4 Multiple Locus VNTR (Variable Number of Tandem Repeat) Analysis

Gilles Vergnaud, Christine Pourcel

4.1 Introduction

The present chapter will review the current state of the art in the field of bacterial strain typing through the use of tandem repeat polymorphism. We will first go through a brief overview of multiple locus VNTR analysis (MLVA) typing and then describe how to set-up or enrich a MLVA assay. We will also review representative examples of the currently proposed MLVA assays and discuss the methods used for MLVA data analysis. Finally, we will compare MLVA to other approaches, and discuss issues related to standardisation and possibilities offered by the internet in terms of shared databases for MLVA (MLVA web services).

4.2 MLVA Origins

The recognition of tandem repeats as often highly polymorphic loci is more than 20 years old. In the early 1980s, a number of laboratories trying to develop the first drafts of the human genetic map were characterising so-called restriction fragment length polymorphisms (RFLPs). Southern blots carrying DNA from large human families were systematically hybridised with DNA probes recognising a single locus in the human genome. RFLPs were bi-allelic and the maximum polymorphism information content (PIC) index (calculated as 1.0 minus the sum of the squares of allelic frequencies) was 0.5. One probe yielded an astonishing result, with multiple alleles and a PIC value well above 0.5. Detailed molecular analysis demonstrated that the observed polymorphism was the result of variations in the number of units in a tandem repeat. The first tandem repeats characterised were satellite DNAs. These tandem repeats cover megabases of DNA; and they

Gilles Vergnaud: Division of Analytical Microbiology, Centre d'Etudes du Bouchet, B.P. 3, 91710 Vert le Petit, France, E-mail: gilles.vergnaud@igmors.u-psud.fr

Christine Pourcel: GPMS laboratory, Institute of Genetics and Microbiology, University Paris XI, 91405 Orsay cedex, France

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

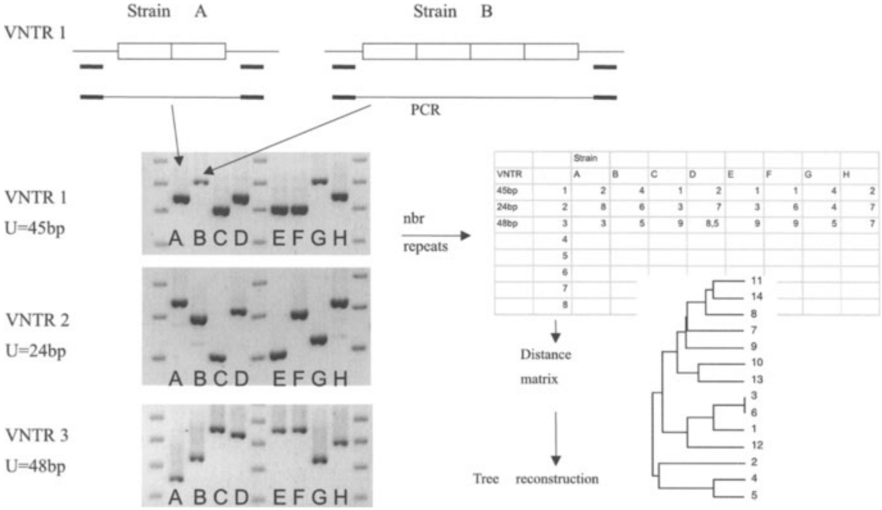
represent a sufficiently large portion of some eukaryote genomes to be able to produce a “satellite” band on caesium chloride density gradients, as soon as the repeat unit has a nucleotide composition slightly different from the genome average. For this historical reason, the small tandem repeats (in the kilobase range) analysed by Southern blotting were called minisatellites and, later, even smaller structures were called microsatellites. Tandem repeat structures cover a number of different situations in terms of origin, mode of evolution, mutation rate and function (when identified). When used for typing purposes, one key feature is the associated length polymorphism. Polymorphic tandem repeats are most often called VNTRs, which includes polymorphic mini- and microsatellites (for a review, see Vergnaud and Denoeud 2000). Towards the end of the 1980s, the advent of the PCR technology made possible the large-scale typing of the shorter tandem repeats to the extent that eventually the human genetic linkage map was essentially based upon microsatellite typing (Weissenbach et al. 1992). The second immediate application of highly polymorphic markers was individual identification; and tandem repeats polymorphism is still the basis of current forensic methods for DNA-based identification in humans. The assay is strictly speaking a multiple locus VNTR analysis, but the MLVA acronym was coined years later in the field of microbiological molecular epidemiology and forensics.

4.3
MLVA Set-up and Enrichment

Tandem repeats were also identified in prokaryotes during the 1980s and the polymorphism associated with a few specific genes, investigated for other reasons, was described. Multiple locus tandem repeats variability was shown to be promising for bacteria typing by Southern blotting and hybridisation with a GC-rich tandem repeat probe (Ross et al. 1992) or even an oligonucleotide probe (Marshall et al. 1996) as previously done in human genomics (Vergnaud 1989). It is the availability of large-scale sequence data which opened the way to PCR-based MLVA assays. The method was applied initially to *Haemophilus influenzae* (van Belkum et al. 1997) with an assay comprising five tetranucleotide microsatellites. However, all bacterial species are not equally amenable to MLVA typing and the first step in setting-up a MLVA assay is to evaluate the potential of MLVA typing for the species of interest.

4.3.1 Evaluation of the Potential Interest of MLVA for a Given Species

The identification of tandem repeats from sequence data is easily achieved owing to the availability of genome sequence data and software for sequence analysis (Benson 1999) and even unfinished, low-coverage genome sequence data can be used. Taking advantage of these resources, Le Flèche et al. (2001, 2002) and Denoeud and Vergnaud (2004) have developed and made available a tandem repeat database as part of a first "MLVA web service". The initial release in year 2001 contained 36 bacterial genomes, compared to the more than 200 genomes available in the latest update. In addition, the database also includes genome comparison results when two or more strains from the same (or sufficiently genetically close) species have been sequenced: the tandem repeats with a different size in the two strains are automatically identified (Denoeud and Vergnaud 2004). This greatly facilitates the identification of candidate polymorphic loci, as shown for instance by Ramisse et al. (2004). In order to avoid the duplication of work by independent groups and to limit the giving of different names to the same locus (as recalled by, for instance, Le Flèche et al. 2002), the database includes links to tandem repeats which have already been investigated and given names in the literature. These resources are accessible over the internet (http://minisatellites.u-psud.fr) and can also be set-up locally.



		Strain							
VNTR		A	B	C	D	E	F	G	H
45bp	1	2	4	1	2	1	1	4	2
24bp	2	8	6	3	7	3	6	4	7
48bp	3	3	5	9	8,5	9	9	5	7
	4								
	5								
	6								
	7								
	8								

Fig. 4.1. Schematic representation of a MLVA scheme. Primers are chosen on both sides of VNTR loci and PCR products are electrophoresed (here on agarose gel) together with size markers. The amplicon size is converted into a repeat number. Multiple markers are analysed in the same way, a distance matrix is generated and a clustering analysis is produced

Candidate tandem repeats can then be tested on a few diverse strains. Less than ten strains will usually be sufficient and polymorphism can be easily evaluated on agarose gels, so that tens of loci can be quickly tested at low cost in a couple of weeks (Fig. 4.1).

