Research article

A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*

Philippe Le Flèche[1,2], Yolande Hauck[2], Lucie Onteniente[2], Agnès Prieur[1,2], France Denoeud[2], Vincent Ramisse[1], Patricia Sylvestre[1], Gary Benson[3], Françoise Ramisse[1] and Gilles Vergnaud*[1,2]

Address: [1]Centre d'Etudes du Bouchet, BP3, 91710 Vert le Petit, France, [2]Génomes et Minisatellites, Institut de Génétique et Microbiologie, Bat 400, Université Paris XI, 91405 Orsay cedex, France and [3]Department of Biomathematical Sciences, Box 1023, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, USA

E-mail: Philippe Le Flèche - lefleche@igmors.u-psud.fr; Yolande Hauck - Yolande.Hauck@igmors.u-psud.fr; Lucie Onteniente - Lucie.Onteniente@igmors.u-psud.fr; Agnès Prieur - Agnes.Prieur@igmors.u-psud.fr; France Denoeud - France.Denoeud@igmors.u-psud.fr; Vincent Ramisse - Vincent.Ramisse@ceb.etca.fr; Patricia Sylvestre - psylvest@pasteur.fr; Gary Benson - benson@ecology.biomath.mssm.edu; Françoise Ramisse - f.ramisse@freesurf.fr; Gilles Vergnaud* - Gilles.Vergnaud@igmors.u-psud.fr
*Corresponding author

Published: 30 March 2001

BMC Microbiology 2001, **1**:2

Received: 19 February 2001
Accepted: 30 March 2001

This article is available from: http://www.biomedcentral.com/1471-2180/1/2

Abstract

Background: Some pathogenic bacteria are genetically very homogeneous, making strain discrimination difficult. In the last few years, tandem repeats have been increasingly recognized as markers of choice for genotyping a number of pathogens. The rapid evolution of these structures appears to contribute to the phenotypic flexibility of pathogens. The availability of whole-genome sequences has opened the way to the systematic evaluation of tandem repeats diversity and application to epidemiological studies.

Results: This report presents a database ([http://minisatellites.u-psud.fr]) of tandem repeats from publicly available bacterial genomes which facilitates the identification and selection of tandem repeats. We illustrate the use of this database by the characterization of minisatellites from two important human pathogens, *Yersinia pestis* and *Bacillus anthracis*. In order to avoid simple sequence contingency loci which may be of limited value as epidemiological markers, and to provide genotyping tools amenable to ordinary agarose gel electrophoresis, only tandem repeats with repeat units at least 9 bp long were evaluated. *Yersinia pestis* contains 64 such minisatellites in which the unit is repeated at least 7 times. An additional collection of 12 loci with at least 6 units, and a high internal conservation were also evaluated. Forty-nine are polymorphic among five *Yersinia* strains (twenty-five among three *Y. pestis* strains). *Bacillus anthracis* contains 30 comparable structures in which the unit is repeated at least 10 times. Half of these tandem repeats show polymorphism among the strains tested.

Conclusions: Analysis of the currently available bacterial genome sequences classifies *Bacillus anthracis* and *Yersinia pestis* as having an average (approximately 30 per Mb) density of tandem repeat arrays longer than 100 bp when compared to the other bacterial genomes analysed to date. In both cases, testing a fraction of these sequences for polymorphism was sufficient to quickly develop a set of more than fifteen informative markers, some of which show a very high degree of polymorphism. In one instance, the polymorphism information content index reaches 0.82 with allele length covering a wide size range (600-1950 bp), and nine alleles resolved in the small number of independent *Bacillus anthracis* strains typed here.

Background

The polymorphism associated with tandem repeats has been instrumental in mammalian genetics for the construction of genetic maps and still is the basis of DNA fingerprinting in forensic applications. Tandem repeats are usually classified among satellites (spanning megabases of DNA, associated with heterochromatin), minisatellites (repeat units in the range 6-100 bp, spanning hundreds of base-pairs) and microsatellites (repeat units in the range 1-5 bp, spanning a few tens of nucleotides).

More recently, a number of studies have supported the notion that tandem repeats reminiscent of mini and microsatellites are likely to be a highly significant source of very informative markers for the identification of pathogenic bacteria even when these pathogens are recently emerged, highly monomorphic species [1–5]. This probably reflects the important contribution of tandem repeats to the adaptation of the pathogen to its host. Tandem repeats appear to contribute to phenotypic variation in bacteria in at least two ways. Tandem repeats located within the regulatory region of a gene can constitute an on/off switch of gene expression at the transcriptional level [6,7]. Similarly, tandem repeats within coding regions with repeat units length not a multiple of three can induce a reversible premature end of translation when a mutation changes the number of repeats (reviewed in [8–10]). In other instances, the repeated unit length is a multiple of three, and the tandem repeat contributes to a coding region. In such cases, variations in the number of copies modify the gene product itself [11].

Mutation mechanisms of micro and minisatellites have been studied in some detail in eukaryotes, essentially human and yeast (reviewed in [12]). In brief, the data obtained so far suggest that microsatellites mutate by replication slippage processes; mutation rates depend upon the efficiency of mismatch repair mechanisms and an internal heterogeneity within the array strongly stabilizes the tandem repeat. In contrast, minisatellites mutate predominantly as the result of the repair of a double strand break initiated within, or very close to, the tandem repeat. In eukaryotes at least, these events can be of replicative origin [13], or can be genetically controlled, and specifically induced, during meiosis, at double strand breaks hot-spots. Minisatellite mutation rate in eukaryotes appears to be insensitive to mismatch repair efficiency, and internal heterogeneity is compatible with a high mutation rate [12, 14].

In bacteria, loci containing a tandem repeat from the microsatellite class (repeat unit sizes of 1-8 bp) have been called simple sequence contingency loci [8]. Altered number of repeats allows for reversible on and off states of expression for the corresponding gene. The mutation rate of a tetranucleotide (microsatellite) tract in *Haemophilus influenzae* is higher than 10^{-4} and contributes to the adaptation of the pathogen to its hosts as the infection progresses [15]. In such an extreme situation, the microsatellite is of limited value for strain identification, epidemiological and phylogenetic studies. The tandem repeat array is composed of perfect copies of the elementary unit, and different alleles are observed in a single culture. In contrast, the phylogenetic identity of minisatellite alleles of identical size can usually be further checked by DNA sequencing, since the repeated units are often not perfect [16]. The pattern of variants along the array provides an additional level of allele identification and phylogenetic information. In addition, tandem repeats with longer repeat unit length can be relatively easily typed in the size range of a few hundred base-pairs using ordinary horizontal gel electrophoresis.

In this report, we will first describe the use of a tandem repeats database for bacterial genomes ([http://minisatellites.u-psud.fr]) and briefly compare the general characteristics of tandem repeats in a number of bacterial genomes for which the sequence has been determined and made publicly available. We will then show how this tool can easily be applied to the rapid characterization of new highly polymorphic markers in two pathogens, *Y. pestis* and *B. anthracis*.

Both *Y. pestis* (causative agent of plague) and *B. anthracis* (causative agent of anthrax) are recently emerged clones of respectively Y. pseudotuberculosis [17] and *B. cereus* [18]. In the case of *Y. pestis*, a high resolution typing tool based on RFLP (Restriction Fragment Length Polymorphism) analysis of IS100 locations has already been developed [17]. However this technology is more demanding than PCR typing, which justifies the development of such an assay. In the case of *B. anthracis*, polymorphisms were initially identified essentially using AFLP (Amplified Fragment Length Polymorphism) typing [19]. Subsequent analyses demonstrated that the most informative fragments in AFLP patterns resulted from tandem repeat array length variations (five minisatellite loci were characterized in this way [2]).

