., 1993). An advantage of DGGE is that selected bands can be sequenced. Thus, the presence of a particular phylotype can be monitored in the samples studied. In addition, DGE can be applied directly to environmental samples or specimens because it is PCR based. However, DGGE requires a specific electrophoresis equipment and it is not possible to run several gels per run.

High-resolution melting (HRM) analysis

Like DGE, HRM analysis is aimed at discriminating DNA alleles through melting temperature analysis. HRM uses real-time PCR amplification and melting curve analysis. The melting profile of a PCR product depending on its GC content, length, sequence, and heterozygosity, is conveniently monitored with saturating dyes that fluoresce in the presence of double-stranded DNA. HRM is not a banding pattern-based method sensu stricto but we classified it in this category of methods because it is based on PCR amplification and allows detection of sequence variants without sequencing or hybridization procedures (Gundry et al., 2003). Using HRM, single nucleotide polymorphisms (SNPs) can be genotyped without probes and more complex regions can be typed with unlabelled hybridization probes. HRM can be combined with various PCR strategies for bacterial strain genotyping. For Mycoplasma synoviae, HRM was used to detect 10 sequence variants in the vlha gene (Jeffery et al., 2007). Using the clustered, regularly interspaced short palindromic repeats (CRISPR) locus of Campylobacter jejuni as target, HRM exhibited a discriminatory power comparable to that of the gold standard genotyping method, i.e. PFGE (Price et al., 2007). When combined to amplification of VNTRs, HRM provided similar results to MLVA for B. anthracis (Fortini et al., 2007). The advantages of HRM are its homogeneity, rapidity (1–5 min), applicability to environmental samples, and the use of closed tubes. It has a similar or superior sensitivity and specificity to methods that require physical separation.

However, HRM requires real-time PCR equipment and may not be equally discriminatory for all bacterial species. In addition, comparability of results among laboratories may be limited.

Enzymatic digestion following DNA amplification

PCR-RFLP

PCR-RFLP, which involves RFLP analysis of a specific locus amplified by PCR, overcomes the shortcomings of RFLP (Wichelhaus et al., 2001) (Fig. 1). Because this method is a PCR-based genotyping method, it can be used with DNA taken directly from human specimens and environmental samples. In addition, the limited number of restriction fragments resulting from RE digestion of the PCR amplicons can be separated and visualized directly by gel electrophoresis without the need for probe hybridization. PCR-RFLP analysis with capillary electrophoresis (PRACE) shows dramatically higher resolution than conventional PCR-RFLP (Ho et al., 2004; Chang et al., 2007b). Therefore, PCR-RFLP

\textit{hyopneumoniae} (Minion et al., 2004), which limits the implementation of MLVA. The size difference in a VNTR locus may not always reflect the real number of tandem repeats because insertions or deletions in the amplified region can also give rise to the same size difference. Therefore, sequencing of the amplicons is necessary in this case. However, with the increase in the number of bacterial genome sequences available for a particular species, it may be possible to improve MLVA through the rational selection of suitable tandem repeat loci and primer design (Johansson et al., 2004; Lindstedt, 2005).
is a simple, rapid, and nonradioactive approach to detect DNA polymorphism. It has been used frequently for typing a variety of bacteria including Brucella species (Al et al., 2005) and M. tuberculosis (Hayward, 1995; Cohn & O’Brien, 1998).

The fact that genetic information in PCR-RFLP comes from a single locus often limits the discriminatory power of this technique. Multilocus-based PCR-RFLP analysis may enhance the discriminatory power.

**Amplified fragment length polymorphism (AFLP)**

AFLP, which was first described by Vos and Zabeau in 1993, combines the accuracy of RE analysis with the precision of PCR, making it a highly sensitive method for detecting DNA polymorphism (Vos et al., 1995). In AFLP analysis, digestion of genomic DNA with two REs (usually MseI and EcoRI) and ligation of restriction fragments with end-specific adapters, is followed by selective amplification. The regions amplified with primers consisting of the adapter sequences at the 5’ end and extended with a variable number of 3’ nucleotides (usually 1–3 bp) chosen by the user (Fig. 1). DNA fragments can be detected by conventional gel electrophoresis or automated DNA sequencer, as in other DNA banding pattern-based genotyping methods (Vos et al., 1995; Mueller & Wolfenbarger, 1999; Zhao et al., 2000) (Fig. 1). The fluorescent AFLP (FAFLP) method uses fluorescently labeled primers for PCR amplification and an automated DNA sequencer for fragment detection. FAFLP shows a much higher resolution than conventional AFLP detection systems, with the ability to detect size differences as small as 1 bp (Lindstedt et al., 2000a; Zhao et al., 2000; Mortimer & Arnold, 2001). Fragments are usually scored as either present or absent, and the presence or absence of bands in the second sample is referred to as polymorphism. Computer analysis of the number and size of fragments generated from each strain allows phylogenetic analysis (Lindstedt et al., 2000b; Mortimer & Arnold, 2001).

The discriminatory power of AFLP analysis is mainly determined by the combination of REs and the number of selective nucleotides in the primers. In silico analysis of bacterial genomic sequences can be used to select the most informative combinations of ERs and primer sequences (Table 2) (Keto-Timonen et al., 2003; Rombauts et al., 2003; Bikandi et al., 2004; Kivioja et al., 2005).