4.3.2 MLVA Validation

After this quick screening has been achieved, it is necessary to precisely identify the need and to define and collect an appropriate reference strain collection. Ideally, the reference collection to be used should have been already characterised and typed using the currently recognised typing methods, so that MLVA can immediately be compared in terms of typeability, reproducibility, relevance and discriminatory efficiency. In particular, different distance coefficients and clustering methods can be evaluated and the dendrograms obtained can be compared with the known epidemiological relations between the strains. Often a few tens of relevant strains will be sufficient for this phase of setting-up an assay. Then the strength and validity of the assay increases as many more strains are genotyped and similarity coefficients and clustering methods are fully tested and validated. Once strains have been selected, the PCR-amplification of tandem repeat loci using primers flanking the array and the measuring of the PCR product length are relatively standard (summarised in Fig. 4.1). Any equipment able to measure a DNA fragment length with sufficient resolution depending on the repeat unit size can be used. Maximum resolution means that the method used must be able to confidently resolve PCR products differing by one repeat unit. Sophisticated equipments such as DNA sequencing machines are able to do this; and such machines may even be necessary for typing arrays with short or very short repeat units, or relatively long alleles. The majority of current needs can be satisfied by methods with a lower resolution, in particular agarose gels and ethidium bromide staining. The MLVA assay can then be run with very ordinary equipment and at very low cost in terms of consumables and equipment. A typical agarose gel MLVA typing set-up consists of a control strain and size marker, each loaded a number of times on each gel (usually 4–7 times, depending upon gel size) in order to be able to take into account and compensate for both intra- and inter-gel electrophoresis variations, as described for instance by Pourcel et al. (2004). In any case, and whatever method is used, the resulting data can be compared and merged only if the appropriate quality control procedures, common reference strains and identical allele assignment conventions are used. Figure 4.2 illustrates a typical MLVA set-up based on agarose gel electrophoresis.

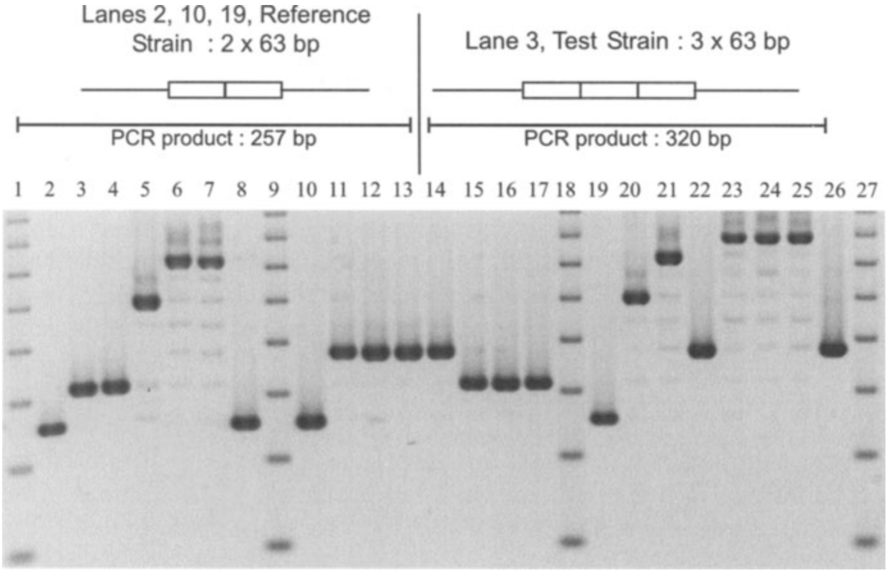


Fig. 4.2. MLVA typing on agarose gel. A tandem repeat from *Brucella* was amplified across 20 strains. Seven lanes are dedicated to controls, i. e. a size marker (four lanes: 1, 9, 18, 27) and a reference strain (three lanes: 2, 10, 19). The size marker used here is a 100-bp ladder (bands from 100 bp to 800 bp are shown). The repeat unit is 63 bp long. Such gels can be easily read manually: six different alleles are observed, comprising 2–8 repeat units

4.3.3
Data Management

The end-product of the assay is typing data, expressed in repeat copy number. These very simple files can be easily merged to produce integrated databases from different sources. When running small-scale projects, limited to a few tens of strains and/or when the biological characterisation of strains is limited to MLVA typing, there is little need for a real database management system. Careful double-checked manual reading and typing into a text file are appropriate (Fig. 4.2). However, when running larger projects and when different kinds of data must be stored and eventually merged for analysis, dedicated data management software is needed. The most widely used such software is Bionumerics (Applied-Maths) which acts as a warehouse for the storage of any biological data and also contains a collection of powerful tools for data analysis.

4.4 Existing First-generation MLVA Assays

MLVA is still quite new. So far, no official standard has been defined for any of the bacteria which will be presented below or are listed in Table 4.1. One reason for this is that MLVA assays are still in the development phase, in terms of the number of markers and strains tested, but it is very likely that, in the coming years, such standards will emerge, at least for the most actively investigated bacteria. Another reason is that the resolution of a MLVA assay can be increased by adding markers (Fig. 4.3), but requirements in terms of resolution depend upon the epidemiological question being asked. The investigation of local outbreaks for instance will benefit from the use of tandem repeats with a high mutation rate, in addition to a routine MLVA assay for strain typing. In other words, the single term MLVA assay will often cover probably two or three complementary panels of markers. In some cases, the use of a few markers will be quite sufficient to cover the need. Table 4.1 lists the bacteria for which MLVA assays have been published so far. In many instances, only one study has been reported, often including only a few markers and a limited number of strains. In other cases, much more work has already been done. Interestingly, a significant fraction of the more thorough investigations is related to pathogens which represent potential biological warfare agents. In this area of technological development, as in others before, it may be so that defence-driven projects related to microbial forensics will contribute to and speed up the development of epidemiological tools for many other pathogens which represent significant human health issues.

