Tuberculosis 95 (2015) S159–S166

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

Clustured regularly interspersed short palindromic repeats (CRISPR) genetic diversity studies as a mean to reconstruct the evolution of the *Mycobacterium tuberculosis* complex

Christophe Sola*

Institut de Biologie Intégrative de la Cellule (I2BC), CEA, CNRS, Université Paris-Saclay, Orsay, France

Keywords:
Molecular evolution
CRISPR
Tuberculosis
Spoligotyping
Genomics

SUMMARY

The natural history of tuberculosis may be tackled by various means, among which the record of *molecular scars* that have been registered by the *Mycobacterium tuberculosis* complex (MTBC) genomes transmitted from patient to patient for tens of thousands years and possibly more. Recently discovered polymorphic loci, the CRISPR sequences, are indirect witnesses of the historical phage-bacteria struggle, and may be related to the time when the ancestor of today's tubercle bacilli were environmental bacteria, i.e. before becoming intracellular parasites. In this article, we present what are CRISPRs and try to summarize almost 20 years of research results obtained using the genetic diversity of the CRISPR loci in MTBC as a perspective for studying new models. We show that the study of the diversity of CRISPR sequences, thanks to «*spoligotyping*», has played a great role in our global understanding of the population structure of MTBC.

1. Introduction

The history of infectious diseases is by nature a dynamic process. What kind of infectious diseases did our ancestors suffer from, compared to today's populations [32]? When did commensalism/mutualism between host and bacteria started and what makes a pathogen agent successful? Is *Mycobacterium tuberculosis* only 15,000 years old, 40,000–70,000 years or 3 million years old? Can we superimpose human and bacterial migrations if we study the right model? Were the food regimens responsible for «ancestral tuberculosis» infection in early times? What was the role of phages during the evolution of MTBC? So many questions and so few answers.

The history of tuberculosis is also interesting in that a social network of between 180 and 400 persons was required for host-pathogen coexistence maintenance (i.e., for endemicity to be transmitted vertically similar to human genetic disease). Hence, one way to study this co-evolution is to model retrospectively in collaboration with paleodemographers and human geneticists how many isolated human communities (of what effective size?) could have (i) existed, (ii) survived, (iii) admixed and expanded today according to different hypotheses bearing on 70,000–1 million years of human kind evolution. As we see, in any case, the study of tuberculosis evolution can be tackled from many angles, anthropological, demographical, ecological, genetic, microbiological, epidemiological, and also from the perspective of public health, since too many people in Asia, in Africa and elsewhere still suffer from a disease that is a mirror of all our difficulties to create a fair, sustainable and equitable world for all.

2. An history of discovery, and current CRISPR research

Repeated DNA was discovered in the 60s in higher organisms and later in bacterial genomes. Such sequences are also related to dynamic evolution and lateral gene transfer of genomes. The term “junk DNA” was often used to describe repeated sequences, which were often non-coding, repetitive, associated with peculiar characteristics (micro, minisatellites, insertion sequences) and hard to reconcile with any physiological function. However, from its invention the very concept of “junk DNA” was debated between evolutionists. The first CRISPR to have been described was found upstream the isoenzyme of the alkaline phosphatase gene (*iap*) in *Escherichia coli* [37]. A similar region within a Bacille Calmette Guérin (BCG) strain was identified [35]. In 1997, the structural study

* Bât. 400, 91405 Orsay-Cedex, France. Tel.: +33 169154648; fax: +33 169157296.
E-mail address: christophe.sola@u-psud.fr.

http://dx.doi.org/10.1016/j.tube.2015.02.029
1472-9792/

of the CRISPR locus of MTBC -designated as "the Direct Repeat locus"- was made possible with the invention of a very innovative macro-array typing technique *Spoligotyping* [39].

According to PubMed, CRISPR bibliographical references increases exponentially (doubling every 20 months from 2007 to 2012). The name CRISPR is attributed to Jensen et al., who generalized the *in silico* observation of Repetitive DNA sequences involving a Direct Repeat (21–47 bp) interspersed by unique sequences, the spacers (26–72 bp) [38]. CRISPRs also possess a leader sequence and a family of CRISPR-associated genes or *cas* coding for enzymes active on DNA/RNA metabolism. CRISPR/*cas* systems are involved in the adaptative immune response against phages through an «*arms race*», and may serve as epidemiological markers and since two years, as genome engineering tools. Long before the immunological role of CRISPR was discovered, differences in the number of these palindromic repeats had been used as markers for studies of the molecular epidemiology of pathogenic bacteria, including *M. tuberculosis* [10].