AFLP offers great flexibility in the number of loci that can be amplified simultaneously in one PCR reaction. Typically, 50–100 restriction fragments are coamplified in one fingerprint, depending on the complexity of the genome, the primers used, and the resolution of gel electrophoresis. Thus, relatively few primer pairs are used to visualize a large number of loci in AFLP analysis. In comparison with other genotyping methods, AFLP has been described as being as discriminatory as PFGE, RAPD, and RFLP, and more discriminatory than MLST (Zhao et al., 2000; Feberwee et al., 2005; Gzyl et al., 2005; Melles et al., 2007). It requires far less amount of DNA because multiple bands are derived from across the entire genome. No prior knowledge of DNA sequence is required for AFLP analysis (Vos et al., 1995; Mueller & Wolfenbarger, 1999). In addition, the reproducibility of AFLP is similar to that of other DNA banding pattern-based methods. Finally, AFLP assays are cost-effective and can be automated (Duim et al., 1999) (Applied Biosystems).

The disadvantages of AFLP include the fact that automated analysis equipment may be required due to the many fragments involved and the huge quantity of information generated. Another disadvantage of AFLP that cannot be solved so easily is that the DNA template should not be a mixture of DNAs from various organisms, which prevents the use of AFLP on DNA taken directly from human specimens and environmental samples (Mortimer & Arnold, 2001).

**DNA sequencing-based methods (Fig. 2)**

DNA sequence, the nucleotide order in a genomic fragment, is the original genetic information of an organism and can be used directly for differentiation and phylogenetic analysis of bacterial strains. The most significant advantages of DNA sequencing-based genotyping over DNA banding pattern-based methods is its high reproducibility because it relies on unambiguous DNA sequences that can easily be stored in online databases and compared among laboratories. GenBank, the largest DNA sequence database, stores huge quantities of genomic sequences as well as locus-specific sequences from almost all known bacteria, and is the most frequently used sequence database by molecular microbiologists. DNA sequencing technology has profoundly influenced microbiology (Salser, 1974; Hall, 2007). DNA sequencing-based genotyping of bacteria has significantly contributed to many aspects of genotyping by identifying SNPs, sequence deletions or insertions (including sequence duplications, such as VNTRs), and genes under positive selection.

Various types of differences among DNA sequences may be identified. SNPs are DNA sequence variations that occur when a single nucleotide – A, T, C, or G – differs between members of a given species. This is the most common type of genetic variation (Schork et al., 2000). In addition, deletion, insertion, and duplication events occur widely in bacterial genomes and are regarded as important evolutionary mechanisms. These changes carry the same genotypic weight as SNPs (Lupski & Weinstock, 1992; van et al., 1998; Ogata et al., 2000; Drancourt et al., 2004). DNA sequencing is an efficient way to identify these genetic changes in order to differentiate and phylogenetically classify bacterial strains. Compared with other SNP genotyping methods, sequencing is particularly suited for identification of multiple SNPs within a small region of DNA. In addition to detecting with...
A high sensitivity and specificity sequence differences, DNA sequencing may also allow the evaluation of the evolutionary forces that led to these differences. Relying on the neutral theory of molecular evolution, a high ratio of nonsynonymous over synonymous changes inferred from original DNA sequences can indicate positive selection, which confers a fitness benefit and will thus increase the frequency of that change (Kimura, 1977). A well-known example of positive selection in action is the development of antibiotic resistance in bacteria. When exposed to antibiotics, most bacteria die quickly, but some may experience mutations that make them less susceptible to treatment. Positive selection acts on any mutation that contributes to antibiotic resistance and increases the frequency of those mutations (Aminov & Mackie, 2007). In addition, DNA sequencing-based identification of genes under positive selection is also an effective approach to identify virulence in bacteria (Chen et al., 2006).

**Two main strategies in DNA sequencing**

Currently, there are two main strategies for DNA sequencing: the traditional Sanger method and the newly developed pyrosequencing method.

**The Sanger method**

The Sanger method is also referred to as dideoxy sequencing or chain termination DNA sequencing. In this method, the DNA sequence of a single-stranded template DNA is determined using DNA polymerase to synthesize a set of polynucleotide fragments of different lengths through the use of dideoxynucleotides that interrupt the elongation step of DNA amplification (Sanger et al., 1977). Since its description in 1974, the Sanger method has become the most widely used DNA sequencing method. However, recently, new DNA sequencing methods offering much higher throughput at lower cost have revolutionized DNA sequencing. Among these developments, pyrosequencing is the most remarkable.

By comparison with pyrosequencing, Sanger sequencing is more expensive and requires more DNA quantity (several micrograms) and a cloning step for large sequencing projects such as genome sequencing.

**Pyrosequencing**

In contrast to the conventional Sanger method (Sanger et al., 1977), pyrosequencing is a nonelectrophoretic sequencing method based on real-time quantitative detection of pyrophosphate released following nucleotide incorporation.
into a growing DNA chain during DNA synthesis (Nyren, 1987; Ronaghi et al., 1998). The current state of pyrosequencing technology leads to an c. 100-fold increase in throughput and a 10-fold reduction in cost over the conventional Sanger method. In 2004, the Roche (454) company first commercialized a high throughput sequencer using the pyrosequencing technology. Following shearing, genomic DNA fragments are hybridized to the surface of agarose beads, and then amplified by emulsion PCR and sequenced, without any cloning step (Margulies et al., 2005). The current output of the Roche instrument is c. 100 Mb of sequence data per 7-h run, with an average read length of c. 250 bp. More recently, Illumina ( Solexa) developed a pyrosequencer that uses glass-attached oligonucleotides that are complementary to specific adapters previously ligated onto DNA library fragments. Sequences are obtained by amplification using an isothermal polymerase, and Sanger-like sequencing. The Illumina method produces c. 40–50 million sequence reads of 32–40 bp simultaneously. The latest pyrosequencing competitor is Applied Biosystems. Their SOLID instrument uses a sequencing process catalyzed by DNA ligase. Oligonucleotide adaptor-linked DNA fragments are coupled with magnetic beads that are covered with complementary oligonucleotides. Each bead-bound DNA fragment is amplified using emulsion PCR, and sequencing is obtained using a thermostable ligase. The SOLID instrument produces 3–4 Gb of sequence data with an average read length of 25–35 bp.