▶ **Fig. 4.3.** Comparison of the discrimination power of a MLVA analysis with 5, 13 or 19 VNTR. A collection of 50 strains from the *M. tuberculosis* complex (MTBC) were typed using 19 markers; and phylogenetic trees were produced using the data from either five markers (ETRs, *left panel*), 13 markers (5 ETRs + 8 MIRUs, *middle panel*) or 19 markers (5 ETRs + 8 MIRUs + 6 Mtubs, *right panel*). The strains were independently assigned to a MTBC group by classic biochemical assays and microdeletion typing (codes: *CAN* "*M. canettii*" strains, *EAI* East Africa/India, *BOV-AFRI M. bovis* and some *M. africanum* strains, *AFRI1* the rest of *M. africanum* type 1 strains, *BEIJ* Beijing strains, *MOD-CDC* the group of modern *M. tuberculosis* strains, including the reference CDC1551 strain, *MOD-H37* the group of modern *M. tuberculosis* strains, including the reference H37Rv strain).When all 19 markers are used in the analysis, 50 genotypes are identified (numbered from 1 to 50). The clustering fits with the independent classification. When 13 markers are used, the discrimination is slightly reduced (46 different genotypes identified). The clustering achieved is still reasonable, with a few inconsistencies: genotype 16 (EAI strain) is grouped with "*M. canettii*" strains and three genotypes from modern *M. tuberculosis* strains are incorrectly assigned (genotypes 44, 49, 40) to the Beijing group of strains. When only five markers are used (*left panel*), 36 different genotypes are identified, which is still relatively high, but the clustering achieved is of little value

We will discuss in more details the application of MLVA for epidemiology- or phylogeny-related investigations of five representative species: *Mycobacterium tuberculosis, Bacillus anthracis, Yersinia pestis, Brucella* sp, and *Legionella pneumophila*, for which enough data exist to assess the validity of the technique, or which illustrate specific points of interest.

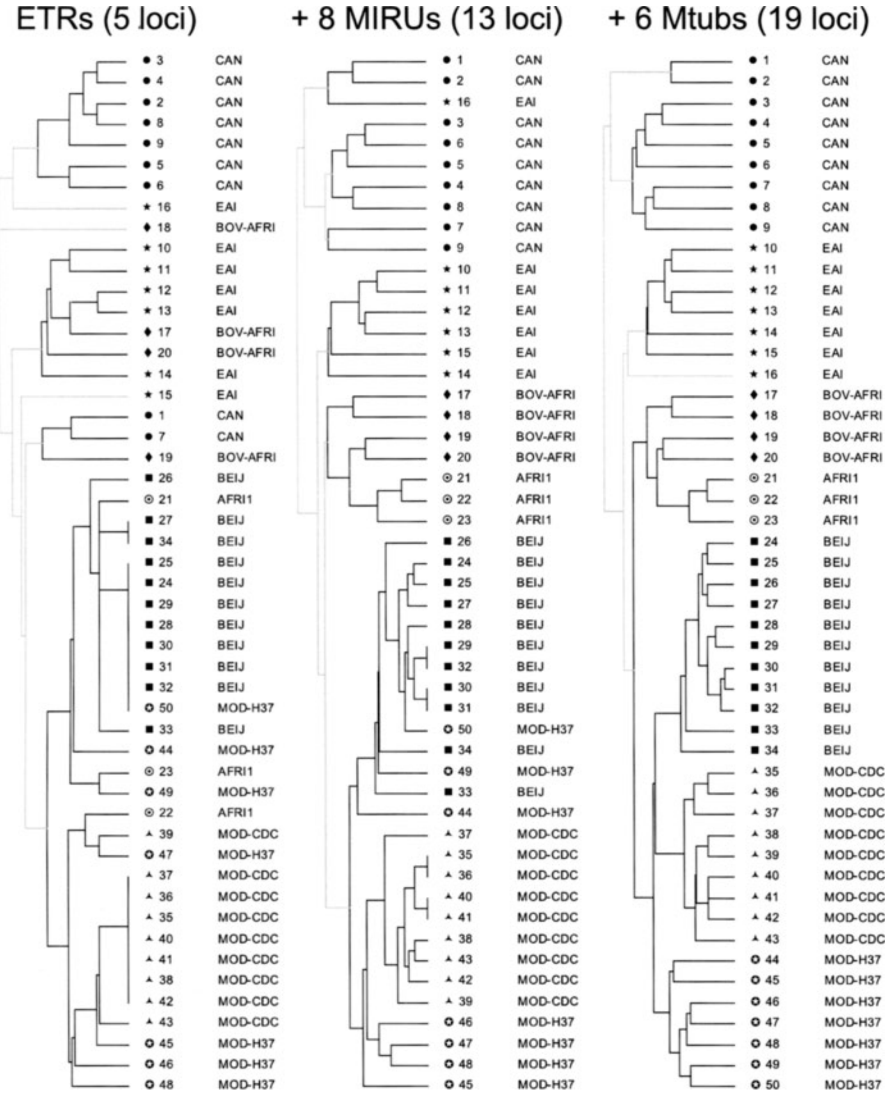


Table 4.1. List of MLVA development reports. Methods: *seq* gel-based sequencing machine, *capi* capillary electrophoresis, *aga* agarose gel

Bacteria[a]	VNTR loci[b]	Repeat (bp)[c]	Isolates	Method	Reference
Bacillus anthracis	8	2–36	426	seql	Keim et al. (2000)
	24 (18)	9–78	32	agarose	Le Flèche et al. (2001)
Bordetella pertussis	6	5–15	198	seq	Schouls et al. (2004)
Borrelia sp.	10	2–21	41	seq	Farlow et al. (2001)
	8	8	22	capi	Bricker et al. (2003)
Candida albicans[a]	3	4	100	seq	Botterel et al. (2001)
Enterococcus faecalis	7	141-393	83	aga	Titze-de-Almeida et al. (2004)
Enterococcus faecium	6	121-279	392	aga	Top et al. (2004)
Escherichia coli O157	7	6–18	81	sequencing	Noller et al. (2003)
	7	6–30	73	capi	Lindstedt et al. (2003)
Francisella tularensis	6	2–21	56	seq	Farlow et al. (2001)
	25	2–23	192	seq	Johansson et al. (2004)
Hemophilus influenzae	5	3–6	20	aga	van Belkum et al. (1997)
Legionella pneumophila	6	18–125	78	aga	Pourcel et al. (2003)
Leptospira interrogans	7	34–77	51	aga	Majed et al. (2005)
Mycobacterium avium	6	53	73	aga	Bull et al. (2003)
	5	20–70	50	aga	Overduin et al. (2004)
M. leprae	5	2–3	12	sequencing	Truman et al. (2004)
	9	1-27	4	seq	Groathouse et al. (2004)
M. tuberculosis	7	15–79	25	aga	Frothingham and Meeker-O'Connell (1998)
	12 (10)	53	31	aga	Supply et al. (2000)
	6	69	100	aga	Skuce et al. (2002)
	21 (8)	9–58	90	aga	Le Flèche et al. (2002)
Pseudomonas aeruginosa	7	6–115	89	aga	Onteniente et al. (2003)

Table 4.1. (continued)

Bacteria[a]	VNTR loci[b]	Repeat (bp)[c]	Isolates	Method	Reference
Salmonella typhimurium/ typhi	8	6–189	102	capi	Lindstedt et al. (2003)
	5	7–26	61	aga	Liu et al. (2003)
	10 (7)	3–20	99	aga	Ramisse et al. (2004)
Staphylococcus aureus	7	48–159	16	aga	Hardy et al. (2004)
Xylella fastidiosa	7	7–9	27	aga	Coletta-Filho et al. (2001)
Yersinia pestis	25	9–60	3+180	aga	Le Flèche et al. (2001), Pourcel et al. (2004)
	42 (35)	1–45	24+156	seq	Klevytska et al. (2001), Achtman et al. (2004)

