Infection, Genetics and Evolution 12 (2012) 819–826

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Alignment of multiple complete genomes suggests that gene rearrangements may contribute towards the speciation of Mycobacteria

J.C. Garcia-Betancur [b,1], M. Carmen Menendez [a,1], P. Del Portillo [b], M.J. Garcia [a,*]

[a] Departamento de Medicina Preventiva, Facultad de Medicina, Universidad Autonoma de Madrid, st/Arzobispo Morcillo, 4, 28029 Madrid, Spain
[b] Corporación Corpogen, Carrera 5 No. 66A-34, Bogota, Colombia

ARTICLE INFO

Article history:
Available online 8 October 2011

Keywords:
Genome comparison
Species delineation
DNA–DNA hybridization
Gene rearrangements
Multiple whole-genome alignments
Genomic Signature

ABSTRACT

To more accurately define the taxonomic relationships among species belonging to the genus *Mycobacterium* we have applied and compared three complete genome sequence comparison procedures to existing systems. These included a nucleotide sequence comparison including both coding and no-coding regions of the genome and two genomic-order comparisons using MAUVE and M-GCAT software to provide comparative gene synteny. These methods clearly differentiated a panel of genomes from reference mycobacterial species. Overall, the speciation of bacteria through determination of gene rearrangements were consistent with the gold standard method for species definition in bacteria, DNA–DNA hybridization however within the context of this system, individual components of the *Mycobacterium tuberculosis* complex (MTBC) did not show sufficient diversity to classify them as a separate species. The high number of gene rearrangements observed between the species tested suggests that gene reorganization of the genome represents an important contributor to speciation within the genus *Mycobacterium* and other related genera. The absence of rearrangements amongst MTBC supports their consideration as a single genospecies. Some gene rearrangements provided clear internal synteny between genomes of mycobacterial strains belonging to a same species and we suggest these could be used to classify subspecies.

1. Introduction

Species definition of bacteria was originally based on morphological and physiological traits. However the subsequent development of molecular tools and molecular markers has allowed a more rational and phylogeny-based classification to emerge.

The genus *Mycobacterium* is one of the most numerous genera among Actinobacteria, including more than 140 different species (http://www.bacterio.cict.fr/m/mycobacterium.html). The definition of Mycobacterial species changed significantly as a result of developments in conserved gene sequence alignments such as the 16SrRNA gene (Tortoli, 2003). Unfortunately this gene was found to be too highly conserved amongst some species ($\geqslant$99%), making differentiation by this method of little practical use in those cases. Other conserved genes such as *rpo*B (Adékambi et al., 2003; Adékambi and Drancourt, 2004) and multiple gene -concatenated sequence comparisons were thus developed (Devulder et al., 2005; Mignard and Flandrois, 2008) which successfully allowed differentiation of the majority of the recognized species, with the notable exception of *Mycobacterium tuberculosis*.

1.1. M. tuberculosis complex and species differentiation

The epithet "*Mycobacterium tuberculosis* complex" (MTBC), although not formally accepted as a taxonomic nomenclature, is currently applied to identify a group of bacteria very closely related to *M. tuberculosis*. Classical members of the MTBC are *M. tuberculosis*, *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium microti* and "*Mycobacterium cannetii*", this last considered as a variant of *M. tuberculosis* (Van Soolingen et al., 1997; Wayne, 1982). More recently *Mycobacterium caprae* (Aranaz et al., 2003), *Mycobacterium pinnipedii* (Cousins et al., 2003) and "*Mycobacterium mungi*" (Alexander et al., 2010) have been added.

Remarkably all MTBC have identical nucleotide sequences of the main target conserved genes used for species differentiation in mycobacteria, including 16SrRNA and *rpo*B. Thus, phylogenetic trees derived from sequence comparison of single- (Takewaki et al., 1994; Tortoli, 2003) or concatenated-gene sequences (Devulder

Abbreviations: MTBC, *Mycobacterium tuberculosis* complex; rRNA, ribosomal RNA; *rpo*B, RNA polymerase subunit B gene; DDH, DNA–DNA hybridization; δ^*, delta differences; LCBs, Local Colinear Blocks.

* Corresponding author. Tel.: +34 91 4972753; fax: +34 91 4975353.
E-mail addresses: juan.c.betancur@stud-mail.uni-wuerzburg.de (J.C. Garcia-Betancur), carmen.menendez@uam.es (M.C. Menendez), pdelportillo@corpogen.org (P. Del Portillo), mariaj.garcia@uam.es (M.J. Garcia).

[1] Contributed equally to the paper.

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.09.024

et al., 2005; Mignard and Flandrois, 2008) resolve to a single unique branch. MTBC have thus been considered as a single genospecies, even though they do not conform to the standard taxospecies nomenclature (Rossello-Mora, 2006).

Several studies analyzing the phylogeny and phylogeography of the MTBC (Dos Vultos et al., 2008; Gagneux and Small, 2007; Mostowy et al., 2002) have shown that genome plasticity has resulted in subtle separations among MTBC members (Brosch et al., 2001; Huard et al., 2003). However, the pivotal characteristic that allows the differentiation among the MTBC members is their host -specificity and host-distribution. *M. bovis, M. microti, M. caprae, M. pinnipedii* and "*M. mungi*" infect bovine, mouse, goat, seal and mongoose respectively whilst *M. tuberculosis* remains almost exclusively of human origin. Indeed discrete host-distribution when employed as a single character of differentiation cannot be considered a robust taxonomic marker to differentiate among bacteria. However a previous analysis of the distribution of the genome region RD9 suggested that the various members of the MTBC could represent *host-adapted* ecotypes (Smith et al., 2006).