To date, pyrosequencing has been used for several applications of genotyping. Sequencing the highly polymorphic regions in 16S rRNA gene by pyrosequencing demonstrated to be a rapid and inexpensive approach for identification and differentiation of bacteria (Jonasson et al., 2002, 2007). In a recent article, Dethlefsen et al. (2008) identified 5700 different bacterial taxa in stool specimens from patients before and after antibiotic therapy. Detection of antibiotic resistance has also benefited from pyrosequencing. Mutation detection in antibiotic resistance genes of M. tuberculosis by pyrosequencing has been shown to be a cost-effective, highly accurate, and high-throughput method for drug resistance screening (Arnold et al., 2005; Jureen et al., 2006). Grouping, typing, and subtyping bacteria with specific virulence or of epidemiological interest may also be obtained. Pyrosequencing-based SNP detection in three genes, siaD, porB, and porA, allow a rapid and specific characterization of Neisseria meningitidis strains (Diggle & Clarke, 2004). Pyrosequencing of genomic fragments has been used to genotype Neisseria gonorrhoeae and B. anthracis (Diggle & Clarke, 2004; Uneno et al., 2004; Wahab et al., 2005).

In comparison with Sanger sequencing, pyrosequencing, although being less expensive and having a higher output, is limited by short read lengths (25–250 bp), small numbers of samples that can be run simultaneously, and difficult sequence assembly. The ability to generate DNA sequences rapidly and the potential for scale-up make pyrosequencing a valuable option for SNP genotyping (Clarke, 2005).

Gene sequencing

16S rRNA gene

16S rRNA gene is highly conserved among bacteria and other kingdoms because rRNA is essential for the survival of all cells due to its involvement in protein synthesis (Woese, 1987; Hillis et al., 1991). Consequently, amplification and sequencing of the 16S rRNA gene is widely used for identification and phylogenetic classification of prokaryotic species, genera, and families (Woese, 1987; Stackebrandt & Goebel, 1994). At the strain level, however, 16S rRNA gene is too conserved to be useful, with the exception of few species such as N. meningitidis (Sacchi et al., 2002), Mycoplasma agalactiae and Mycoplasma bovis (Konigsson et al., 2002). In practice, 16S rRNA gene sequencing is mostly used for bacterial species identification.

Other genes

Some less-conserved genes, especially those under positive selection, have often been used for species identification and, in some rarer cases, for bacterial strain typing. Most of these genes encode bacterial surface proteins or virulence factors. The choice of appropriate genes may vary according to the species.

Encoding the β subunit of RNA polymerase, rpoB appears to exist in only one copy in most bacteria and has widely been used for bacterial species identification and subtyping (Mollet et al., 1997; Adekambi et al., 2006). The rpoB sequence offers higher discriminatory power than 16S rRNA gene (De & De, 2004). DNA sequencing of the variable regions in the slpA gene, which encodes a surface layer protein of C. difficile, were found to be an ideal alternative to serotyping because the sequence variations are identical within a given serogroup but divergent between serogroups (Karjalainen et al., 2002; Kato et al., 2005). The set genes (set2, set5, and set7), which encode S. aureus superantigen-like proteins, were found to exhibit a higher discriminatory power than MLST based on seven genes (Aquije-Alves et al., 2006). Sequence analysis of the ompA gene, which encodes a major antigenic outer membrane protein of Rickettsia species, was demonstrated to be an important tool for identifying and subtyping these bacteria (Fournier et al., 1998, 2003). Sequencing of the X region of the protein A gene (spa) is widely used for subtyping methicillin-resistant S. aureus (MRSA) strains and proved to be nearly as discriminatory as PFGE (Shopsin et al., 1999; Koreen et al., 2004; Deurenberg et al., 2007). For more informative
discrimination of MRSA strains, spa is strongly suggested to be combined with other gene loci or other genotyping methods (Hallin et al., 2007; Kuhn et al., 2007).

Although genes encoding surface proteins or virulence factors evolve very quickly, interpretation of the typing results may be misleading, especially in evaluating the bacterial population structure or in the case of a long-term epidemiological survey. Combination of genes encoding surface antigens or virulence factors with housekeeping genes may be a rational approach for bacterial strain typing, particularly for bacterial species with a high rate of genetic recombination (Cai et al., 2002).

**MLST**

MLST is a method using DNA sequencing to uncover allelic variants in several conserved genes (usually seven genes), and is currently one of the most popular genotyping methods for characterizing bacterial strains (Maiden et al., 1998). MLST examines multiple housekeeping genes whose sequences are constrained because of the essential function of the proteins they encode; the variation observed in these sequences is therefore neutral or nearly neutral.

