Available online at www.sciencedirect.com
ScienceDirect
Biochimie 90 (2008) 660–668
BIOCHIMIE
www.elsevier.com/locate/biochi

Review

On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing

Ibtissem Grissa [a,*], Patrick Bouchon [a], Christine Pourcel [a], Gilles Vergnaud [a,b]

[a] *Univ Paris-Sud, Institut de Génétique et Microbiologie, Orsay F-91405, France; CNRS, Orsay F-91405, France*
[b] *Division of Analytical Microbiology, Centre d'Etudes du Bouchet, 5 rue Lavoisier, 91710 Vert le Petit, France*

Received 16 June 2007; accepted 19 July 2007
Available online 28 July 2007

Abstract

The control of bacterial pathogens requires the development of tools allowing the precise identification of strains at the subspecies level. It is now widely accepted that these tools will need to be DNA-based assays (in contrast to identification at the species level, where biochemical based assays are still widely used, even though very powerful 16S DNA sequence databases exist). Typing assays need to be cheap and amenable to the designing of international databases. The success of such subspecies typing tools will eventually be measured by the size of the associated reference databases accessible over the internet. Three methods have shown some potential in this direction, the so-called spoligotyping assay (*Mycobacterium tuberculosis*, 40,000 entries database), Multiple Loci Sequence Typing (MLST; up to a few thousands entries for the more than 20 bacterial species), and more recently Multiple Loci VNTR Analysis (MLVA; up to a few hundred entries, assays available for more than 20 pathogens).

In the present report we will review the current status of the tools and resources we have developed along the past seven years to help in the setting-up or the use of MLVA assays or lately for analysing Clustered Regularly Interspaced Short Palindromic Repeats called CRISPRs which are the basis for spoligotyping assays.

Keywords: VNTR; MLVA; CRISPR; Tandem repeats; Bacterial pathogens; Molecular epidemiology; Genotyping; Databases; MLVA web service; CRISPRfinder; CRISPRdb

1. Introduction

Although identification of bacteria at the species level is usually sufficient for short-term health care response, the tracing of bacterial pathogens (microbial forensics) or the identification of emerging clones escaping prophylactic or therapeutic strategies require much more precise strain identification tools. Epidemiological investigations rely upon molecular assays providing accurate and rapid differentiation of bacterial strains using some specific sites of genetic variability such as particular repeats or mutations. Various bacterial strain typing techniques have been described in the literature to differentiate strains but not all methods are equally applicable to every species. In addition many techniques cannot provide a portable result and therefore the strain genetic profile cannot be easily coded and stored into databases that can be exchanged between laboratories. This is in particular the problem of methods relying on restriction enzyme polymorphism analysed by gel electrophoresis such as restriction fragment length polymorphism (RFLP) and pulsed field gel electrophoresis (PFGE). Other pattern-producing techniques such as the random amplification of polymorphic DNA (RAPD) or amplified fragment-length polymorphism (AFLP) are respectively notably not reproducible enough or technically too demanding to allow accurate or convenient inter-laboratory comparisons of profiles. Multiple loci sequence typing (MLST) is a highly accurate, reproducible

* Corresponding author: Tel.: +33 1 6915 3001; fax: +33 1 6915 6678.
E-mail address: ibtissem.grissa@igmors.u-psud.fr (I. Grissa).

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.07.014

and portable method but is not adapted to the typing of the most highly homogenous species such as *Mycobacterium tuberculosis* and its current cost prevents its systematic use.

Multiple loci variable number of tandem repeats (VNTR) analysis (MLVA) is a typing technique based on the polymorphism of certain tandemly repeated DNA sequences. VNTRs consist in consecutive occurrences of a DNA repeat unit, and they are found in all organisms, prokaryotes as well as eukaryotes. Although their biological function and evolution mechanism is not fully understood, they have diverse practical applications including strain identification in bacterial epidemiology. In a typical MLVA assay, a few to more than twenty VNTRs, distributed over the entire bacterial genome, are analysed, and a code corresponding to the number of repeats at each locus is determined. This code is easily stored into databases and can be used for strain clustering and epidemiological studies.