[a] with the exception of *C. albicans*

[b] number of loci proposed for MLVA (number of new loci)

[c] repeat unit size range explored in the report

4.4.1 *Mycobacterium tuberculosis*

This is the bacterium for which MLVA has been the most extensively used to date and for which a large body of data is available. VNTR markers have been described by different teams and used alone or in combination. Particularly interesting markers were the exact tandem repeats (ETRs; Frothingham and Meeker-O'Connell 1998), multiple interspersed repetitive units (MIRUs; Supply et al. 2000), QUBs (for Queen's University of Belfast; Skuce et al. 2002) and Mtubs (Le Flèche et al. 2002). ETRs are 53–79 bp long and only ETRA is located within an ORF. The allelic profiles are reproducible and stable and VNTR typing was proposed to be useful for strain differentiation and evolutionary studies. MIRUs are tandem duplications of 53 bp except for MIRU04, a 77 bp repeat, which in fact corresponds to ETRD. MIRU31 corresponds to ETRE. Most are present in regions separating genes. In contrast, QUBs are mostly located inside genes. ETRA, QUb11a and QUB11b are present in the same protein, pUCB, a protein of the PPE family (O'Brien et al. 2000). They show a very high level of polymorphism. Additional informative markers were described by Le Flèche et al. (2002), in particular Mtub21 and Mtub39, both localised in intergenic regions. The size of the repeats in these different VNTRs is such that agarose gel-based MLVA can be performed. However, automated procedures are commonly used (Supply et al. 2001; Spurgiesz et al. 2003) and their utility in clinical mycobacteriology analysis was recently demonstrated (Allix et al. 2004).

MLVA was compared to classic typing methods for *M. tuberculosis*: spoligotyping, which investigates the polymorphism of a single locus, the DR locus, and IS typing, usually performed by RFLP analysis (Sun et al. 2004). The most recent studies concluded that the resolution of MLVA compares favourably with the other techniques when a sufficient number of informative markers are used, i. e. more than the most frequently used set of 12 MIRUs. For instance, MLVA appears to be the best method to investigate the diversity inside the important "Beijing" family, a recently emerged group of strains. MLVA assay was also a key assay in describing the group of "*M. canettii*" as a single entity (Fabre et al. 2004). ETRA, a very informative marker for the complete *M.tuberculosis* complex, shows a single allele in "*M. canettii*," an allele which has been found only in two *M. tuberculosis* strains belonging to the more ancient family from East Africa/India (Pourcel, unpublished data). This family can be identified on the basis of a specific allele of MIRU24, an otherwise very poorly informative marker (Sun et al. 2004). However, although many reports suggest that MLVA may be the new gold standard technique for typing inside the *M. tuberculosis* complex, more needs to be done to define a common assay allowing comparison of data between laboratories. Some markers are commonly used (with sometimes different names), whereas others are only used by some laboratories. In addition, there is a wrong assumption that some markers, because they do not seem informative inside a subgroup, should not be used although they are clearly useful when a large population of strains is studied. In contrast, a marker such as QUB-11a, a highly polymorphic repetition, can be useful in epidemic situations because of rapid modifications but is probably not stable enough for phylogenetic studies. In recent reports, it was proposed that VNTR typing should be used in combination with IS6110RFLP, a rather cumbersome technique necessitating the preparation of high quality DNA. Instead, the addition of several VNTR markers, already described in the literature, bringing the total number to 19, should be sufficient for high resolution analysis. Figure 4.3 shows a clustering analysis performed on a collection of 50 strains of the *M. tuberculosis* complex using either five loci (the five ETRs), 13 loci (the five ETRs and eight MIRUs) or 19 loci (the previous markers plus six Mtubs). The strains were selected from our collection to contain representatives of the major *M. tuberculosis* families ("Modern", "Beijing", ancient East Africa/India), plus some *M. bovis*, *M. africanum*, and "*M. canettii*"strains; and a similar pattern was observed even when more strains were used. Interestingly, these major groups are well defined by biochemical assays or by the independent tools (micro-deletion typing) described by Marmiesse et al. (2004). Typing with five markers is clearly not robust, even if the discriminatory power is already very good (36 genotypes resolved). With 13 markers (the ten most relevant MIRUs, the ETRs), a much nicer clustering is achieved with still some inconsistencies: one East Africa/India strain is grouped with the *M. canettii* group and the Beijing and Modern clusters are poorly defined. Forty-six genotypes are resolved. The panel of 19 markers proposed by Fabre et al. (2004) correctly clusters the strains and 50 genotypes are resolved. If necessary, the typing assay could be extended to 25 easily typable VNTRs, by using some QUB markers and additional markers uncovered by sequence comparison between the *M. bovis* genome and *M. tuberculosis* (Le Flèche, unpublished data).

4.4.2 *Bacillus anthracis*

B. anthracis is a highly monomorphic species, recently emerged from the *B. cereus/thuringiensis* group through the acquisition of two virulence plasmids. The main reason for the development of a MLVA assay in this dangerous pathogen is microbial forensics. This bacterium is no longer a significant health problem but is a potential bioterrorist agent, as illustrated by the 2001 events. MLVA-based genotype databases have been the key tools to identify the precise strain which was used in the bioterrorist event. The first MLVA assay was built upon a number of contributions. An extensive search for DNA polymorphisms eventually led to the finding that tandem repeats were a major source of polymorphism in this organism. The assay comprised eight markers, two of which were located on the virulence plasmids (Keim et al. 2000); and 426 isolates were typed. A number of these isolates were collected during a single outbreak, or corresponded to reference strains conserved for a number of years in different laboratories. With very few exceptions, the genotypes were indeed identical, which demonstrated that most tandem repeats are sufficiently stable to define strains. Eighty-nine genotypes were resolved. Whereas some genotypes were restricted to geographic regions, others were found to be widely distributed.

Taking advantage of the availability of large-scale sequence data, the MLVA assay was later expanded by adding 18 new markers (Le Flèche et al. 2001; http://bacterial-genotyping.igmors.u-psud.fr/). All of these markers are located on the chromosome within ORFs. Some of the encoded proteins are components of the outer layers of the spore. Bams13, for instance, is a 9-bp repeat located within the *BclA* gene which shows a 500-bp size difference between the largest and the smallest alleles (Le Flèche et al. 2001). The collagen-like BclA protein is the main component of the *B. anthracis* exosporium; and the Bams13 length polymorphism is directly related to the exosporium size (Sylvestre et al. 2003). The typing of 32 isolates confirmed the existence of the two main clusters, A and B, identified by (Keim et al. 2000) and showed the existence of additional clearly distinct branches, represented by isolates from West Africa. Much work still needs to be done, using MLVA in combination with other DNA analysis methods, on both *B. anthracis* and *B. cereus* to, for instance, identify the geographic origin of *B. anthracis*. The currently proposed assay comprises 24 loci which can be typed by agarose gels or by capillary electrophoresis, has a much higher resolution than the earlier 8-markers assay and represents a good first-level typing assay for phylogenetic investigations. To investigate local outbreaks, microsatellites (i.e. tandem repeats with the shortest repeat units) might constitute a second-level MLVA set. Eventually, it can only be hoped that all new isolates identified in the world will be genotyped and the genotypes submitted to common databases, as illustrated by the prototype (http://bacterial-genotyping.u-psud.fr).