Typically, fragments of 450–500 bp of seven genes are sequenced and each different sequence for a given gene is attributed a number. Each strain is, therefore, assigned a seven-number allelic profile designated as sequence type (ST) (Maiden et al., 1998; Spratt, 1999). MLST is suitable for long-term investigation of bacterial population structures, particularly when subtyping bacterial species with a high rate of genetic recombination, such as N. meningitidis (Maiden et al., 1998), S. pneumoniae (Enright & Spratt, 1998), and Enterococcus faecalis (Ruiz-Garbajosa et al., 2006). MLST analysis indicates that recombinational replacements in many species contribute more to clonal diversification than do point mutations (Feil & Spratt, 2001). DNA sequences are easily stored in online databases, which allow convenient exchange of strain typing data both within and between laboratories. These databases also facilitate the global epidemiological survey of bacterial infections (Feil & Spratt, 2001; Jolley et al., 2004; Aanensen & Spratt, 2005). So far, MLST has been applied to more than 23 bacterial species and is regarded as a reference genotyping method for many bacteria. Detailed information can be found on the MLST homepage, http://www.mlst.net (Table 2).

However, MLST also has drawbacks. First, alleles are assigned to a numbering system that is not representative of the actual gene sequence, which makes the phylogenetic analysis of tested strains poorly credible (Clarke, 2002). Second, the use of highly conserved housekeeping genes in MLST often fails to detect the variability of closely related strains. Finally, sequencing of seven genes is costly and time consuming.

**Sequencing of noncoding DNA**

16S–23S rRNA gene internal transcribed spacer (ITS)

ITS, a special genomic region separating 16S and 23S rRNA genes in prokaryotic microorganisms, consists mainly of noncoding sequences and is the first noncoding sequence used for bacterial strain typing (Gurtler & Stanisich, 1996; Sadeghifard et al., 2006). ITS varies not only in sequence and length but also in the number of alleles and their positions on the chromosome (Garcia-Martinez et al., 1999; Sadeghifard et al., 2006). The high polymorphism of the ITS locus makes it useful for identifying and subtyping bacteria (Gurtler & Mayall, 1999). While size variation of ITS resulting from either the amplification or digestion of amplicons by REs is useful for bacterial strain typing, sequencing ITS can provide a more comprehensive analysis of polymorphisms for strain typing. However, more than one copy of ITS is often found in the bacterial genome, and those ITS sequences within a given strain may differ from one another, making direct DNA sequencing of ITS in some bacterial species difficult. Cloning-based sequencing can resolve this problem but makes it a time-consuming method (Garcia-Martinez et al., 1999).

One advantage of using ITS for studying phylogeny, molecular evolution, or population genetics is that ITS can be amplified easily by picking primers in the conserved region of 16S rRNA and 23S rRNA genes (Gurtler & Stanisich, 1996; Gurtler & Mayall, 1999). In comparison with 16S rRNA gene, ITS is more variable and exhibits greater resolution for subtyping bacteria at the strain level. For the Bartonella genus, partial ITS amplification and sequencing has proven to be a sensitive tool for differentiating B. henselae, Bartonella claridgeiae, and Bartonella bacilliformis isolates (Birtles et al., 2000; Houpikian & Raoult, 2001). However, ITS is mainly used for species or subspecies identification, and less for strain typing. In addition, it is not suitable for bacteria from the order Rickettsiales because their 16s and 23s rRNA genes are not contiguous.

**Multispace typing (MST)**

MST, a novel genotyping system first applied to Y. pestis in 2004, is a DNA sequencing-based genotyping method that uses as typing markers highly variable intergenic spacers (Drancourt et al., 2004). MST assumes that noncoding DNA sequences, which are subject to less selection pressure than genes, vary more than genes and so are preferable for bacterial strain typing (Drancourt et al., 2004). The intergenic spacers used as typing markers in MST are chosen to be those noncoding sequences varying the most between aligned genomes of bacterial strains within one species, or between closely related species if only one genome sequence is available for a given species.
For example, the genome sequences of the closely related \textit{B. henselae} and \textit{Bartonella quintana} were compared to select the most variable intergenic spacers, based on the hypothesis that the intergenic spacers varying the most between closely related species would also vary the most among all strains of the species (Li et al., 2006b). MST based on six highly variable intergenic spacers identified 19 MST types among 36 \textit{Y. pestis} strains from three biovars detected in dental pulp from patients deceased from plague in the second and third pandemics (Drancourt et al., 2004). This work demonstrated the great potential of MST for bacterial strain typing.

With rapid accumulation of bacterial genome sequences and the increasing number of species with more than two genome sequences available, the highly variable intergenic spacers selected by genome comparison are becoming more representative and powerful for bacterial strain typing. After amplification and sequencing, any DNA sequence variation in a spacer can provide a unique ST. Combination of STs from each studied spacer provides an MST genotype. The phylogenetic organization of studied strains, however, was inferred from concatenation of all selected intergenic spacers (Fig. 3).

MST has been applied successfully to several human pathogens, including \textit{Y. pestis} (Drancourt et al., 2004), \textit{Rickettsia conorii} (Fournier et al., 2004), \textit{Rickettsia prowazekii} (Zhu et al., 2005), \textit{Rickettsia sibirica} (Fournier et al., 2006), \textit{Coxiella burnetii} (Glazunova et al., 2005), \textit{B. henselae} (Li et al., 2006b, 2007), \textit{B. quintana} (Foucault et al., 2005) and \textit{Tropheryma whipplei} (Li et al., 2008). Comparison of MST and MLST showed the former to have higher resolution (Fournier et al., 2004; Li et al., 2006b). Because MST is a PCR-based technique, it has the potential to be used directly on noncultured samples. Indeed, MST has been successful in subtyping \textit{Y. pestis} from dental pulps c. 1500 years old (Drancourt et al., 2004). Recently, MST was also used to evaluate the genetic diversity of \textit{B. henselae} and \textit{R. conorii} directly from human specimens (Li et al., 2007). Another advantage of MST is that primers can be chosen in conserved regions of the flanking genes, making amplification of highly variable spacers easy.