Results and discussion

Use of the tandem repeats database

To date, 36 bacterial genome sequences from 32 species have been released in the public domain and are included in the database (Figure 1A; the nine archaebacteria genomes sequenced to date are presented in an other page, which can be accessed from [http://minisatellites.u-psud.fr/]). As many other sequencing projects are under way ([http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html] ; [http://www.tigr.org/tdb/mdb/

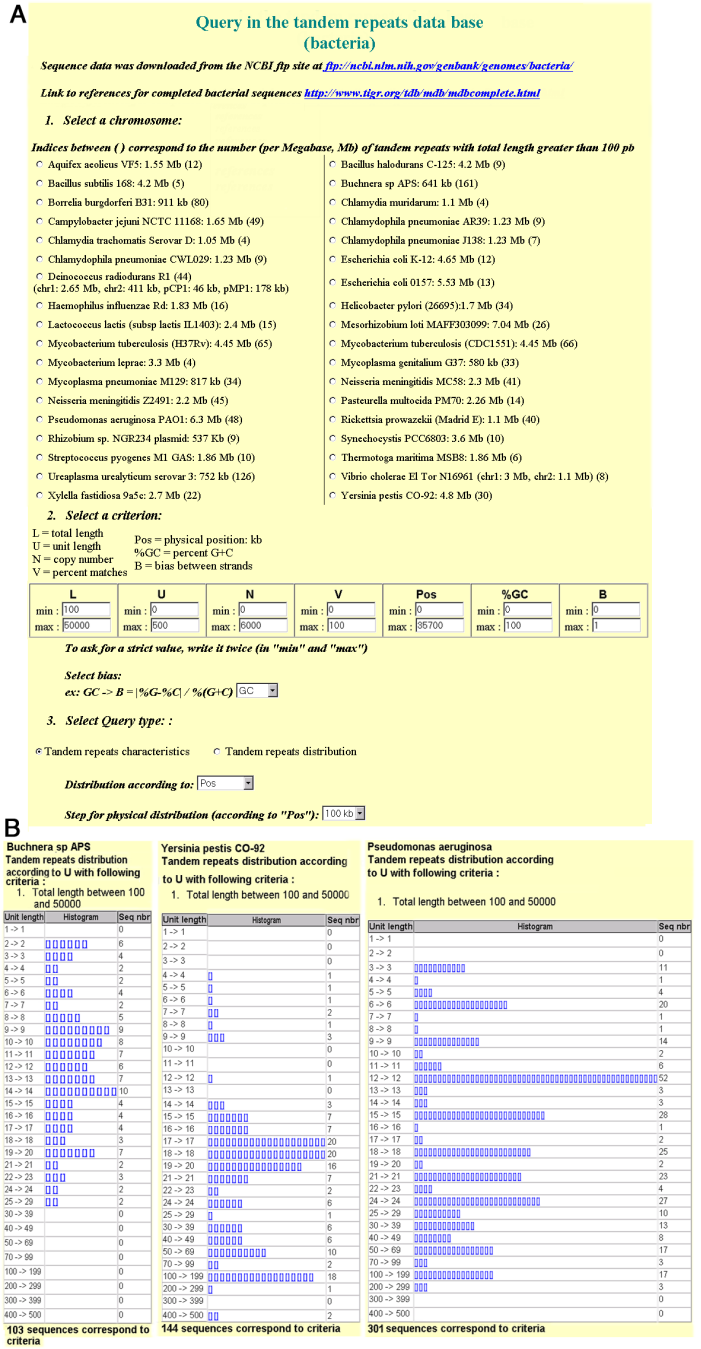

Figure 1

Querying the tandem repeats database 1A: bacterial tandem repeats main page Bacteria species are listed in alphabetical order. The name of the strain used for sequencing is indicated after the species name and before the genome size (expressed in megabase). The rightmost figure indicates the density (per Mb) of tandem repeat arrays longer than 100 bp. The search for tandem repeats can be restricted according to a combination of criteria, including total array length (L), repeat unit length (U), number of repeats (N), internal conservation of the repeats (V), position (expressed in kilobase) on the genome (Pos), GC content of the array (%GC), strand bias (B). Three different biases can be evaluated, GC bias, AT bias and Purine-Pyrimidine bias. The bias reflects strand asymmetry of the repeat sequence. The search output can either present a list of characteristics of the tandem repeats fulfilling criteria, ordered according to their position on the genome, or classify the tandem repeats according to a selected structural parameter. 1B: examples of queries in three genomes All tandem repeat arrays spanning more than 100 base-pairs are classified according to repeat unit length. The query was run on *Buchnera sp.* (left panel), *Yersinia pestis* (middle panel) and *Pseudomonas aeruginosa* (right panel).

mdbinprogress.html] ; [http://www.sanger.ac.uk/Projects/Microbes/]), the database will be regularly updated. The collection of tandem repeats present in a given genome can be queried according to a combination of criteria, total tandem repeat array length (L), repeat unit length (U), number of repeats (N), percentage of conservation of the repeats along the array (V), position on the genome (Pos), average GC percent of the repeats (%GC), strand bias in nucleotide composition (B) (these values have been precomputed using the Tandem Repeats Finder software described in [20]). The results shown on Figure 1B use the "Tandem Repeats Distribution according to repeat unit length" option (Figure 1A). Three genomes were searched for tandem repeat arrays longer than 100 base-pairs (L ≥ 100). The genomes selected illustrate three different behaviors. On the right panel, *Pseudomonas aeruginosa* shows a very striking bias towards minisatellites with a motif length multiple of three. On the left and middle panels of Figure 1B, B*uchnera sp* and Y. *pestis*, show no such bias. The overall density of tandem repeat arrays longer than 100 base-pairs varies in the different genomes. *Buchnera sp.* contains 103 such loci, for a total genome size of 641 kb, which corresponds to a density per megabase of 161. *Pseudomonas aeruginosa*, with a total genome length of 6.3 Mb, has a density of 48. *Y. pestis* has an intermediate value of 30. Figure 2 summarizes the values observed in the 32 species. Ten non pathogenic species are presented in the upper part, 22 pathogenic species on the lower part. The species are ordered from top to bottom according to increasing genome size. The dark bars indicate for each genome the density per megabase of tandem repeat arrays longer than 100 bp. The clear bars reflect the excess of tandem repeats with unit length a multiple of three. A wide range of situations is observed, with a remarkable excess of tandem repeats multiples of three in *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa*, presumably reflecting a significant contribution of tandem repeats to coding regions in these two bacteria.

As a quick illustration of the use of this database to facilitate the development of genotyping tools for bacterial genomes, we have evaluated the polymorphism associated with tandem repeats from *Y. pestis* on one hand and *B. anthracis* on the other (in this second instance, the genome sequence has not been completed yet and does not appear on the publicly accessible Tandem Repeats Database page, Figure 1A).

Application to Y. pestis

Figure 3A presents the result of a query run on *Y. pestis*, to identify tandem repeats with repeat units longer than 9 base-pairs repeated at least 7 times in the strain which has been sequenced (CO-92 biovar Orientalis).Sixty-four tandem repeats fulfill these criteria (an additional group

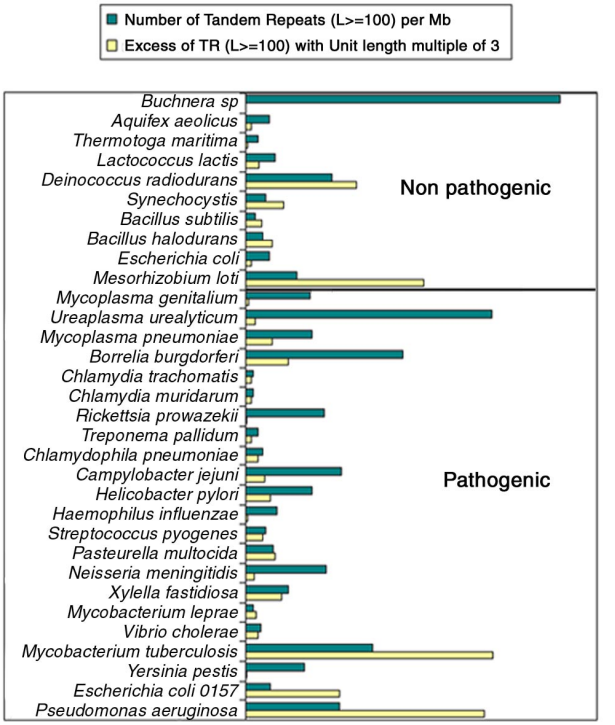

Figure 2
Relative frequency of tandem repeats within bacterial genomes The ten non-pathogen species are listed on top. Within each category, species are ordered according to genome size (smallest genome on top). The density of tandem repeat arrays longer than 100 bp is plotted for each species (dark bars). The clear bars reflect the excess (χ^2 values) of tandem repeats with a repeat unit length multiple of three.

of forty-nine have 6 copies of the motif; the twelve loci with the highest internal conservation were also included in this study). The output includes links to individual alignment files, as produced by the Tandem Repeat Finder software [20]. The alignment file also includes 200 base-pairs of flanking sequence from each side of the tandem repeat, from which primers can be selected for PCR amplification. Figure 3B shows an annotated extract of one alignment file. The positions of the primers selected for subsequent PCR amplification are underlined. Three *Y. pestis* (representing the Antiqua, Medievalis, and Orientalis biovars [17]) and two *Y. pseudotuberculosis* strains were used for the initial identification of minisatellites sufficiently polymorphic to be of interest for further studies. Table 1 summarizes the PCR conditions used for each polymorphic locus and the results obtained. A total of 76 tandem repeats were tested. PCR amplification failed in 6 cases. Twenty one loci are monomorphic in the five *Yersinia* strains typed here. Forty-nine of the loci are polymorphic (Table 1). Twenty-five of these are polymorphic among the *Y. pestis* strains.