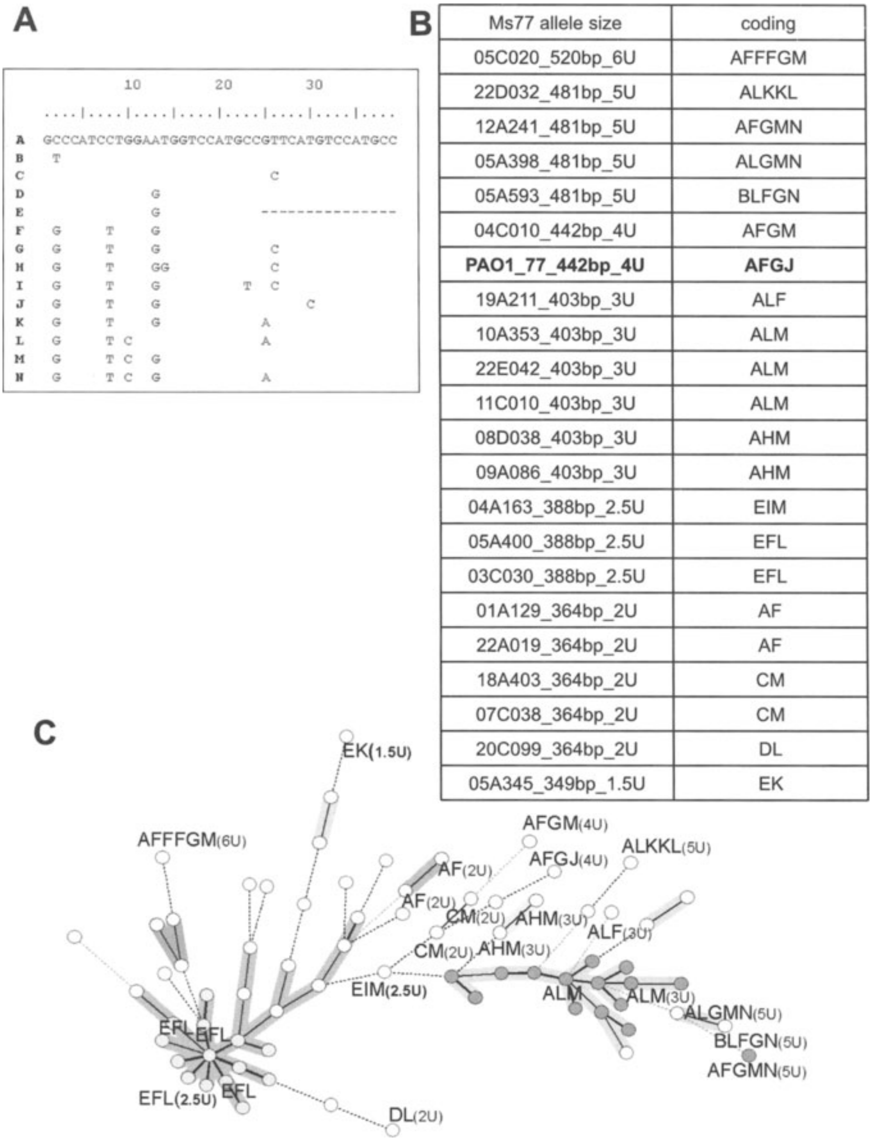
4.4.3
Yersinia pestis

The first report of the analysis of VNTR polymorphism at one locus in *Y. pestis* by Adair et al. (2000) suggested that these sequences could be a useful source of polymorphism in this very monomorphic species. Indeed, the works of Le Flèche et al. (2001) and Klevytska et al. (2001) confirmed that a MLVA scheme could be used to efficiently genotype *Y. pestis* strains. The choice of markers by the two teams was very complementary, in part due to the different electrophoresis techniques used and only seven markers were common to the two sets (Pourcel et al. 2004). Markers with small repetitions, of the microsatellite class, were favoured by the Keim laboratory, whereas in our laboratory we chose markers with repetitions larger than 12 bp, to allow for agarose gel separation of alleles. The selection of markers on the basis of the repetition size can have consequences on their discriminatory efficiency, as the mechanisms of variability of microsatellite and minisatellite (more than 9 bp long) can be different. More recently, significantly larger and more diverse collection of strains have been analysed using the two sets of markers (Achtman et al. 2004; Pourcel et al. 2004). Previous classifications were essentially based upon biochemical assays, which define the three classically recognised biovars: Antiqua, Medievalis and Orientalis. The MLVA clustering is usually in agreement with this rough classification but provides a much higher discrimination. The results obtained clearly distinguish between Antiqua strains from Asia and Antiqua strains from Africa. Interestingly, a few abnormalities were uncovered, with some Medievalis strains clustering among Antiquas. Further investigations demonstrated that the Medievalis phenotype resulted in these strains from different mutation events inactivating the *napA* gene. Although other very high resolution typing methods had been used for some years before, including IS typing by Southern blotting, these methods were unable to detect these inconsistencies, which illustrates the power of MLVA typing.

4.4.4 *Brucella sp.*

The single study published so far on MLVA typing in the *Brucella* genus illustrates some aspects of MLVA set-up and marker selection. Bricker et al. (2003) investigated the polymorphism associated with a family of octameric tandem repeats located within a mobile element present in multiple copies in the *Brucella* genome. The mobile element is small enough to use at least one primer located outside of it, so that each locus can be amplified and analysed independently from the others. The *Brucella* genus is separated in a number of species, not for genetic reasons (the genus is very highly homogeneous) but because of some features of the associated disease, a strong host specificity within mammals and varying virulence in human (Moreno et al. 2002). Each species can be further separated in a few biovars by biotyping, which is a combination of biochemical assays, phage typing, serotyping and growth in the presence of specific dyes. Biotyping data necessitates the manipulation of live bacteria, has a low resolution and is sometimes ambiguous, so that alternative, DNA-based assays would clearly be of interest. The 8-loci MLVA assay proposed by Bricker et al. (2003) is very discriminatory, highly reproducible and all strains investigated can be fully typed for the eight loci. However, the clustering of strains deduced from the MLVA typing data does not fit with the biotype or even with the species assignment. The best explanation for this behaviour and inconsistency is that the tandem repeat loci used have such a high mutation rate, within a limited allele size range, that many alleles have an identical size in spite of a different evolutionary origin (homoplasy). Because the repeat array is perfect, such alleles will be strictly identical and cannot be distinguished even by sequencing, as can be done when internal heterogeneity exists (see paragraph below; Fig. 4.4). Such an assay cannot replace the existing tools and is limited to the investigation of local outbreaks.