In 2005, Mojica and colleagues suggested that the spacers are derived from foreign genetic elements [44]. Using *Streptococcus thermophilus*, Bolotin et al. demonstrated that CRISPR spacers have foreign phage or plasmid origins; they also suggested a negative correlation between the number of spacers and the sensitivity to bacteriophages [7]. Using *Yersinia pestis*, thanks to an outbreak collection obtained in Vietnam, Pourcel et al. independently showed that the majority of spacers corresponded to fragments of a prophage and that the insertion of spacers was polarized on the side where the *cas* genes were located [45]. At the same time a theoretical paper of the NCBI[1] suggested that the CRISPR systems were likely to work by an RNA interference mechanism with functional analogies to eukaryotic systems and they suggested a potential mode of action [43]. The experimental proof of the link between CRISPR content and phage susceptibility came from research done in Danisco®, an industrial producer of ferments for the dairy industry. Barrangou et al. confirmed that the content in spacers of the CRISPR could be linked to the sensitivity to phages using various phage challenges. They noticed that, if the bacteria had sequences perfectly identical to the phage, the bacteria was resistant, whereas a single mutation in the spacer allowed the phage to lyse the culture. BIM (Bacteriophages insensitive mutants) are generated by acquiring spacers identical to protospacers motifs in phages. The phages evade by creating SNPs mutants in their protospacers. An «*arms race*» is born [2]. CRISPR sequences are always associated to *cas* proteins. These proteins have either helicases, nucleases or RNA or DNA binding activities. A core of 6 cas proteins was first defined; certain proteins are highly conserved, such as cas1 and cas2 whereas others are very polymorphic and are sometimes missing in some systems. The mode of action of CRISPR/cas involves three steps: adaptation, expression and interference [42], see also Figure 1.

2.1. Phase I: adaptation

In type I and type II CRISPR systems selection of proto-spacers in invading nucleic acid probably depends on a proto-spacer-adjacent motif (PAM); how the PAM or the proto-spacer is recognized was still unclear in 2012. After the initial recognition step, cas1 and cas2 incorporate the proto-spacers into the CRISPR locus to form spacers. During the expression stage, the CRISPR locus containing the spacers is expressed, producing a long primary CRISPR transcript (the pre-CRISPR-RNA). The CRISPR of *M. tuberculosis* is peculiar in the sense that it is probably non functional anymore since cas1 and cas2 are missing. Consequently it is likely that it cannot incorporate proto-spacers anymore.

2.2. Phase II/expression

The CRISPR-associated complex for antiviral defense (or Cascade) complex binds the pre-CRISPR-RNA, which is then cleaved by the cas6e or cas6f subunits, in subtype I-E or I-F, respectively, resulting in CRISPR-RNAs with a typical eight nucleotide repeat fragment on the five prime-end and the remainder of the repeat fragment, which generally forms a hairpin structure, on the third prime flank. Type II systems use a trans-encoded small RNA (trans-CRISPR-RNA) that pairs with the repeat fragment of the pre-CRISPR-RNA, followed by cleavage within the repeats by the housekeeping RNase III in the presence of cas9 (formerly csn1 or csx12). Subsequent maturation might occur by cleavage at a fixed distance within the spacers, catalyzed by cas9. In type III systems, cas6 is responsible for the processing step, but the CRISPR-RNAs seem to be transferred to a distinct cas complex (called csm in subtype III-A systems and cmr in subtype III-B systems). In subtype III-B systems, the third-prime end of the CRISPR-RNA is trimmed further.

2.3. Phase III/interference

During the interference step, the invading nucleic acid is cleaved. In type I systems, the CRISPR-RNA guides the Cascade complex to targets that contain the complementary DNA, and the cas3 subunit is probably responsible for cleaving the invading DNA. The PAM probably also plays an important part in target recognition in type I systems. In type II and type III systems, no cas3 ortholog is involved. In type II systems, cas9 loaded with CRISPR-RNA targets invading DNA with a trans-acting CRISPR RNA guide in a process that requires the PAM.

The two subtypes of CRISPR–Cas type III systems target either DNA (subtype III-A systems) or RNA (subtype III-B systems). In type III systems, a chromosomal CRISPR locus and an invading DNA fragment are distinguished either by base pairing to the five prime repeat fragment of the mature CRISPR-RNA (resulting in no interference) or by no base pairing (resulting in interference). Filled triangles represent experimentally characterized nucleases, and unfilled triangles represent nucleases that have not yet been identified.

The molecular diversity of these CRISPR loci make them excellent targets to define bacterial strain identity, of an exquisite precision, indirectly allowing clues on the natural history of a bacterial disease and its evolutionary genetics to be drawn. The mechanisms by which CRISPR might gain or loose spacers can be deduced to proceed as follows:

- By addition of spacers with changing phage pools: a new repeat-spacer unit is added in between the leader and the previous unit within the CRISPR, however this CRISPR evolution strongly depends on the *cas* proteins enzymatic activities available in a given organism, in particular cas1 and cas2.
- By deletion of spacers by mechanisms that could promote recombination or deletion between and within spacers. At least three mechanisms could be involved [1]: insertion sequence transposition are known drivers of CRISPR evolution in MTC depending on the activities and number of copies of transposon and insertion sequences [2]; homologous recombination may of course happen since CRISPR are ideal structures for recombination events [3]; slipped-strand mispairing could also happen.

[1] National center for Biotechnology Information.