A

Yersinia pestis CO-92
Tandem repeats distribution according to U with following criteria :

1. Unit length between 9 and 500
2. Copy number between 7 and 6000

Unit length	Histogram	Seq nbr
1 -> 1		0
2 -> 2		0
3 -> 3		0
4 -> 4		0
5 -> 5		0
6 -> 6		0
7 -> 7		0
8 -> 8		0
9 -> 9	[] [] [] []	4
10 -> 10	[]	1
11 -> 11	[]	1
12 -> 12	[] [] [] []	4
13 -> 13	[]	1
14 -> 14	[] [] [] []	4
15 -> 15	[] [] [] [] [] [] []	7
16 -> 16	[] [] [] [] [] []	6
17 -> 17	[] [] [] [] [] [] [] [] [] [] [] [] [] [] []	14
18 -> 18	[] [] [] [] [] [] [] [] [] [] []	11
19 -> 20	[] []	2
21 -> 21	[]	1
22 -> 23	[] []	2
24 -> 24	[] []	2
25 -> 29	[]	1
30 -> 39		0
40 -> 49	[]	1
50 -> 69	[] []	2
70 -> 99		0
100 -> 199		0
200 -> 299		0
300 -> 399		0
400 -> 500		0

64 sequences correspond to criteria

B

```
                *     **    * **
    3057961 CCACCGGATCCAACAGTG (observed unit: type a)
          1 CCGCCGGAAACAACGGCA (consensus)

                *     ** *  ****
    3057979 CCACCGGATCCGACATTG (type b)
          1 CCGCCGGAAACAACGGCA

                *  *  **     ***
    3057997 CCACCAGATCCAACGTTG (type c)
          1 CCGCCGGAAACAACGGCA

                *  *  **     **
    3058015 CCACCAGATCCAACGTTA (type d)
          1 CCGCCGGAAACAACGGCA

                         *
    3058033 CCGCCGGAAACCACGGCA (type e)
          1 CCGCCGGAAACAACGGCA

                         *
    3058051 CCGCCGGAAACCACGGCA (type e)
          1 CCGCCGGAAACAACGGCA

                   *
    3058069 CCGCCAGAAACAACGGCA (type f)
          1 CCGCCGGAAACAACGGCA

                   *  *
    3058087 CCGCCAGAGACAACGGCA (type g)
          1 CCGCCGGAAACAACGGCA

                         *
    3058105 CCGCCGGAAACCACGGCA (type h)
          1 CCGCCGGAAACAACGGCA


    3058123 CCGCCGGAAACAACGGCA (type i)
          1 CCGCCGGAAACAACGGCA


(type g-g-f-i-i-i-h-g-g-i-f-i-i-i-i-f-i-f)


    3058465 CCGCCGGAAACAACGGCA (type i)
          1 CCGCCGGAAACAACGGCA

                   *
    3058483 CCGCCAGAAACAACGGCA (type f)
          1 CCGCCGGAAACAACGGCA

                   *
    3058501 CCGCCAGAAACAACGGCA (type f)
          1 CCGCCGGAAACAACGGCA

                      * *
    3058519 CCGCCGGAACCGACGCGAACA (type j)
          1 CCGCCGGAAACAACG-G--CA

                        **      *
    3058540 CCGCCAGGGACGCAGACGCCA (type k)
          1 CCGCC--GGAAACA-ACGGCA


    3058561 CCGCC
          1 CCGCC


Consensus pattern (18 bp):
CCGCCGGAAACAACGGCA

Left flanking sequence: Indices 3057761 -- 3057960
TTGTTGGTAAAAGTAAGGTTATTATTCAGGGCGGTGAAGTTCACATTAACTCTGATGATTGTTGT
GGTCGTAAAACGGGCAATGGCGGCGGGGCATTCGGTGCCTTTGCGAGCCTGGCAGGTTTATTGCC
GCTGGCGGGTGCTATTGCCTTACCGCTACCGTTACCCTTGTTGCCAATAGTGGTGCCACCAGTGA
CATCG


Right flanking sequence: Indices 3058566 -- 3058765
TTGATTTTATAGATAAAAGGTAAGCATGTTCTGCGTGGGAAACACACGTAGAGCATGCGGTCTCT
GATTTTTCGAACGGTTAGAATTTAATTGAGGTGACGTATGTTCAGTGCGGCAAAAATGGGCGATG
TGACCAGCGATGGTGGTGCGATATGTACCGGTTCATGCGATGTGTTCATTAACAATATTCCGGCC
TCTAA
```

Figure 3
Selection procedure of minisatellites for *Y. pestis* 3A: Sixty-four tandem repeats have at least 7 units longer than 9 base-pairs. Panel A presents the distribution of these 64 loci according to repeat unit length. Each rectangle is an hyperlink to an alignment file. The rectangle indicated by the arrow is linked to the file illustrated in panel B. 3B: This is an annotated alignment file. The file corresponds to Yp3057ms09 (Table 1 and Figure 4; Yp : *Yersinia pestis*; 3057 : position on the genome, expressed in kilobases; MS09 : MiniSatellite index). The consensus pattern of 18 base-pairs is aligned to each motif. Annotations of the file are inserted within brackets. Although this minisatellite is very polymorphic, eleven different motifs (labeled a-k) are observed in the sequenced allele. The first four and last two copies are most diverged and rare. Four types of motifs (f, g, h, i) constitute most of the array. For convenience, 18 motifs have been removed from the alignment file and replaced by their letter code. The last two copies are 21 base-pair long instead of 18. The end of the alignment file (panel B, bottom) provides sequence data flanking the tandem repeat array. The positions of the primers chosen for PCR amplification of this locus (Table 1) are shown underlined.

Seven present a different allele in each of the five *Yersinia* strains, thirteen have a different allele in each of the three *Y. pestis* strains. Gel images for the 25 loci polymorphic among *Y. pestis* are shown in Figure 4. As can be seen, the repeat unit size and the overall length of the PCR products are such that tandem repeats differing by a single repeat unit can be distinguished by simple agarose gel electrophoresis.

Application to B. anthracis

Given the relatively low overall size of most bacterial tandem repeats, tandem repeat search can be run even on unfinished sequences. Tandem Repeats Finder was applied to *B. anthracis* sequence obtained from The Institute for Genomic Research through the website at [http://www.tigr.org] . The sequence was recovered as approximately 1000 contigs, for a total amount of slightly more than 5 Mb. Thirty tandem repeats have at least 10 copies of a repeat unit longer than 9 base-pairs. Fourteen of them are polymorphic among the 31 *B. anthracis* strains typed here (Table 2). Twenty-seven different genotypes are identified. Polymorphism information content (PIC) indexes based on the 27 genotypes vary from 0.07 to 0.82. Nine PIC values are above 0.5. Eight alleles are identified for CEB-Bams30, in a size range 270-900 base-pairs (Figure 5). In this case, the resolution of the largest alleles would probably be improved by using an automated DNA sequencer, and more alleles might be resolved. There are clear gaps in the size range coverage shown in Figure 5, and it is likely that the typing of additional strains would uncover new alleles. The genotyping data obtained was used to construct a phylogenetic tree based upon the Neighbor-Joining method ([http://www.infobiogen.fr]). In order to be able to correlate the tree obtained here with earlier studies [2], 5 minisatellites and one microsatellite reported previously were also typed. Figure 6 presents the data obtained and the resulting tree, using the nomenclature previously proposed [2]. Six *Bacillus cereus* strains have also been included and used as an outgroup in the analysis. Occasionally *B. cereus* strains will not amplify (scored as 0 in Figure 6) or will give weak amplification signals (Figure 5, last six lanes on the right). The proposed tree is in good agreement with earlier results. In particular, the A and B clusters are well defined. We have apparently no representatives for the A1b and A3a group, whereas strains 9533 and 9502 to 9505 appear to define a new branch. The correspondence between allele numbering and allele size is indicated in Table 3.

Correlations between polymorphism and structural characteristics of minisatellites

We have looked for correlations between on one hand the number of alleles and polymorphism of the minisatellites, and on the other, simple structural characteristics of the tandem repeats in the sequenced strain : motif size, number of motifs, total length, conservation of the motifs along the array (percent identity), GC content, strand bias. In the case of *B. anthracis*, a highly significant correlation (0.01 level) is observed between polymorphism and both total length and GC content. This is not true for *Y. pestis* in which a strong correlation is seen between the number of alleles and the conservation of the motifs (Figure 7).