Hopefully, the existence of many additional tandem repeat sequences in the *Brucella* genome indicates that a MLVA assay will eventually be developed for this species (Le Flèche et al., in preparation).



A

```
            10        20        30
   ....|....|....|....|....|....|....|....|....
A  GCCCATCCTGGAATGGTCCATGCCGTTCATGTCCATGCC
B   T
C                           C
D               G
E               G           ---------------
F   G     T    G
G   G     T    G            C
H   G     T    GG           C
I   G     T    G         T  C
J   G     T    G                C
K   G     T    G           A
L   G     T C              A
M   G     T C  G
N   G     T C  G           A
```

B

Ms77 allele size	coding
05C020_520bp_6U	AFFFGM
22D032_481bp_5U	ALKKL
12A241_481bp_5U	AFGMN
05A398_481bp_5U	ALGMN
05A593_481bp_5U	BLFGN
04C010_442bp_4U	AFGM
PAO1_77_442bp_4U	**AFGJ**
19A211_403bp_3U	ALF
10A353_403bp_3U	ALM
22E042_403bp_3U	ALM
11C010_403bp_3U	ALM
08D038_403bp_3U	AHM
09A086_403bp_3U	AHM
04A163_388bp_2.5U	EIM
05A400_388bp_2.5U	EFL
03C030_388bp_2.5U	EFL
01A129_364bp_2U	AF
22A019_364bp_2U	AF
18A403_364bp_2U	CM
07C038_364bp_2U	CM
20C099_364bp_2U	DL
05A345_349bp_1.5U	EK

Fig. 4.4. Analysis of the internal variation inside the *P. aeruginosa* ms77 marker. **a** Sequence of the different motifs and letter code. **b** Coding of the ms77 allele in 22 different strains, showing the internal variability. **c** Phylogenetic relationship between strains using the minimum spanning tree representation and corresponding ms77 alleles (Onteniente et al. 2003; Onteniente 2004)

4.4.5 *Legionella pneumophila*

VNTR analysis in *L. pneumophila* led to the description of three informative markers that could be used for strain comparison in epidemic situations and the description of some additional markers that were amplified only in a subset of strains (Pourcel et al. 2003). Indeed, comparison of the genome sequence of the strains Philadelphia, Paris and Lens revealed that 13% of the DNA was strain-specific (Cazalet et al. 2004; Chien et al. 2004). By comparison of repeated sequences in the three sequenced genomes, new VNTRs were selected and primers chosen to match all the strains (http://minisatellites.u-psud.fr/comparison/; Pourcel, in preparation). Thus for species such as *L. pneumophila* which show a very high intraspecies variability, the availability of several sequenced genomes is necessary to set a MLVA assay. However, the number of available markers for a *L. pneumophila* MLVA assay remains limited. Sequencing of a large collection of alleles for three markers shows an important internal variability, resulting in homoplasy (Pourcel et al. 2003; Pourcel, unpublished data). Sequence data add to the resolution of the assay and open the way to an analysis of the mechanism of evolution of repeated sequences.

4.4.6 Other Bacteria

For a number of bacteria such as *Mycobacterium avium* subsp. *paratuberculosis*, *Pseudomonas aeruginosa*, *Salmonella enterica* subsp. *enterica* (including *typhi* and *typhimurium*), *Staphylococcus aureus*, etc (see Table 4.1) VNTR markers were identified and tested on sometimes relatively small collections of strains; and much still needs to be done before MLVA can become a standard procedure. However, in many instances and although only a limited number of markers were used, the authors believe that the resolution of the assay was comparable to that of other more complex assays. The major problem is the frequent use of microsatellites (2-bp to 8-bp repeat units) which tend to be unstable, as reported in several studies, with especially high homoplasy levels. In addition, they necessitate the use of sequencing gels or methods with equivalent resolution.

4.5 Validating and Analysing MLVA Data

A number of aspects specific for tandem repeat analysis must be kept in mind. Firstly, tandem repeat loci can be very variable in terms of mutation rates, some loci having an extreme mutation rate while others are monomorphic. At present, this behaviour cannot be predicted from the sequence itself and will have to be experimentally measured by eventually typing hundreds of strains, as was done previously for human forensics-related projects. Regarding human forensics and paternity analyses for instance, hundreds of individuals of different ethnic origins have been genotyped in order to estimate both reliable allele frequencies and mutation rates for each locus employed in an assay. A number of different processes have been shown to drive mutation events in tandem repeats, including replication slippage and double-strand break repair (Debrauwère et al. 1999; Vergnaud and Denoeud 2000). Highly polymorphic markers which often result from a higher rate of mutation events will usually have a high homoplasy level. Such markers are sometimes called "highly informative", which is not necessarily correct. On the contrary, a MLVA assay based solely on such markers would probably be unable to cluster strains according to their true historical proximity, as illustrated previously with *Brucella.* Diversity indexes such as Simpson's index, promoted by Hunter and Gaston (1988), although very useful for comparing the discriminatory power of assays, do not measure the relevance of the discrimination which is achieved by a given marker, or combination of markers. Eventually, it will probably make sense to consider that two strains which differ at one highly variable marker are more similar than two strains differing at a moderately variable marker. Such more sophisticated distance coefficients cannot be developed until many strains have been typed, so that it should not come as a surprise to get a feeling that MLVA typing is at least in some instances not yet mature. MLVA will clearly take a major place among the epidemiological tools available to type a number of major bacterial pathogens, but this field of investigation is at present a very quickly evolving and competitive area of research and development.

Because many strains from different countries will have to be typed, it is essential that the MLVA data be carefully validated with appropriate controls, so that data sets from different laboratories can be merged. The main reason for this is that PCR products containing tandem repeats may occasionally show an electrophoretic behaviour which can give incorrect size measurements. Abnormalities are not random, but are locus-dependant, i.e. only a few loci may contain a repeat unit sequence bias or length which may modify the migration behaviour. The effect will be different if PCR product sizes are run as double-strand DNA (as for instance on an agarose gel) or single-strand DNA (as usually done on sequencing machines). The effect may also be proportional to the number of repeat units. For instance, marker Bams01 from *Bacillus anthracis* (Le Flèche et al. 2001) comprises a 21-bp very highly purine-rich repeat unit and runs significantly more slowly than expected from sequencing data. This suggests that the repeat unit is slightly kinked and that this is amplified for larger alleles because a size of 21 bp represents exactly two DNA helix turns.

This difficulty is easily circumvented, by using a few reference DNAs with well characterised repeat copy numbers. When such sets are not easily sharable, or have not been defined, they can be organised locally once and for all by sequencing alleles from a few representative strains.