1.2. DNA–DNA hybridization and bacterial species definition

DNA–DNA hybridization (DDH) has been applied to analyze bacterial genomic relationships since the early seventies (Seidler and Mandel, 1971) and derives the DNA–DNA similarity of two closely related genomes by measuring the degree of stability of hybrid DNA formed when the two sample genome strands are combined (Rossello-Mora, 2006). This procedure is rarely applied nowadays because of problems with reproducibility and workability (Coenye et al., 2005). Nevertheless, when stringent conditions of hybridization are used, DDH is still considered the molecular gold standard in bacterial species definition within a genus. This is because the taxonomic information for comparison of complete nucleotide sequence in a bacterial genome (Goris et al., 2007) is more accurate than sequences from selected gene markers. DDH has thus higher resolution power than 16S rRNA sequencing (Keswani and Whitman, 2001; Stackebrandt and Goebel, 1994; Stackebrandt et al., 2002; Coenye et al., 2005; Rossello-Mora and Amann, 2001) is valid for a majority of bacterial genera (Kusunoki et al., 1991; Leao et al., 2009; Wayne et al., 1996) and has been successfully applied in the description of novel species within the genus *Mycobacterium* (Brown et al., 1999; Domenech et al., 1997; Jiménez et al., 2004; Murcia et al., 2006).

Using DDH, members of the MTBC show a level of similarity consistent with a single species but are clearly distinguishable from other mycobacterial species (Gross and Wayne, 1970; Baess, 1979; Kusunoki et al., 1991) including the closely related *Mycobacterium marinum* (Tonjum et al., 1998).

1.3. Whole-genome sequence comparison

The increasing availability of fully-sequenced genomes has permitted the application of modern computational algorithms to derive highly detailed genomic comparisons and should, in a near future, replace the DDH for species definition of prokaryotes (Goris et al., 2007). Amongst the available systems of whole-genome nucleotide sequences comparisons, the Genomic Signature method has shown a high correlation to DDH values (Coenye et al., 2005). This method determines the relative intragenomic invariance of oligonucleotide composition by measuring the distribution and relative abundance of di- or tetra-nucleotide within any genome sequence (Karlin et al., 1997). Relative abundance values obtained are more similar throughout the genomes from closely related organisms than comparing the distantly related ones.

Another procedure compares the gene location in the genomes or the physical co-localization of genetic loci in the chromosome (gene synteny) within an individual or species. This type of comparison derives degrees of changes in the gene order along the chromosome and is partially independent of the gene sequence (Coenye et al., 2005). Few algorithms can successfully cope with multiple whole-genome alignments, since the size, complexity and non-collinearity, of large-scale nucleotide sequences are often too great to derive accurate alignments. Nevertheless, successful implementations have been achieved in software such as MAUVE (Darling et al., 2004) and M-GCAT (Treangen and Messeguer, 2006).

With the aim of gaining insight into the taxonomic relationships of mycobacteria at the species level, we applied several complete genome comparison procedures to a panel of sequenced mycobacterial species genomes including members of the MTBC.

2. Design and methods

Table 1 shows the bacterial strains used in this study and their genome identification numbers to which three approaches based on whole-genome sequence characterization were applied.

2.1. Genomic Signature comparison

The Genomic Signature procedure determines the relative abundance of di- or tetra-oligonucleotide along the genome sequence, including coding- and non-coding regions, calculated for disjoint contigs covering the entire genome (Karlin et al., 1997). Similarities among genomes are represented as δ^* differences. This method has the following advantages over other procedures in that it uses the

Table 1
Bacterial strains and complete genome accession numbers.

Bacterial species name	Strain	Reference ID
Mycobacterium tuberculosis	H37Rv ^{T}	NC_000962.2
"	H37Ra	NC_009525.1
"	CDC15151	NC_002755.2
"	F11	NC_009565.1
"	KZN 1435	NC_012943.1
M. bovis	AF 2122/97	NC_002945.3
M. bovis BCG	Pasteur 1173P2	NC_008769.1
M. bovis BCG	Tokio 172	NC_012207.1
M. africanum	GM041182	NC_015758.1
"*M. canettii*"	CIPT 140010059	NC_015848.1
M. avium subsp. hominisuis	104	NC_008595.1
M. avium sbsp. paratuberculosis	K10	NC_002944.2
M. marinum	M	NC_010612.1
M. ulcerans	Agy99	NC_008611.1
M. abscessus	ATCC 19977 ^{T}	NC_010397.1
M. smegmatis	mc2	NC_008596.1
M. vanbalenii	PYR-1	NC_008726.1
M. gilvum	PYR-GCK	NC_009338.1
M. leprae	TN	NC_002677.1
M. sp.	JLS	NC_009077.1
M. sp.	MCS	NC_008146.1
M. sp.	KMS	NC_008705.1
Corynebacterium diphteriae	NCTC 19129	NC_002935.2
C. glutamicum	ATCC13032 ^{T}	NC_006958.1
"	R	NC_009342.1
C. aurimucosum	ATCC700975	NC_012590.1
C. efficiens	YS 314 ^{T}	NC_004369.1
C. jeikeium	K 411	NC_007164.1
C. kroppenstedtii	DSM 44385 ^{T}	NC_012704.1
C. pserudotuberculosis	FRC 41	NC_014329.1
C. urealyticum	DSM 7109^T	NC_010545.1
Rhodococcus erythropolis	PR4	NC_012490.1
R. jostii	RHA1	NC_008268.1
R. opacus	B4	NC_012522.1
Streptomyces avermitilis	MA-4680^T	NC_003155.4
S. coelicolor	A3(2)	NC_003888.3
S. griseus subsp. griseus	NBRC 13350	NC_010572.1
S. scabiei	87.22	NC_013929.1

Superscript "T" stands for type strain.