Conclusions

We limited here our investigation of tandem repeats to minisatellites, i.e. repeat units longer than 9 base-pairs, so as to avoid simple sequence contingency loci [8] of limited epidemiological value, and to facilitate the typing of alleles with agarose gel electrophoresis. However, simple sequence contingency loci are also represented in the database and are of great interest for molecular pathogenicity studies [6–8]. The use of the tandem repeats database was demonstrated here on two of the most genetically homogeneous human pathogenes, *Y. pestis* and *B. anthracis*. There is consequently a possibility that a common database format for identification and epidemiological analyses of pathogens amenable to minisatellite typing be developed. As more data becomes available on polymorphism associated with tandem repeats, it will be added to the database presented here in order to avoid duplication of work and nomenclature.

Bacterial species differ very significantly in the density of tandem repeats within their genome, and also in their use of tandem repeats. Some species have a very strong excess of tandem repeats with repeat units length which are multiple of three, the most striking examples being *M. tuberculosis* and *P. aeruginosa*. Polymorphism in such tandem repeats is likely to modulate the protein structure rather than gene activity. In *M. tuberculosis*, all tandem repeats with total length (L) higher than 100 bp and 9 or 15 base-pairs long units are located with ORFs [21]. An important proportion of these tandem repeats correspond to the so-called PE and PPE multigene families [21].

In the two species studied here, tandem repeat polymorphism is strongly correlated with one or more of the sequenced allele characteristics, as illustrated in Figure 7. In *Yersinia pestis* a strong correlation is observed between number of alleles observed and homogeneity of the tandem array. In *Bacillus anthracis*, the strongest correlations are with total array length and GC content. It appears that the correlations are not the same in the two species, so that at present at least, the polymorphism associated with a tandem repeat cannot be inferred from its primary sequence. In particular, and in contrast to what is known for microsatellites (1-5 bp repeat units),

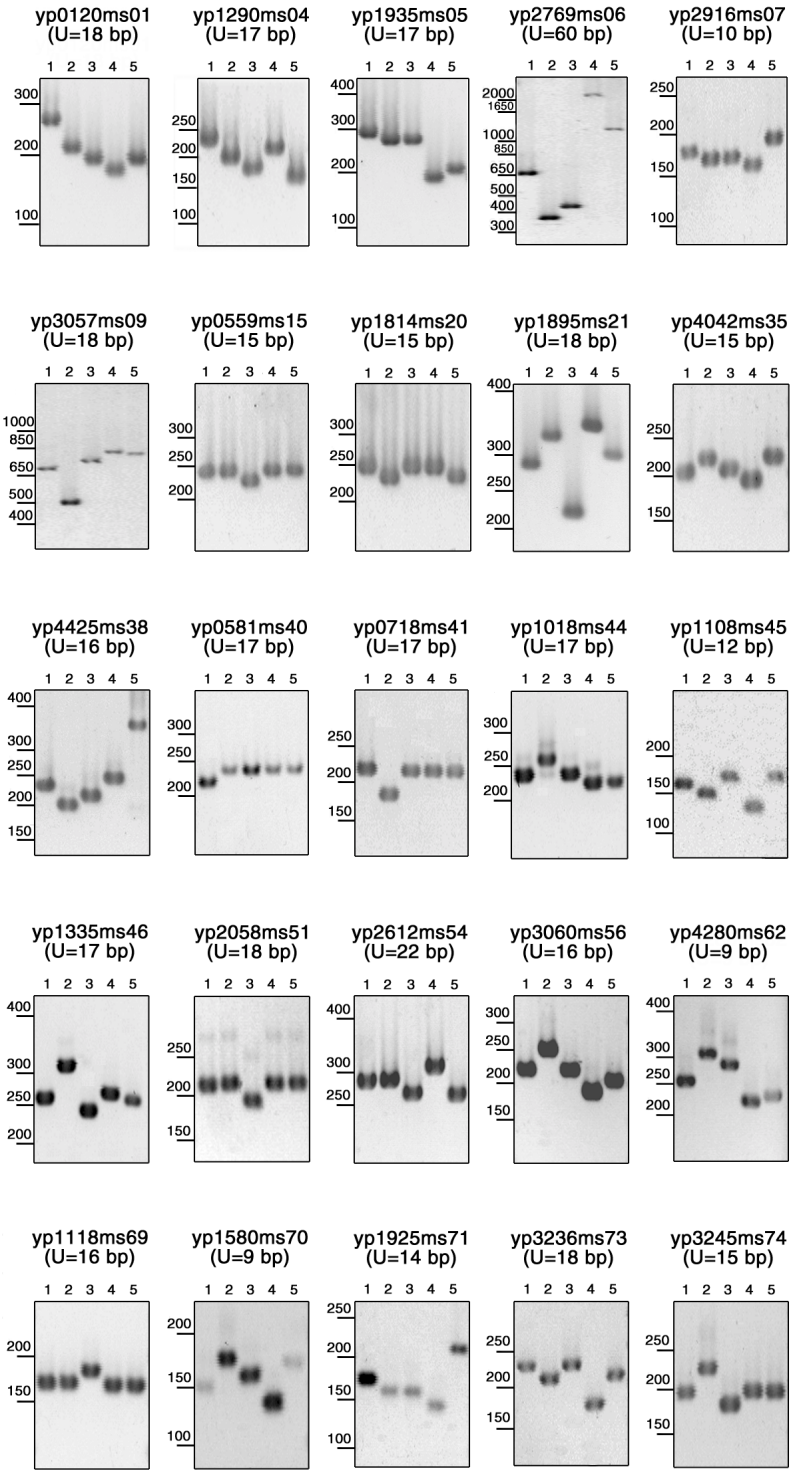

Figure 4

Images of PCR amplification of the twenty-five minisatellites polymorphic in the *Y. pestis* strains DNA from three reference *Y. pestis* strains representing each of the main biovars, antiqua (lane 1), medievalis (lane 2) and orientalis (lane 3) and two *Y. pseudotuberculosis* strains (lanes 4 and 5) have been PCR amplified and an aliquot of the products has been run on 2% horizontal agarose gels as described. The length of the minisatellite motifs (U) and the size range is indicated on each panel. Yp2916ms07 has one of the shortest (10 bp) unit. Four alleles are clearly distinguished between the 150 and 200 bp marker fragments.

some of the minisatellites are highly polymorphic in spite of a poor internal homogeneity of the sequenced allele, as is also the case for minisatellites in the human genome [12]. However, more systematic allele sequencing will be required to demonstrate that polymorphism is not associated with a subclass of alleles showing a higher internal homogeneity. Similarly, allele sequencing will be required to formally establish that the allele size variations observed are indeed (as is likely) the consequence of variations in the number of repeats.

Five among the *B. anthracis* markers described here (Ceb-Bams1, 3, 7, 13 and 30) are highly polymorphic with PIC values (or Nei's index) above 0.7. In this respect, it is important to observe that the length of the allele observed for Ceb-Bams1 in the Ames strain is not of the size expected from the sequence data (Table 2). This may result either from a high mutation rate at Ceb-Bams1 or from a sequencing error. The expected allele size corresponds to allele 4 (Table 3), which is unlikely for the Ames strain because Ceb-Bams1 allele 4 is observed only in cluster B strains (Figure 6) and Ames is well apart of cluster B [2]. A similar situation is observed for Ceb-Bams28, for which the expected product does not correspond to any existing allele in the collection of strains typed. In this case however, the locus is moderately polymorphic, with a PIC value of 0.26 and only three alleles observed (Table 2), so that a sequencing error is the most likely interpretation. This issue could be easily solved by typing with Ceb-Bams1 and Ceb-Bams28 the very strain which has been used for the sequencing project.

It is interesting to observe that, although the magnitude of allele size difference has not been taken into account when building the distance matrix, the resulting phylogenetic tree proposed in Figure 6 tends to group together strains with alleles of similar size at these most variable loci. This is reminiscent of observations made in *H. influenzae* [1] and suggest that mutation events are predominantly small size changes. Here again, more detailed studies involving full allele sequencing should now help understand the succession of events producing a population of alleles.

Materials and methods

Bacterial genomes DNA sequences

Finished sequences in the public domain were recovered by ftp from the NCBI or the Sanger center sites ([http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html] ; [http://www.sanger.ac.uk/Projects/Microbes/]). Preliminary sequence data for *B. anthracis* was obtained from The Institute for Genomic Research through the website at [http://www.tigr.org] .