While intergenic spacers are considered as rapidly evolving markers, MST has succeeded in establishing reliable correlations of MST genotypes with geographic distribution, clinical manifestations, and epidemiology of strains (Drancourt et al., 2004; Fournier et al., 2004; Zhu et al., 2005; Arvand & Viezens, 2007; Li et al., 2006b, 2007). In addition, MST enabled

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{mst_diagram.png}
\caption{Outline of MST.}
\end{figure}
population structure of currently the most suitable genotyping tool for evaluating the diversity of this complex species and suggests that MST is from human specimens offers further insight into the genetic recent study applying MST to database, MST-Rick, which provides a localBLAST program, enables anyone to compare his own spacer sequences and determine MST genotypes (Table 2).

**Genome sequencing**

By giving access to the complete genetic content of bacterial strains, genome sequencing is the ultimate genotyping method. Both the Sanger and pyrosequencing methods are currently used for genome sequencing, either separately or combined. The former technology has been the mostly used worldwide since the genome of *Haemophilus influenzae* was sequenced in 1995 (Fleischmann *et al.*, 1995). Genome sequencing using the Sanger method requires the construction of a cDNA library and produces 650–800-bp reads, with a maximum output of 0.44 Mb per 7-h run (Fleischmann *et al.*, 1995; Binnewies *et al.*, 2006). In contrast, pyrosequencing produces short reads (25–250 bp) but with an output of up to 3–4 Gb per run.

Because of the short read length, making sequence assembly difficult, pyrosequencing may currently be preferred for genome resequencing than for *de novo* assembly, although a combination of Sanger sequencing and next generation sequencing has been proposed as a valuable alternative for this latter purpose. Recently, pyrosequencing was used for rapid genome sequencing of *F. tularensis*, a potential bioterrorism bacterial species, in order to identify the detailed genetic differences between a clinically isolated strain and other *F. tularensis* strains (La *et al.*, 2008). Pyrosequencing also proved to be useful for metagenomic studies (Eckburg *et al.*, 2005).

However, although sequencing techniques have been greatly improved since the first bacterial genome was sequenced in 1995 (Fleischmann *et al.*, 1995), with the current possibility of obtaining a genome sequence from a single bacterial cell (Ishoey *et al.*, 2008), genome sequencing remains as yet too expensive and time demanding to be used in routine genotyping. In contrast, both non-sequence-based and sequence-based genotyping methods may benefit from the study of genomic sequences, such as *‘in silico’* design of macrorestriction profiles for enzymes used in PFGE, RFLP analyses or microarrays, identification of SNPS, tandem repeats (VNTRs), and variable intergenic spacers (MST).

**DNA hybridization-based methods (Fig. 4)**

DNA hybridization is widely used for detecting DNA mutations because it requires sequence complementarity to occur. Two main elements, probes and targets, are involved in DNA hybridization. Probes are DNA fragments of known sequences, whereas targets are the free nucleic acids whose identity and abundance are detected by hybridization with fluorescently labeled probes based on their complementarity to the probes.

A DNA array enables genomic DNA to be tested for its ability to hybridize to hundreds or tens of thousands of DNA fragments or oligonucleotides (spots) arrayed on a substrate. Substrates can be membranes, glass, plastic, silicon wafers, or metal alloys. An array is a very powerful tool for studying the transcriptosome as well as the genetic diversity of bacteria. Two classes of DNA arrays, microarrays and macroarrays, that differ in the size and number of spots on the supports are used. Macroarrays are usually printed on nylon membranes c. 8 × 12 cm in size and contain up to 5000 spots, each > 300 μm in diameter. A microarray, in contrast, is much denser than a macroarray, with up to 1 000 000 spots per array.

**DNA macroarrays**

Macroarrays offer a rapid, specific, and cost-efficient genotyping method without the need to purchase expensive
equipment, and has proven particularly effective for detecting genes involved in antibiotic resistance (Johansen et al., 2003; Zhang et al., 2007). In a recent study, one group rapidly detected mutations in rpoB, katG, inhA, and ahpC, which are associated with rifampin and isoniazid resistance in M. tuberculosis. The macroarray approach in this study showed high sensitivity and specificity and could even be directly applied to DNA extracted from clinical specimens (Zhang et al., 2007).

Spoligotyping, detection of the polymorphism in the direct repeat variants (DRVs) loci of the mycobacterial chromosome, is a PCR-based macroarray technique for simultaneous detection and differentiation of MTC bacteria (Groenen et al., 1993). The DVR loci, comprising a cluster of well-conserved 36-bp direct repeats (DRs) and spacers of 35–41 bp interspersed throughout DRs (van Embden et al., 2000), show considerable strain-to-strain polymorphism (Fang et al., 1998). Amplification of the DRVs using primers complementary to the DRs, followed by the hybridization of amplicons to the oligonucleotides derived from the spacer sequences (probes), generates distinct spoligotyping patterns. Spoligotyping has been proven to be a rapid, reproducible, informative, and effective technique for use on DNA taken directly from human specimens or environmental samples without the need for prior culture. It has been widely used in epidemiological studies of M. tuberculosis infections, especially in low-income countries (Goyal et al., 1997; Hayward & Watson, 1998; Heyderman et al., 1998). An international spoligotyping database, SpolDB4, suggested the existence of geographical genetic clones within MTC populations and provided a large-scale conceptual framework for the global TB Epidemiologic Network (Brudey et al., 2006) (Table 2). For Corynebacterium diphtheriae, a macroarray-based spoligotyping method based on hybridization analysis of two CRISPR loci differentiated 20 strains into three spoligotypes (Mokrousov et al., 2008). Eighty S. aureus isolates were classified within 52 genotypes using a DNA macroarray containing 465 intragenic amplicons (Trad et al., 2004).