MLVA is nowadays increasingly replacing or at least completing traditional genotyping methods, providing a different or complementary point of view in *M. tuberculosis*, *Bacillus anthracis*, *Yersinia pestis*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Legionella pneumophila* investigations thanks to its design easiness, low cost and portability (see refs. [1,2] for reviews). MLVA is best applied within a highly homogeneous group of strains, typically with genomes showing an average similarity well above 98%. Some MLVA assays have been developed in species with an internal genome homogeneity in the 95%–98% range (as illustrated for instance in *L. pneumophila* [3]), but the design of primers, and the level of homoplasy at VNTR loci, introduce specific technical challenges.

In parallel, the particular polymorphic structures called CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat); or TREP (Tandem REPeat) [4], SRSR (short regularly spaced repeats) [5,6], DVRs (direct variant repeats) [7], LCTR (long clusters of tandem repeats) [8], SPIDR (spacers interspersed direct repeats) [9] have been used in some bacteria for genotyping (Fig. 1). They were firstly detected in *Escherichia coli* [10] and then in about 40% of bacterial genomes and almost all archaea. A CRISPR consists in exact repeat sequences (DR for direct repeat) of approximately 24–48 bases long separated by unique spacers of similar length [11]. It has been shown that CRISPRs may be used for evolutionary studies due to their mechanism of acquisition/deletion of motifs (a repeat and a spacer) [12,13]. The CRISPR region was widely used to genotype *M. tuberculosis* strains [14–18] and to a much lesser extent for *Streptococcus pyogenes* [19] *Y. pestis* [12], *Corynebacterium diphtheriae* [20,21] and *Campylobacter jejuni* [22,23]. Presumably because of strong structural constraints, the DR sequence is more conserved than the surrounding genomic elements, so that a CRISPR typing assay (in which a single PCR amplification with primers corresponding to the DR sequence is sufficient to amplify the whole set of spacers) might in theory be applied on a larger evolutionary scale than MLVA.

In this review, we describe the current set of bioinformatics tools helping in MLVA and CRISPR assay setting-up and analysis. Before starting bench work, biologists need to accomplish *in silico* pre-processing and post-processing phases for developping the typing assay and analyzing the results [2]. The pre-processing phase consists in the identification of genetic markers from sequence data, which requires the use of bioinformatics resources. Software for sequence analysis must be developed or adapted to detect correctly and exclusively the markers of interest from completed or unfinished genomes. In addition, dedicated databases archiving the markers must also be created and regularly updated from publicly available sequenced genomes and publications. Such databases provide platforms gathering information about the markers and allowing their comparison in related species to facilitate the assay design. After the genotyping has been completed, additional tools are necessary to store, analyze and compare the findings.

We will describe the current state of three databases (Minisat_db 3.0 [24], the Genotyping page version 2.0 [25] and CRISPRdb 1.0 [13]) and some web-accessible tools (CRISPRFinder [26], Spacers Dictionary Creator…). The databases have been developed using MYSQL 4.1 (http://www.mysql.com). The administration and querying process use PHP (http://www.php.net) and Perl CGI scripts.

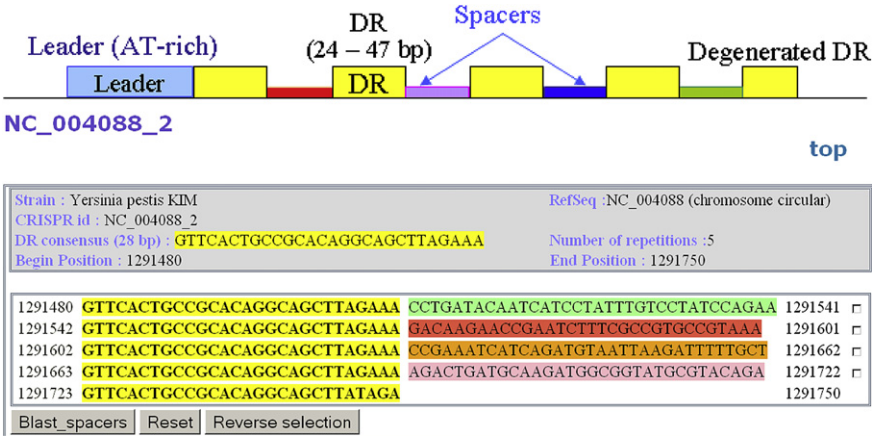

Fig. 1. Schematic representation of a CRISPR locus and output of the CRISPR database for one of the three *Y. pestis* CRISPR loci [12].