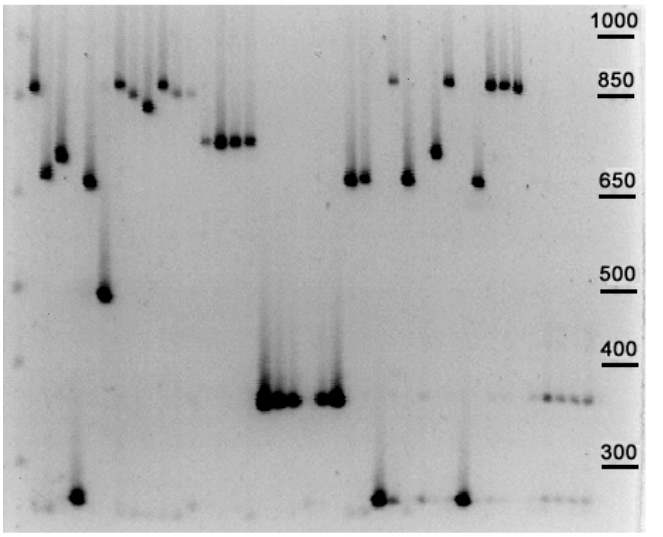

Figure 5
PCR amplification of *B. anthracis* minisatellite CEB-Bams30 DNA from *B. anthracis* and *B. cereus* (six rightmost lanes) was amplified using primers for CEB-Bams30 (Table 2). The PCR products were run on a 40 cm long 2% ordinary agarose gel.

DNA preparation

All strains used here are part of the collection maintained by the Centre d'Etudes du Bouchet (CEB). They originate either from the CIP (Collection Institut Pasteur, [http://www.pasteur.fr/]) or from AFSSA (Agence Française de Sécurité Sanitaire des Aliments, [http://www.afssa.fr/] , Dr Josée Vaissaire). DNA from each isolate was obtained by large-batch procedures or by the simplified procedure as described in [2]. In addition, 15 µg of DNA from the *B. anthracis* Ames strain were kindly provided by Dr Mats Forsman, FOA, Sweden.

Minisatellite PCR amplification and genotyping

PCR reactions were performed in 15 µl containing 1 ng of DNA, 1x Long Range Reaction Buffer 3 (Roche-Boehringer), 1 unit of Taq DNA polymerase, 200 µM of each dNTP, 0.3 µM of each flanking primer. The Taq DNA polymerase was either prepared essentially as described in [22] or purchased from Qbiogen or Roche-Boehringer. The 1x LongRange Buffer 3 is 1.75 mM $MgCl_2$, 50 mM Tris-HCl pH9.2, 16 mM $(NH_4)2SO_4$.

PCR reactions were run on a Perkin-Elmer 9600 or a MJResearch PTC200 thermocycler. An initial denaturation at 96°C for five minutes was followed by 34 cycles of denaturation at 96°C for 20 seconds, annealing at 60°C for 30 seconds, elongation at 65°C for 1 minute, followed by a final extension step of 5 minutes at 65°C. In few cases, other annealing temperatures and/or elongation times were used (see tables 1 and 2). Five microliters of

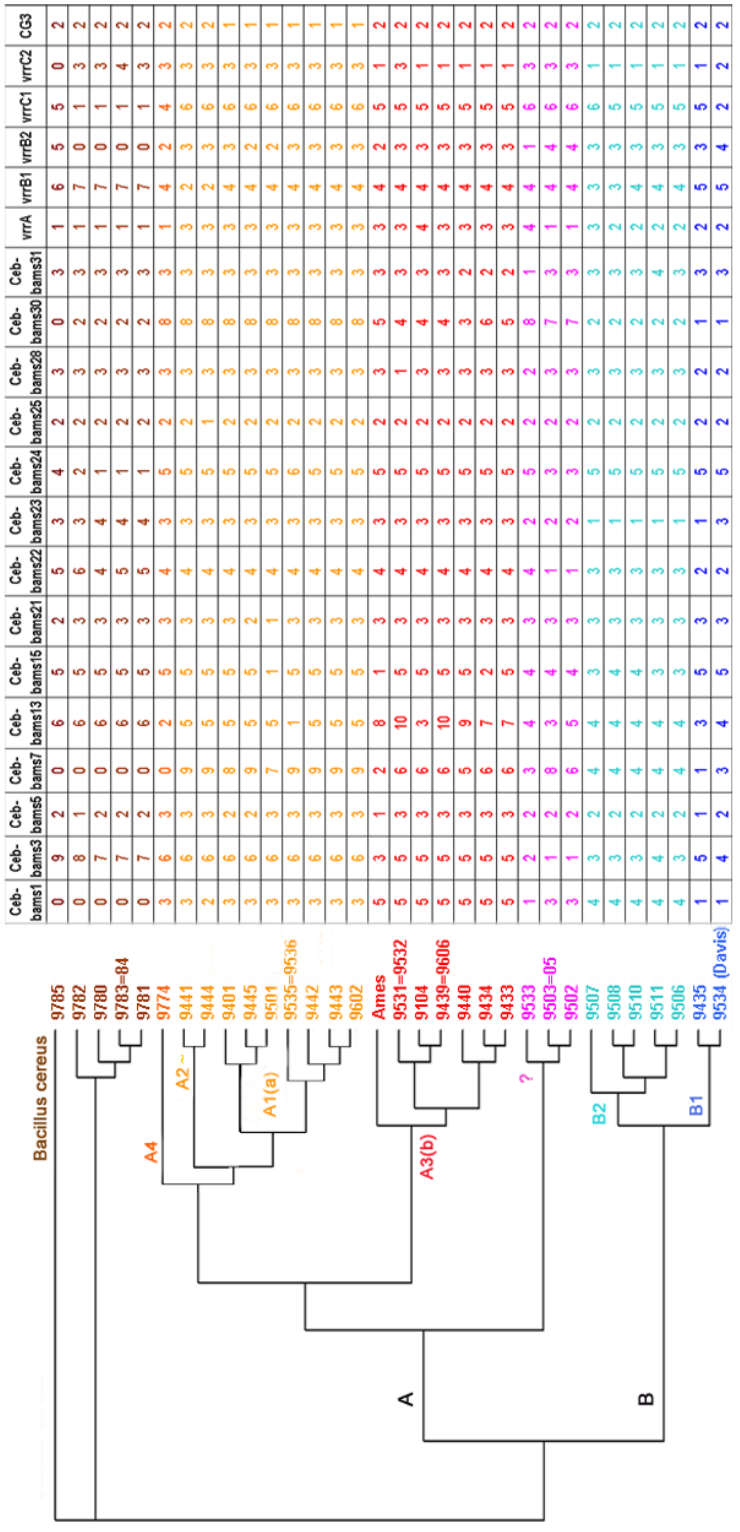

Figure 6
***Bacillus anthracis* phylogenetic tree** The genotype of each strain for the polymorphic minisatellites is given (size estimates for each allele are given in Table 3). "0" indicates a failure of the PCR amplification. This is most often associated with *B. cereus* strains, and probably reflects in these cases sequence divergence in the flanking sequence. The phylogenetic tree was produced using the Neighbor-Joining method as available on-line at [http://www.infobiogen.fr.]