By comparison with microarray, DNA macroarray is less expensive but suffers from a lower discriminatory power due to the small number of characters studied.

DNA microarrays

DNA microarrays are specially treated microscope slides (chips) that carry an ordered mosaic of sequences representing most or all of the genes of an organism. Since its development in 1995 (Schena et al., 1995), the DNA microarray technique has quickly become an efficient tool for investigating the bacterial transcriptome (DeRisi et al., 2000; Khodursky et al., 2000) and genetic diversity (Hacia, 1999; Garaizar et al., 2006). In contrast to Southern blotting and macroarray techniques, where a limited number of probes are transferred onto nitrocellulose or nylon membranes, a DNA microarray includes tens of thousands of probes arrayed directly on the support and it therefore constitutes a high-throughput genotyping method (Garaizar et al., 2006). However, compared with other locus-limited genotyping methods, the DNA microarray remains more expensive. In addition, satisfactory quality controls are difficult to achieve in DNA microarray analysis because many factors affect nucleic acid hybridization reactions. The probe design in a DNA microarray is mainly based on reference DNA sequences, including prior knowledge of the polymorphisms derived from other genotyping studies. However, comparison of multiple genomes within a given bacterial species reveals numerous ‘accessory genes’ (Fraser-Liggett, 2005; Tettelin et al., 2005; Lefebure & Stanhope, 2007), some of which may be absent in the reference genome. As a result, DNA microarray analysis may underestimate the genetic diversity among the bacterial population.

There are two kinds of DNA microarrays: cDNA microarray and oligonucleotide microarray, which differ in the size of the probes used. A cDNA microarray uses cDNA as probes and is often used to identify the presence or absence of genes in a technique called comparative genomic hybridization. An oligonucleotide microarray, on the other hand, uses short oligonucleotides as probes and is usually used to identify SNPs.

cDNA microarrays

The probes in a cDNA microarray are complete genes amplified by PCR or taken from cDNA libraries (Duggan et al., 1999). A whole-genome microarray is a powerful tool for investigating comprehensive variation at the genomic level (Dorrell et al., 2001; Garaizar et al., 2006). In addition, cDNA microarrays are useful for identifying genes associated with serotypes, virulence, antibiotics resistance, epidemics, and host diversity. These kinds of genes are difficult to identify by most other genotyping methods (Garaizar et al., 2002; Borucki et al., 2004). Once these genes are identified by microarray analysis, they may serve as probes for individual discrimination of unknown bacteria. In a powerful illustration of cDNA microarrays, an array including a series of housekeeping genes, virulence genes, and antibiotic resistance-associated genes was used for the simultaneous identification, characterization, and differentiation of S. aureus, E. coli, and Pseudomonas aeruginosa (Cleven et al., 2006). Specific loci-based cDNA microarrays are cheaper, easier to use, and give results that are easier to interpret compared with whole-genome cDNA microarrays.

Because cDNAs or partial cDNAs were used as probes in the cDNA microarray and the hybridization is often under conditions of lower stringency, a family of genes that share
significant sequence identity (> 70%) may cross-hybridize with the probes, leading to unspecific hybridization.

**Oligonucleotide microarrays**

The probes of an oligonucleotide microarray are short nucleic acid sequences with c. 30–70 nucleotides spotted or directly synthesized in high density on the solid substrate (up to 1 000 000 oligonucleotides per chip) (Lockhart et al., 1996; Iwasaki et al., 2001; Iwasaki et al., 2002). All of the possibilities at the polymorphic sites within an oligonucleotide can be deduced from the results of an oligonucleotide microarray. In fact, well-designed oligonucleotide microarrays using short nucleotides (10–20 mers) as probes can provide investigators with actual sequence data. These are termed resequencing microarrays (Hacia, 1999; Dunman et al., 2004; Zwick et al., 2005). DNA resequencing helps determine the order of nucleotides in a DNA fragment by hybridization with the genomic reference sequence (Zhan & Kulp, 2005). Oligonucleotide microarray-based resequencing is a high-throughput method for SNP detection and short sequence deletions by dividing a gene sequence into a number of oligonucleotide probes (Lockhart et al., 1996; Kato-Maeda et al., 1996; Iwasaki et al., 2002). The whole-genome resequencing microarray has been well-designed oligonucleotide microarrays using short nucleotides (10–20 mers) as probes can provide investigators with actual sequence data. These are termed resequencing microarrays (Hacia, 1999; Dunman et al., 2004; Zwick et al., 2005). DNA resequencing helps determine the order of nucleotides in a DNA fragment by hybridization with the genomic reference sequence (Zhan & Kulp, 2005). Oligonucleotide microarray-based resequencing is a high-throughput method for SNP detection and short sequence deletions by dividing a gene sequence into a number of oligonucleotide probes (Lockhart et al., 1996; Kato-Maeda et al., 1996; Iwasaki et al., 2002). The whole-genome resequencing microarray has been shown to be a rapid, highly discriminatory, and cost-effective method for strain typing compared with the Sanger method of sequencing whole genomes (Mockler et al., 2005; Zwick et al., 2005).

Locus-specific oligonucleotide microarrays are also used for genotyping. The complexity of *E. coli* serotypes makes antibody-based typing very difficult. Oligonucleotide microarrays containing the *E. coli* genes for O- and H-antigens are valuable alternatives to classical serotyping because they allow single colonies to be processed within one working day (Liu & Fratamico, 2006; Ballmer et al., 2007).