MLVA data can then be held and exchanged in simple text files. Each strain is described by a succession of values corresponding either to numbers of repeat units or to allele sizes expressed in base pairs. Although the former format is usually preferred, the latter can sometimes not be avoided for some rare tandem repeats in which combinations of repeat unit lengths coexist. This is observed for instance in *Legionella pneumophila* ms4 (Pourcel, unpublished data). The allele size will then depend on the set of primers used, so that the data should be carefully corrected if different primer sets are employed by different research groups.

Expressing the data in terms of repeat copy numbers is for this reason usually more appropriate. However, even in this case, rules have to be defined, because tandem repeat arrays often do not contain a perfect copy number. The true copy number can be for instance 2.5, 3.5, 4.5, etc., which for simplicity will be coded as 2, 3, 4, or alternatively 3, 4, 5. This is illustrated by ETRD (alias MIRU04), for instance, from *Mycobacterium tuberculosis*. Since different conventions were used initially, published data must be converted before data from different groups can be merged. In publications, the conventions used need to be clearly described. It is convenient to refer to a sequenced genome, especially when the corresponding strain is widely available, which is usually the case. Then the comprehensive description of a tandem repeat can be summarised for instance by Bams30_9bp_727bp_57U in which Bams30 is the locus name (Le Flèche et al. 2001), 9 bp is the repeat unit length, 727 bp is the PCR product size expected in the reference strain using the primers referred to in the given report and 57U is the corresponding repeat unit number.

The resulting data matrix can be imported into data-mining tools or into more conventional biology-oriented clustering methods. The currently preferred method to measure similarities between two strains is the simple counting of the number of markers at which the two strains differ (divided by the total number of markers and expressed as a percentage). This is a very crude similarity measure which gives the same weight to all markers. It also considers that alleles which differ by one repeat unit are not evolutionarily closer than alleles which differ by many repeat units. The two assumptions are often wrong but, in spite of this, the resulting clustering analyses make sense (Fig. 4.3, right). This is because the use of multiple markers compensates for variable homoplasy levels at individual markers.

Figure 4.4 shows the sequence variability of *Pseudomonas aeruginosa* marker Ms77 (Fig. 4.4a) and the encoding of the different alleles (Fig. 4.4b). The clustering analysis shown in Fig. 4.4c uses the minimum spanning tree method and clearly demonstrates that strains possessing alleles with the same number of repeats but with different codes are correctly clustered when several VNTRs are typed (Onteniente 2004). This also shows that additional information can be obtained by sequencing alleles in species with important variability. As larger MLVA data sets will be available, containing hundreds of genotypes, it is likely that different similarity measures will be developed to take more precisely into account, first, the evolutionary rate and homoplasy level (which can be indirectly deduced in part from the HGDI values) and, second, the mode of evolution of each individual marker.

Once MLVA data has been produced and collected, it is easy to set-up shared internet resources, for instance MLVA web services, as exemplified at http://bacterial-genotyping.igmors.u-psud.fr/ (Le Flèche et al. 2002) and http://www.mlva.umcutrecht.nl (Top et al. 2004).

4.6 MLVA Compared to Other Methods

MLVA does not provide a molecular clock. It is clear at least in some instances that the mutation of tandem repeats directly influences the phenotype of the corresponding strains, so that these mutations are not neutral and probably contribute to the adaptation of the species to its environment, in a reversible way. When MLVA is to be used for evolutionary studies, other sequence-based methods with a lower resolution will usually be employed in combination, as illustrated by Achtman et al. (2004) and Fabre et al. (2004). MLVA applies to sub-species typing. In many pathogens of interest, tandem repeat polymorphism analysis, including MLVA, will complement the existing tools. In some instances, it will even become the gold standard, which does not mean that it will replace existing methods, each one often providing a different and complementary point of view. The *M. tuberculosis*, *B. anthracis*, *Y. pestis* studies and a few others are clearly among these. One key feature of MLVA typing is that its low cost opens the possibility of an almost systematic typing, not limited to the few hundred of strains (at best) included in research projects. In addition to the importance of this aspect for clinical epidemiology, the possibility to quickly check the identity of a strain is also very important for the maintenance of strain collections, in particular when dangerous pathogens or precious strains are involved.

Acknowledgements.

Work on the typing of bacterial pathogens in our laboratory is supported by University Paris XI and by the French Délégation Générale pour l'Armement.

References

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci USA 101:17837–17842

Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P (2000) Diversity in a variable-number tandem repeat from *Yersinia pestis*. J Clin Microbiol 38:1516–1519

Allix C, Supply P, Fauville-Dufaux M (2004) Utility of fast mycobacterial interspersed repetitive unit-variable number tandem repeat genotyping in clinical mycobacteriological analysis. Clin Infect Dis 39:783–789

van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H (1997) Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. Infect Immun 65: 5017–5027

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Botterel F, Desterke C, Costa C, Bretagne S (2001) Analysis of microsatellite markers of *Candida albicans* used for rapid typing. J Clin Microbiol 39:4076–4081

Bricker BJ, Ewalt DR, Halling SM (2003) *Brucella* 'hoof-prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs). BMC Microbiol 3:15

Bull TJ, Sidi-Boumedine K, McMinn EJ, Stevenson K, Pickup R, Hermon-Taylor J (2003) Mycobacterial interspersed repetitive units (MIRU) differentiate *Mycobacterium avium* subspecies *paratuberculosis* from other species of the *Mycobacterium avium* complex. Mol Cell Probes 17:157–164

Cazalet C, Rusniok C, Bruggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C (2004) Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet 36:1165–1173

Chien M, Morozova I, Shi S, Sheng H, Chen J, Gomez SM, Asamani G, Hill K, Nuara J, Feder M, Rineer J, Greenberg JJ, Steshenko V, Park SH, Zhao B, Teplitskaya E, Edwards JR, Pampou S, Georghiou A, Chou IC, Iannuccilli W, Ulz ME, Kim DH, Geringer-Sameth A, Goldsberry C, Morozov P, Fischer SG, Segal G, Qu X, Rzhetsky A, Zhang P, Cayanis E, De Jong PJ, Ju J, Kalachikov S, Shuman HA, Russo JJ (2004) The genomic sequence of the accidental pathogen *Legionella pneumophila*. Science 305:1966–1968

Coletta-Filho HD, Takita MA, de Souza AA, Aguilar-Vildoso CI, Machado MA (2001) Differentiation of strains of *Xylella fastidiosa* by a variable number of tandem repeat analysis. Appl Environ Microbiol 67:4091–4095

Debrauwère H, Buard J, Tessier J, Aubert D, Vergnaud G, Nicolas A (1999) Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. Nat Genet 23:367–371

Denoeud F, Vergnaud G (2004) Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource. BMC Bioinform 5:4

Fabre M, Koeck JL, Le Fleche P, Simon F, Herve V, Vergnaud G, Pourcel C (2004) High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*". J Clin Microbiol 42:3248–3255

Farlow J, Smith KL, Wong J, Abrams M, Lytle M, Keim P (2001) *Francisella tularensis* strain typing using multiple-locus, variable-number tandem repeat analysis. J Clin Microbiol 39:3186–3192