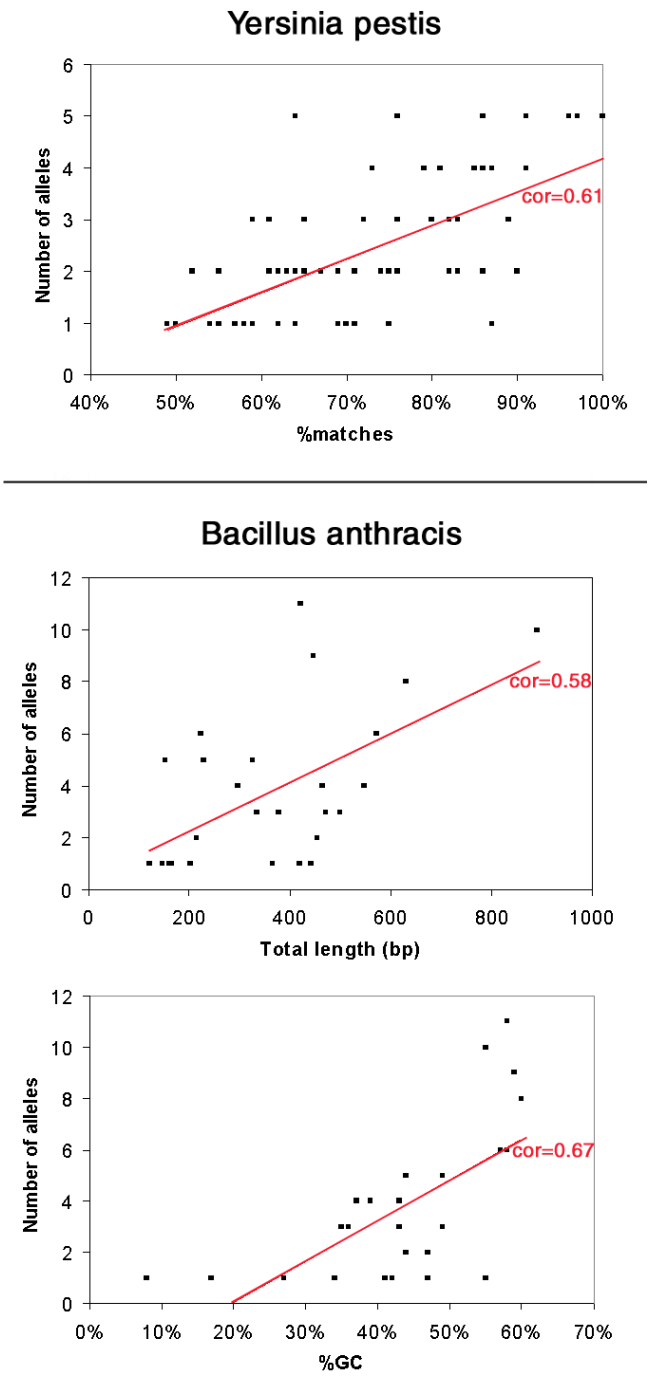

Figure 7
Significant correlation between number of alleles and minisatellites structural characteristics The number of alleles is plotted as a function of Total length and %GC for *Bacillus anthracis*, and %matches for *Yersinia pestis* (the correlations are highly significant at the 0.01 level). Number of alleles for each locus is the total number detected (i.e. *Bacillus anthracis* and *B. cereus*; *Yersinia pestis* and *Y. pseudotuberculosis*).

the PCR products where run on standard 1% or 2% agarose gel (Qbiogen) in 0.5 x TBE buffer at a voltage of 10 V/cm as indicated in Tables 1 and 2. Gel length of 10 to 40 cm were used according to PCR product size and motif length. Gels were stained with ethidium bromide and visualized under UV light. Allele sizes were estimated using as size markers the 1 kb ladder plus (Gibco-BRL which also includes a 100 bp ladder between 100 bp and 500 bp, plus 650, 850 and 1000 bp bands) or the 50 bp ladder (Euromedex) which provides a 50 bp ladder between 50 and 300 bp and a 100 bp ladder from 300 bp to 1000 bp.

Data analysis

Tandem Repeats Finder analysis:

Sequences were processed using the Tandem Repeats Finder software ([http://c3.biomath.mssm.edu/trf.html]). The output was processed to eliminate duplicates before being imported in a database (running under Access2000, Microsoft Corp.) as described previously [12]. The *B. anthracis* preliminary sequence data file uses FASTA type of headers (i.e. >sequenceId) to separate the independent contigs. The headers were replaced by runs of 10 Ns before running Tandem Repeats Finder.

Blast queries against the M. tuberculosis genome:

The identifications of the open reading frames containing a given tandem repeat from M. tuberculosis were done by running a BLAST search on the dedicated web page at [http://www.sanger.ac.uk/Projects/M_tuberculosis/blast_server.shtml] .

Estimation of the excess of tandem repeats with motif length multiple of three:

A χ^2 test was calculated for the difference between the observed number of tandem repeats with motif length multiple of 3 and the expected number of tandem repeats with motif length multiple of 3 (expected value in the absence of bias being the total number of tandem repeats divided by 3). The χ^2 values vary from 0.01 to 253.5. There is a significant excess ($\chi^2 > 3.841$) for all species but 6 (*Buchnera sp*, *T. maritima*, *H. influenzae*, *M. genitalium*, *R. prowazekii*, *Y. pestis*).

Polymorphism index:

Polymorphism Information Index (PIC) or Nei's diversity index is calculated as $1 - \Sigma (\text{allele frequency})^2$ based upon the unique genotypes.

Phylogenetic reconstruction:

A phenetic approach, based on a distance matrix was used. Distance matrix between strains was obtained by counting the number of differences between the corresponding genotypes. Then, Neighbor Joining cluster

analysis was performed with Phylip [23] accessed at [http://www.infobiogen.fr/] . An outgroup was arbitrary chose among *Bacillus cereus* strains (9785) and input order of species was randomised.

Data (genotypes, distance matrix, phylogenetic tree) are available at [http://minisatellites.u-psud.fr/ASPSamp/Phylogenie/data.htm]

Correlation analysis

Correlations were calculated with the statistical program SPSS: Pearson correlation, and non-parametric correlations (Kendall's tau and Spearman's rho) show similar results.

Table 1: Description of *Yersinia* polymorphic markers

Marker	U	N	% GC	V	Primer sequences	PCR	Expected product length (bp)	Estimated size range (bp)	Number of Alleles in *Y.pestis*	Total number of alleles
Markers polymorphic in *Yersinia pestis* strains										
yp0120ms01	18	8	34	86	**L:** CTAAGCACAATTGTTATGCTGAACC **R:** TACTGAATCTGCTTCATTGTTCAAA		228	180 - 280	3	4
yp1290ms04	17	8	27	96	**L:** CGCTGTTGAAGTTTTAGTGTAAGAA **R:** AAATGTAACCTGCCAAACGTG		230	160 - 240	3	5
yp1935ms05	17	11	36	87	**L:** CCTCAGTTCATTGTGTAAAATCTCA **R:** GTATTAGCGAGATCACAGATGAGC		291	190 - 300	2	4
yp2769ms06	60	9	48	64	**L:** AATTTTGCTCCCCAAATAGCAT **R:** TTTTCCCCATTAGCGAAATAAGTA	90 s	606	370 - 2500	3	5
yp2916ms07	10	9	44	85	**L:** ATACCGCTACGATCAGCCTCTAT **R:** ATTTAATATTGATTTTGGGACTTGC		184	150 - 200	2	4
yp3057ms09	18	33	65	91	**L:** CGTTACCCTTGTTGCCAATAGT **R:** ACGCAGAACATGCTTACCTTTTAT	90 s	682	500 - 820	3	5
yp0559ms15	15	10	30	62	**L:** TTGACCAAGTGTAAAAGGCATAAAT **R:** AAACTATCGCCAGCCATTTTAGTA		237	225 - 250	2	2
yp1814ms20	15	9	47	74	**L:** ACAACCTCAGTTTGCCCTTG **R:** GTAAAGAGCGCAATGATCGTACT		253	230 - 250	2	2
yp1895ms21	18	9	51	76	**L:** GCTTAAAGCAGATTGATACTCACG **R:** CTGCATGTTCACCGGTTCAG		278	220 - 350	3	5
yp4042ms35	15	8	41	59	**L:** CTGTTACCGGTCAAAGTGGATATT **R:** AGGCTCTCCTTATCATTATTTGGTC		204	195 - 225	2	3
yp4425ms38	16	8	41	86	**L:** GTGAGGTATAGCTAAACGGTGATGT **R:** CGCCGTAGATTATTTGTCACTTTAT		233	200 - 380	3	5
yp0581ms40	17	7	28	76	**L:** GCAATCATTCACCTAACCATATCTC **R:** GTGCAATAGGCGTTGTTGTGTA		214	220 - 240	2	2
yp0718ms41	17	7	41	75	**L:** GAAGAAAGCCAGCTAATCTGATG **R:** TAATGAATAGCAACGACAACCAATA		217	180 - 220	2	2
yp1018ms44	17	7	38	61	**L:** CAATTCCAACAGCTATTAATGCAA **R:** GAATTTTCATAACACGTTCTTCCTG		233	220 - 260	2	3
yp1108ms45	12	7	65	79	**L:** GCATCGGAGACTGGGTAAAC **R:** TTTCTGAGGATTTATCGGTGTGAT		161	120 - 170	3	4
yp1335ms46	17	7	33	73	**L:** CAGGTTTTACGTTATTTTCTGAAGG **R:** CAGCATGAAGTATGACGGGTATATTA		252	230 - 310	3	4
yp2058ms51	18	7	37	65	**L:** GGTTTTTACCGATATAAATCCTGAG **R:** GACCAAGAAGTTAAGTTGCTTATCG		207	190 - 210	2	2
yp2612ms54	22	7	28	82	**L:** GTCCACCATTTTCATACTGTCACTT **R:** GCTCTTTGTTCGATTTTATTGAATG		281	250 - 300	2	3
yp3060ms56	16	7	21	81	**L:** AACCGACTGACTCACTTTATATTGG **R:** TTCTTTTCCATTACTCAGCCTGTT		220	180 - 250	2	4
yp4280ms62	9	7	33	60	**L:** TTTAGTCTTGATTAAGCTGCGTTTT **R:** ACGGAAGACAACCTTATTATTGATG		240	220 - 310	3	5
yp1118ms69	16	6	39	82	**L:** GACGTTGCAACTGCAAAAATAAG **R:** ACTTGTTGTGAAGACCATCACTCT		179	165 - 180	2	2
yp1580ms70	9	6	32	97	**L:** AAACCAACGGTTCATATTGAATAAA **R:** CTTCTTCCGCTATTTTCCTACAGA		146	140 - 170	3	5
yp1925ms71	14	6	45	91	**L:** GCTACTCGAATATGAGTTAGCCAAA **R:** ATTGCCATATTGGATGCTAAAATAA		171	145 - 210	2	4
yp3236ms73	18	6	40	89	**L:** AATACCCTGTGGGTGATAATGAAC **R:** ATCGATTTAGGTACCACCAATTCA		225	175 - 230	2	3
yp3245ms74	15	6	44	83	**L:** CCCCGACTTATATCAAGCACTG **R:** AACTGACGATCTTTTTCACTGAGTT		195	180 - 225	3	3