Oligonucleotide microarrays are more discriminatory than a cDNA microarray due to the use of short oligonucleotides as probes. A cDNA microarray, however, can be used with environmental samples or human specimens if the cDNA is amplified beforehand (Zhou, 2003; Roth et al., 2004).

**Selection of appropriate genotyping methods**

An ideal genotyping method would have the following characteristics: applicable to all isolates (validity), capable of differentiating unrelated isolates (discriminability), reproducible both within and between laboratories (reproducibility), rapid, cost efficient, and easy to perform. However, among the many genotyping methods that are currently available, no single method is universally ideal because each one has both strengths and weaknesses, as discussed above.

Two main objectives of genotyping may be studied using distinct methods. Surveillance of local outbreaks of bacterial infections may be obtained using fast-evolving markers such as those studied by MST (Fournier et al., 2004; Li et al., 2006b) or MLVA (van et al., 1998). In contrast, for long-term epidemiological studies, or population studies, conserved and stable marker-based typing methods such as MLST, PFGE, and whole-genome sequencing may be more appropriate (Cooper & Feil, 2004) (Fig. 5).

Discriminatory power, or resolution, which refers to the ability to distinguish bacterial strains from a given species, is one of the most important criteria for selecting a genotyping method. For example, bacterial populations may appear to be identical using phenotypic or low-resolution genotyping analysis, but different using high-resolution genotyping methods. Generally, methods studying fast-evolving markers, such as MST, MLVA, DGE, or HRM are more discriminatory than those studying slow-evolving markers such as MLST. Whole genome-based typing methods such as PFGE, AFLP, RFLP, and genome microarray are more discriminatory than locus-limited typing methods such as ribotyping and PCR-RFLP. However, the discriminatory power is also influenced by many factors, such as the enzymes and primers used and the enzymatic and amplification conditions. In addition, most genotyping methods are strain sensitive, and their discriminatory power may vary according to the species studied.

Reproducibility, the ability to obtain the same result whenever the same test is performed, is also important both within and between laboratories, to allow global epidemiological surveillance of bacterial infections. In addition to methodological reproducibility, the stability of typing methods should also be considered when choosing a typing method. Generally, results from banding pattern-based methods are less reproducible and comparable than sequence-based methods.

Another criterion that should be considered is the purity of the strain(s) to be studied. Given that > 99% of environmental bacteria cannot be cultured, PCR-based genotyping methods, namely PCR-RFLP, MLVA, DGE, HRM, MLST, MST, and cDNA microarrays, are better suited than other methods for polymicrobial environmental or human samples (Kellenberger, 2001; Rappe & Giovannoni, 2003) (Fig. 5). In contrast, although they are PCR-based, AFLP, RAPD, and REP-PCR, using nonspecific primers, are poorly applicable to polymicrobial samples. PFGE, RFLP, oligonucleotide microarrays, and genome sequencing,
require large amounts of high-quality genomic DNA, which means that they are not suited for typing bacteria in polymicrobial samples as well (Fig. 5).

Other important considerations when choosing genotyping methods include the cost of reagents, equipment, and staff; the time necessary to obtain results; and the difficulty of the technique. Generally, in comparison with unambiguous DNA sequencing-based methods, which require an expensive DNA sequencer, methods based on DNA banding pattern are less expensive but are also less discriminatory and reproducible. The DNA microarray is more likely to produce high-throughput strain typing results, but it is also more expensive than banding pattern- and sequencing-based methods.

**Current challenges in bacterial strain typing**

In molecular biology, a bacterial species is defined as a relatively homogeneous population sharing at least 70% hybridization of DNA (Wayne et al., 1987; International Committee on Systematic Bacteriology, 1988), and over 97% 16S rRNA gene sequence similarity (Stackebrandt & Goebel, 1994). Over the last decade, as a result of the increasing number of bacterial genome sequences available, comparative genomics demonstrated that genetic diversity within bacterial species was far greater than previously thought (Fraser-Liggett, 2005; Tettelin et al., 2005; Binnewies et al., 2006; Lefebure & Stanhope, 2007). This also introduced the concept of bacterial pan-genome, including the ‘core genome,’ made of genes shared by all strains in a species, and the ‘dispensable’ or ‘flexible genome’ comprising genes unique in one or some strains. Two strains of the same bacterial species may differ in gene content by as much as 30% and a species may show endless diversity (Fraser-Liggett, 2005; Tettelin et al., 2005). Genomic variations may also involve sequences from genes shared by the majority of strains within a species, with possibility of intragenic variable and conserved fragments. The concept of pan-genome highlights a great challenge of bacterial strain typing, that is, how can a strain typing method evaluate the genetic diversity of a bacterial species? As an example, the genomic diversity of *Streptococcus agalactiae* based on whole genome analysis was found to be inconsistent with MLST sequence types (Medini et al., 2005). This discrepancy may be explained by the fact that housekeeping genes used in MLST are selected from the core genome rather than the pan-genome. Because intraspecies diversity results from point mutations, insertions/deletions, and genetic recombination, the recombination rate should be taken into account when designing a typing procedure for a given species. Some bacteria, such as streptococci, meningococci, *Helicobacter pylori*, and salmonellae are highly dynamic, with high rates of genetic recombination. Therefore, typing methods...
studying few genetic markers may not be sufficient for genotyping these bacteria at the strain level (Medini et al., 2005), and comparison of genomic sequences from several strains may facilitate selection of appropriate targets for a given species.