Frothingham R, Meeker-O'Connell WA (1998) Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 144:1189–1196

Groathouse NA, Rivoire B, Kim H, Lee H, Cho SN, Brennan PJ, Vissa VD (2004) Multiple polymorphic loci for molecular typing of strains of *Mycobacterium leprae*. J Clin Microbiol 42:1666–1672

Hardy KJ, Ussery DW, Oppenheim BA, Hawkey PM (2004) Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation. Microbiology 150:4045–4052

Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 26:2465–2466

Johansson A, Farlow J, Larsson P, Dukerich M, Chambers E, Bystrom M, Fox J, Chu M, Forsman M, Sjostedt A, Keim P (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. J Bacteriol 186:5808–5818

Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J Bacteriol 182:2928–2936

Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J, Keim P (2001) Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. J Clin Microbiol 39:3179–3185

Le Flèche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. BMC Microbiol 1:2

Le Flèche P, Fabre M, Denoeud F, Koeck JL, Vergnaud G (2002) High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. BMC Microbiol 2:37

Lindstedt BA, Heir E, Gjernes E, Kapperud G (2003) DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar *typhimurium* with emphasis on phage type DT104 based on variable number of tandem repeat loci. J Clin Microbiol 41:1469–1479

Liu Y, Lee MA, Ooi EE, Mavis Y, Tan AL, Quek HH (2003) Molecular typing of *Salmonella enterica* serovar *typhi* isolates from various countries in Asia by a multiplex PCR assay on variable-number tandem repeats. J Clin Microbiol 41:4388–4394

Majed Z, Bellenger E, Postic D, Pourcel C, Baranton G, Picardeau M (2005) Characterization of *Leptospira interrogans* sensu stricto serovars by VNTR polymorphism analysis. J Clin Microbiol 43:539–545

Marmiesse M, Brodin P, Buchrieser C, Gutierrez C, Simoes N, Vincent V, Glaser P, Cole ST, Brosch R (2004) Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. Microbiology 150:483–496

Marshall DG, Coleman DC, Sullivan DJ, Xia H, O'Morain CA, Smyth CJ (1996) Genomic DNA fingerprinting of clinical isolates of *Helicobacter pylori* using short oligonucleotide probes containing repetitive sequences. J Appl Bacteriol 81:509–517

Moreno E, Cloeckaert A, Moriyon I (2002) *Brucella* evolution and taxonomy. Vet Microbiol 90:209–227

Noller AC, McEllistrem MC, Pacheco AG, Boxrud DJ, Harrison LH (2003) Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates. J Clin Microbiol 41:5389–5397

O'Brien R, Danilowicz BS, Bailey L, Flynn O, Costello E, O'Grady D, Rogers M (2000) Characterization of the *Mycobacterium bovis* restriction fragment length polymorphism DNA probe pUCD and performance comparison with standard methods. J Clin Microbiol 38:3362–3369

Onteniente L (2004) Etude du polymorphisme associé aux répétitions en tandem pour le typage de bactéries pathogènes: *Pseudomonas aeruginosa* et *Staphylococcus aureus*. PhD thesis, University of Evry, Val d'Essonne

Onteniente L, Brisse S, Tassios PT, Vergnaud G (2003) Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing. J Clin Microbiol 41:4991–4997

Overduin P, Schouls L, Roholl P, van den Zanden A, Mahmmod N, Herrewegh A, van Soolingen D (2004) Use of multilocus variable-number tandem-repeat analysis for typing *Mycobacterium avium* subsp. *paratuberculosis*. J Clin Microbiol 42:5022–5028

Pourcel C, Vidgop Y, Ramisse F, Vergnaud G, Tram C (2003) Characterization of a tandem repeat polymorphism in *Legionella pneumophila* and its use for genotyping. J Clin Microbiol 41:1819–1826

Pourcel C, Andre-Mazeaud F, Neubauer H, Ramisse F, Vergnaud G (2004) Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. BMC Microbiol 4:22

Ramisse V, Houssu P, Hernandez E, Denoeud F, Hilaire V, Lisanti O, Ramisse F, Cavallo JD, Vergnaud G (2004) Variable number of tandem repeats in *Salmonella enterica* subsp. *enterica* for typing purposes. J Clin Microbiol 42:5722–5730

Ross BC, Raios K, Jackson K, Dwyer B (1992) Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. J Clin Microbiol 30:942–946

Schouls LM, van der Heide HG, Vauterin L, Vauterin P, Mooi FR (2004) Multiple-locus variable-number tandem repeat analysis of Dutch *Bordetella pertussis* strains reveals rapid genetic changes with clonal expansion during the late 1990s. J Bacteriol 186:5496–5505

Skuce RA, McCorry TP, McCarroll JF, Roring SM, Scott AN, Brittain D, Hughes SL, Hewinson RG, Neill SD (2002) Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets. Microbiology 148:519–528

Spurgiesz RS, Quitugua TN, Smith KL, Schupp J, Palmer EG, Cox RA, Keim P (2003) Molecular typing of *Mycobacterium tuberculosis* by using nine novel variable-number tandem repeats across the Beijing family and low-copy-number IS6110 isolates. J Clin Microbiol 41:4224–4230

Sun YJ, Lee AS, Ng ST, Ravindran S, Kremer K, Bellamy R, Wong SY, van Soolingen D, Supply P, Paton NI (2004) Characterization of ancestral *Mycobacterium tuberculosis* by multiple genetic markers and proposal of genotyping strategy. J Clin Microbiol 42:5058–5064

Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C (2000) Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol Microbiol 36:762–771

Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C (2001) Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J Clin Microbiol 39:3563–3571

Sylvestre P, Couture-Tosi E, Mock M (2003) Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. J Bacteriol 185:1555–1563

Titze-de-Almeida R, Willems RJ, Top J, Rodrigues IP, Ferreira RF II, Boelens H, Brandileone MC, Zanella RC, Felipe MS, van Belkum A (2004) Multilocus variable-number tandem-repeat polymorphism among Brazilian *Enterococcus faecalis* strains. J Clin Microbiol 42:4879–4881

Top J, Schouls LM, Bonten MJ, Willems RJ (2004) Multiple-locus variable-number tandem repeat analysis, a novel typing scheme to study the genetic relatedness and epidemiology of *Enterococcus faecium* isolates. J Clin Microbiol 42:4503–4511

Truman R, Fontes AB, De Miranda AB, Suffys P, Gillis T (2004) Genotypic variation and stability of four variable-number tandem repeats and their suitability for discriminating strains of *Mycobacterium leprae*. J Clin Microbiol 42:2558–2565

Vergnaud G (1989) Polymers of random short oligonucleotides detect polymorphic loci in the human genome. Nucleic Acids Res 17:7623–7630

Vergnaud G, Denoeud F (2000) Minisatellites: mutability and genome architecture. Genome Res 10:899–907

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M (1992) A second-generation linkage map of the human genome. Nature 359:794–801