Table 1: Description of *Yersinia* polymorphic markers

Markers polymorphic in 5 *Yersinia* strains (monomorphic in *pestis*)

yp0802ms02	**18**	**12**	**49**	**86**	**L:** CTGACACAAAACGAGAGCCTATTT **R:** AGCGTGAGTGGGCTATCAATAC	53°C 1 min	**314**	**240 - 315**	**1**	**2**
yp2925ms08	**15**	**12**	**39**	**63**	**L:** AGCCTTTTTGTTGATTATCAGTCAT **R:** CGATAATAATACTGAATTACCGGATG		**270**	**270 - 290**	**1**	**2**
yp4411ms10	**14**	**8**	**32**	**69**	**L:** ATCATGCTTTTGCCTCAATATAATC **R:** GAAACGCAGTCCCTGTTGTAG		**191**	**190 - 210**	**1**	**2**
yp0813ms16	**17**	**8**	**39**	**64**	**L:** GTTGGTTATCCGACAGTCTTCAATA **R:** GCAATTCGTTATGGCTTAGTAAAAA		**235**	**230 - 270**	**1**	**2**
yp1269ms18	**27**	**9**	**54**	**55**	**L:** GCAAAAGCTGAAGCAGATAAAATAG **R:** AAACCACCAAACAAATCATCAAC		**303**	**220 - 250**	**1**	**2**
yp2196ms22	**20**	**8**	**12**	**55**	**L:** AAACCAACAAGAAAACTGTAACCAC **R:** CATTCACCATTGATGTCCTTAGAC	90 s	**265**	**270 - 1500**	**1**	**2**
yp2324ms24	**19**	**8**	**34**	**65**	**L:** TTCACCGGGTTACCTTAATTACATA **R:** CTACCTTGCTGTCAACACTCGAC		**255**	**215 - 255**	**1**	**2**
yp2331ms25	**17**	**9**	**36**	**76**	**L:** AACGCGTTAATAAAACAATAAAGTG **R:** CAATATCCTTTTACTCAGCCGATG		**181**	**190 - 230**	**1**	**3**
yp2679ms27	**16**	**8**	**20**	**76**	**L:** ATGATTACTGGCAAGAGCACTATGT **R:** AACAAGATCACCTGGTCGTTAAA		**217**	**200 - 220**	**1**	**2**
yp2908ms28	**18**	**8**	**40**	**69**	**L:** GCAGAAATAATCTTCAGGAGAAACA **R:** AGATCGTCGTTAGTCCATGTCAG		**242**	**190 - 290**	**1**	**2**
yp2958ms29	**16**	**8**	**23**	**61**	**L:** AAAATAGTCTGTGTTTCAGCAAAGC **R:** CCTTAAAAACCCTAAGTGGGTAAAA		**215**	**215 - 245**	**1**	**2**
yp3225ms30	**54**	**11**	**51**	**52**	**L:** CAATAATACCATCGTGCGTGATAC **R:** TATTAATGGTGGTGTTAGTCGCTGT		**683**	**680 - 900**	**1**	**2**
yp3532ms31	**14**	**8**	**30**	**67**	**L:** GTTATTTATTTTTGCCCCAACTTGT **R:** TTAGCCTGTTTGTTCTTCAAATAGC		**217**	**215- 245**	**1**	**2**
yp3787ms32	**18**	**8**	**49**	**65**	**L:** CGATAACGTTAATGCCATCAGTAG **R:** GCGCCGGTAAGTTTTGTTTATTA		**218**	**190 - 240**	**1**	**3**
yp3795ms33	**15**	**8**	**43**	**67**	**L:** CCCTTCTTTTTATGCTTGAAGATACT **R:** GTTGAACCACAGGCTGTTGAG		**210**	**210 - 225**	**1**	**2**
yp4371ms37	**18**	**8**	**35**	**82**	**L:** TACTTAGGCATTGTCTCTTCACTCC **R:** CTGAAATTATCAGTAGGTGGTTCGT		**235**	**235 - 255**	**1**	**2**
yp0999ms43	**17**	**7**	**38**	**80**	**L:** ATTCCACCACCAACAATTATCAC **R:** GGTATTCGTATTGAAGATGACATTG		**211**	**220 - 300**	**1**	**3**
yp1962ms50	**18**	**7**	**34**	**71**	**L:** TACCGAGGTATTCCCTGGTCTAAT **R:** AGTTGACTCCCAGTCACTTTTCC		**225**	**225 - 240**	**1**	**2**
yp3734ms59	**16**	**7**	**36**	**69**	**L:** ATTATCATGACCCTTCCAGTGCTAT **R:** CATCAAAATGCCAGGAGAGTAAC		**218**	**200 - 220**	**1**	**2**
yp4338ms63	**17**	**7**	**38**	**72**	**L:** ATTAACGATTTCTTGTCGCTCAGT **R:** AATCAGTAACGGCATGTGTCAGTA		**194**	**190 - 275**	**1**	**3**
yp0549ms66	**18**	**6**	**41**	**83**	**L:** TAAAAGCGTCAACAAAGTAGGTCAT **R:** GTTCCTGTTGTTGAAAATGCTG		**212**	**200 - 220**	**1**	**2**
yp0782ms67	**18**	**6**	**40**	**90**	**L:** TTCCAGGCTAAAGATATTGACTTTG **R:** CTCGGCTTGTTCTACGTTTAATG		**248**	**250 - 270**	**1**	**2**
yp1053ms68	**18**	**6**	**32**	**82**	**L:** CCGTTATCTGGTGAAAGTGAACAG **R:** GTCCGGTAGCCTGATTGTTTATT		**182**	**175 - 205**	**1**	**3**
yp3634ms75	**15**	**6**	**36**	**80**	**L:** ATGTGAGCTTGATTGCTGAGTAGT **R:** TCATATTTTAGGTGTTTTGCCTTTG		**210**	**180 - 210**	**1**	**3**

Some structural characteristics of the tandem repeats are presented : U (unit length), N (number of repeats), %GC, V (% of conservation). PCR and electrophoresis conditions are as described in the material and methods section : annealing temperature is 60°C, elongation time is 60 seconds and gels are 2% agarose except when indicated otherwise. Total number of alleles means number of alleles in 3 *Y. pestis* and 2 *Y. pseudotuberculosis* strains.