Another challenge in microbiology is species identification. Species identification may be regarded as bacterial typing at the species level. Some strain genotyping approaches based on relatively conserved genetic markers such as MLST, ribotyping, 16S rRNA gene-/ITS-sequence typing have been used both for species identification and strain typing. In contrast, fast-evolving markers such as VNTRs, noncoding DNA, and horizontally transferred genes, are not suitable for species identification and may misclassify strains in a wrong species. Polymicrobial specimens, especially those containing unculturable strains, are especially problematic for species determination. Metagenomic studies, notably those using highly conserved genes such as 16S rRNA gene, may be particularly suited for bacterial strain identification in these samples (Fig. 5). After species identification, appropriate strain typing methods may be applied to studied strains.

Unraveling the association between genotypes and phenotypic or epidemiological traits such as bacterial virulence, antibiotic resistance, host adaptation, geographic origin, pandemic, or epidemic outbreaks, is another major challenge of bacterial strain genotyping. Although genotyping is highly discriminatory, it is still important to take epidemiological traits into account when evaluating putative outbreaks of bacterial infections. Otherwise, a false outbreak may be announced. While most typing markers are not directly virulence or antibiotic resistance genes, it is possible to find correlations between specific genotypes and phenotypic or epidemiological traits by analyzing the genetic distances between strains through phylogeny. In addition, the inclusion of a large collection of strains and different typing markers is also helpful for linking genotypes with phenotypical and epidemiological traits (Caws et al., 2008). While phylogenetic profiling is useful for correlating phenotype to genotype as well as for the annotation and identification of functionally related genes, this informative computerized method is currently based on genome sequence analysis (Antonov & Mewes, 2008; Benfey &Mitchell-Olds, 2008). With increasing accumulation of bacterial genotypes generated by various approaches, application of powerful phylogenetic profiling for linking genotypes identified by strain typing with phenotypical and epidemiological traits is a great challenge in the genomic era.

Finally, will whole genome sequencing replace other genotyping tools? Whole genome sequencing provides the full genetic information of studied strains and thus appears as an ideal strain-typing method. Currently, the limitations of whole genome sequencing for genotyping include the cost, requirement for large amounts of high-quality genomic DNA, and the small read length of high throughput sequencing methods. However, due to the continuous reduction of costs, wider application of pyrosequencing, development of strategies that enable sequencing from a single cell, and anticipation that read length will increase, this strategy may become a first-line genotyping tool in the next future.

**Concluding remarks**

Bacterial strain-typing methods have been rapidly evolving over recent years. The field has been moving toward computer-assisted design/analysis and increasing automation, resolution, throughput, and reproducibility. In the genomic era, bacterial genotyping benefits increasingly from bioinformatics, such as computer-based selection of typing markers; design of typing strategy; storage, exchange, and comparison of genotyping data (DNA banding patterns, DNA sequences, microarray profiles); development of improved analysis tools; and phylogenetic analysis. Genotyping databases containing banding patterns, DNA sequences, or DNA microarray profiles are convenient for interlaboratory comparison, retrospective studies, and long-term epidemiological surveillance of bacterial infections. Locus-specific typing methods now use multiple loci rather than only one, and whole-genome analysis allowed the selection of typing markers to become gradually less empirical and more rational. In addition, newly developed detection systems for discriminating DNA fragments, such as flow cytometry-RFLP, combines the reproducibility and reliability of RFLP with the speed and sensitivity of flow cytometry, which dramatically increases the speed and resolution of conventional genotyping (Larson et al., 2000; Ferris et al., 2004). Other combined methods have recently emerged and have been found to be more functional than the component methods on their own (Chang et al., 2007a).

Typing results obtained from one genotyping method can be used to improve others or even to design novel typing systems. As discussed above, SNP detection is based on prior knowledge of DNA polymorphism generated by other genotyping methods based on DNA sequence. Genome sequencing of one bacterial strain not only provides detailed information of intraspecies diversity but also enables selection of locus-specific typing markers and rational design of genotyping strategies. As examples, MLVA and MST (Drancourt et al., 2004; van, 2007) have benefited from the availability of an increasing number of sequenced bacterial genomes, allowing selection of highly polymorphic VNTR loci or intergenic spacers. Genome sequences also facilitate the rational choice of primers for PCR-based genotyping methods. In addition, genomic sequences are necessary for probe design in DNA microarray-based genotyping (Lucchini et al., 2001; Zhou, 2003).
Bacterial species are relatively heterogeneous entities whose strain diversity is important to understand in order to manage the treatment and epidemiology of infections. Currently, this diversity may be evaluated using genotyping methods classified in three categories: DNA banding pattern-based, DNA sequencing-based, and DNA hybridization-based methods. Each genotyping method offers advantages and disadvantages. Consequently, choosing an appropriate genotyping method is not easy and may depend on the objective studied and several variables including typeability, reproducibility, resolution, cost, ease of execution and interpretation, and time to obtain results. In addition, the results of bacterial genotyping should always be analyzed in the light of clinical and epidemiological data, especially when evaluating novel genotyping methods. Genome sequencing, in addition to enabling the rational selection of adequate targets for many genotyping methods, appears today, as the ultimate genotyping method, and current technical improvements and reduction of costs and delays may favor this strategy. However, it is difficult to predict whether this method will become the norm and replace other genotyping methods.

References


deBruijn FJ (1992) Use of repetitive (repetitive extragenic palindromic and enterobacterial repetitive intergeneric consensus) sequences and the polymerase chain reaction to
Bacterial genotyping in the genomic era

911


Keto-Timonen RO, Autio TJ & Korkeala HJ (2003) An improved amplified fragment length polymorphism (AFLP) protocol for...


W. Li *et al*. 2009 Federation of European Microbiological Societies Published by Blackwell Publishing Ltd. All rights reserved