Table 2: Description of *Bacillus anthracis* polymorphic markers

Marker	U	N	% GC	V	Primer sequences	PCR	Expected product length in bp (observed)	Estimated size range (bp)	Number of alleles in B. *anthracis*	Total number of alleles	PIC index
Ceb-Bams 1	21	16	44	88	**L:** GTTGAGCATGAGAGGTACCTTGTCCTTTTT **R:** AGTTCAAGCGCCAGAAGGTTATGAGTTATC		485 (520)	410-520	5	5	0.72
Ceb-Bams 3	15	25	59	73	**L:** GCAGCAACAGAAAACTTCTCTCCAATAACA **R:** TCCTCCCTGAGAACTGCTATCACCTTTAAC	1%	544	480-860	6	9	0.75
Ceb-Bams 5	39	10	49	92	**L:** GCAGGAAGAACAAAAGAAACTAGAAGAGCA **R:**ATTATTAGCAGGGGCCTCTCCTGCATTACC	53°C 60s	307	305-385	3	3	0.56
Ceb-Bams 7	18	49	55	69	**L:** GAATATTCGTGCCACCTAACAAAACAGAAA **R:** TGTCAGATCTAGTTGGCCCTACTTTTCCTC	65°C 1%	1017	600-1950	9	9	0.82
Ceb-Bams 13	9	70	60	79	**L:** AATTGAGAAATTGCTGTACCAAACT **R:** CTAGTGCATTTGACCCTAATCTTGT	120s 1%	814	330-850	8	11	0.79
Ceb-Bams 15	18	12	57	77	**L:** GTATTTCCCCCAGATACAGTAATCC **R:** GTGTACATGTTGATTCATGCTGTTT		409	410-610	5	5	0.59
Ceb-Bams 21	45	11	43	75	**L:** TGTAGTGCCAGATTTGTCTTCTGTA **R:** CAAATTTTGAGATGGGAGTTTTACT		676	540-680	3	3	0.14
Ceb-Bams 22	36	15	39	81	**L:** ATCAAAAATTCTTGGCAGACTGA **R:** ACCGTTAATTCACGTTTAGCAGA		735	590-950	4	6	0.51
Ceb-Bams 23	42	11	37	85	**L:** CGGTCTGTCTCTATTATTCAGTGGT **R:** CCTGTTGCTCCTAGTGATTTCTTAC		653	570-820	3	4	0.49
Ceb-Bams 24	42	11	44	80	**L:** CTTCTACTTCCGTACTTGAAATTGG **R:** CGTCACGTACCATTTAATGTTGTTA		630	335-670	3	6	0.2
Ceb-Bams 25	15	14	45	60	**L:** CCGAATACGTAAGAAATAAATCCAC **R:** TGAAAGATCTTGAAAAACAAGCATT		391	375-390	2	2	0.07
Ceb-Bams 28	24	14	36	70	**L:** CTCTGTTGTAACAAAATTTCCGTCT **R:** TATTAAACCAGGCGTTACTTACAGC		493 (400)	300-400	3	3	0.26
Ceb-Bams 30	27	16	58	78	**L:** AGCTAATCACCTACAACACCTGGTA **R:** CAGAAAATATTGGACCTACCTTCC	120s 1%		200-890	11	11	0.77
Ceb-Bams 31	9	64	58	57	**L:** GCTGTATTTATCGAGCTTCAAAATCT **R:** GGAGTACTGTTTGTTGAATGTTGTTT	1%	772	300-850	4	4	0.32

Some structural characteristics of the tandem repeats are presented : U (unit length), N (number of repeats), %GC, V (% of conservation). PCR and electrophoresis conditions are as described in the material and methods section : annealing temperature is 60°C, elongation time is 60 seconds and gels are 2% agarose except when indicated otherwise. The expected product length is deduced from the sequencing data corresponding to the Ames strain. When the Ames strains typing does not fit with the expected value, the observed value is indicated between (). Only one side of the Ceb-Bams30 minisatellite can be identified in the available Ames sequence. The other side was identified in the course of the independent, partial sequencing of *B. anthracis* strains (Vergnaud and col., unpublished data). Total number of alleles includes alleles observed in the *B. cereus* strains. Polymorphism Information Index (PIC) or Nei's diversity index is calculated as $1 - \Sigma (\text{allele frequency})^2$.

Table 3: Correspondence between *B. anthracis* allele sizes and allele numbering

allele nb marker name	1	2	3	4	5	6	7	8	9	10
Ceb-Bams1	~ 410	~ 430	~ 450	~ 480	~ 520					
Ceb-Bams3	484	514	544	559	574	589	704	734	857	
Ceb-Bams5	307	346	385							
Ceb-Bams7	603	1017	1305	1503	1557	1647	1809	1899	1953	
Ceb-Bams13	328	382	454	481	490	652	742	787	814	850
Ceb-Bams15	409	535	571	589	607					
Ceb-Bams21	541	631	676							
Ceb-Bams22	591	627	699	735	~ 900	~ 950				
Ceb-Bams23	569	611	653	821						
Ceb-Bams24	336	420	462	504	630	672				
Ceb-Bams25	376	391								
Ceb-Bams28	~ 300	~ 375	~ 400							
Ceb-Bams30	266	375	500	660	695	730	760	850 to 900		

Table 3: Correspondence between *B. anthracis* allele sizes and allele numbering

Ceb-Bams31	304	700	772	853			
vrrA	289	301	313	325	337		
vrrB1	184	193	220	229	256	~ 280	~ 290
vrrB2	~ 135	153	162	171	~ 180		
vrrC1	400	502	520	538	583	613	685
vrrC2	532	568	607	660			
CG3	153	158					

Alleles have been numbered in increasing size order. When the allele size (in base-pairs) observed in the Ames strain was in agreement with the size expected according to Ames sequence data, the values indicated in the table assume that alleles differ in size by a multiple of the motif length. These likely values will have to be confirmed by more accurate size estimation tools and allele sequencing. When the allele size in Ames is not as expected (Ceb-Bams1 and Ceb-Bams28), the estimated values are preceded by a ~. The Vrr and CG3 allele sizes were described in [2]; new alleles are indicated by a ~.

Acknowledgements

Minisatellite investigations in the laboratory are supported by grants from Délégation Générale de l'Armement (DGA/DSA/STTC and DGA/DSA/SP-Nuc). Preliminary sequence data for *B. anthracis* was obtained from The Institute for Genomic Research through the website at [http://www.tigr.org] . Sequencing of *B. anthracis* was accomplished with support from Office of Naval Research, Department of Energy, and National Institute of Allergy and Infectious diseases. The *Yersina pestis* sequence data were produced by the *Y. pestis* Sequencing Group at the Sanger Centre and can be obtained from [ftp://ftp.sanger.ac.uk/pub/pathogens/yp/]. We wish to thank the referees for the significant improvements they have suggested.

References

1. van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H: **Variable number of tandem repeats in clinical strains of Haemophilus influenzae.** *Infect Immun* 1997, **65**:5017-27
2. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within Bacillus anthracis.** *J Bacteriol* 2000, **182**:2928-2936
3. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144**:1189-96
4. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.** *Mol Microbiol* 2000, **36**:762-71
5. Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P: **Diversity in a variable-number tandem repeat from Yersinia pestis.** *J Clin Microbiol* 2000, **38**:1516-9
6. van Ham SM, van Alphen L, Mooi FR, van Putten JP: **Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region.** *Cell* 1993, **73**:1187-96
7. Weiser JN, Love JM, Moxon ER: **The molecular mechanism of phase variation of H. influenzae lipopolysaccharide.** *Cell* 1989, **59**:657-65
8. Bayliss CD, Field D, Moxon ER: **The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis.** *J Clin Invest* 2001, **107**:657-66
9. Henderson IR, Owen P, Nataro JP: **Molecular switches - the ON and OFF of bacterial phase variation.** *Mol Microbiol* 1999, **33**:919-32
10. Wang G, Ge Z, Rasko DA, Taylor DE: **Lewis antigens in Helicobacter pylori: biosynthesis and phase variation.** *Mol Microbiol* 2000, **36**:1187-96
11. Wilton JL, Scarman AL, Walker MJ, Djordjevic SP: **Reiterated repeat region variability in the ciliary adhesin gene of Mycoplasma hyopneumoniae.** *Microbiology* 1998, **144**:1931-43
12. Vergnaud G, Denoeud F: **Minisatellites: Mutability and Genome Architecture.** *Genome Res* 2000, **10**:899-907
13. Kokoska RJ, Stefanovic L, Tran HT, Resnick MA, Gordenin DA, Petes TD: **Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t).** *Mol Cell Biol* 1998, **18**:2779-88
14. Debrauwère H, Buard J, Tessier J, Aubert D, Vergnaud G, Nicolas A: **Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks.** *Nat Genet* 1999, **23**:367-71
15. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW, Moxon ER: **The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases.** *Mol Microbiol* 2000, **35**:211-22
16. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes.** *Microbiol Mol Biol Rev* 1998, **62**:275-93
17. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: **Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis [published erratum appears in Proc Natl Acad Sci U S A 2000 Jul 5;97(14):8192].** *Proc Natl Acad Sci U S A* 1999, **96**:14043-8
18. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, Hegna I, Kolsto : **Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis - one species on the basis of genetic evidence.** *Appl Environ Microbiol* 2000, **66**:2627-30
19. Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, Adair DM, Hugh-Jones M, Kuske CR, Jackson P: **Molecular evolution and diversity in Bacillus anthracis as detected by amplified fragment length polymorphism markers.** *J Bacteriol* 1997, **179**:818-24
20. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-80
21. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, *et al*: **Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.** *Nature* 1998, **393**:537-44
22. Engelke DR, Krikos A, Bruck ME, Ginsburg D: **Purification of Thermus aquaticus DNA polymerase expressed in Escherichia coli.** *Anal. Biochem.* 1990, **191**:396-400
23. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166