Infection, Genetics and Evolution 18 (2013) 238–246

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

SNP typing reveals similarity in *Mycobacterium tuberculosis* genetic diversity between Portugal and Northeast Brazil

Joao S. Lopes [a,*], Isabel Marques [a], Patricia Soares [a], Hanna Nebenzahl-Guimaraes [a], Joao Costa [a], Anabela Miranda [b], Raquel Duarte [c,d], Adriana Alves [b], Rita Macedo [e], Tonya A. Duarte [f,g], Theolis Barbosa [h], Martha Oliveira [i], Joilda S. Nery [h], Neio Boechat [f], Susan M. Pereira [g], Mauricio L. Barreto [g], Jose Pereira-Leal [a], Maria Gabriela Miranda Gomes [a], Carlos Penha-Goncalves [a]

[a] *Instituto Gulbenkian de Ciencia, 2781-901 Oeiras, Portugal*
[b] *Instituto Nacional de Saude Dr. Ricardo Jorge, 4150-180 Porto, Portugal*
[c] *Centro Hospitalar de Vila Nova de Gaia/Espinho, 4434-502 Vila Nova de Gaia, Portugal*
[d] *Departamento de Epidemiologia Clinica, Medicina Preventiva e Saúde Pública, Faculdade de Medicina da Universidade do Porto, 4200-319 Porto, Portugal*
[e] *Direccao-Geral da Saude, 1049-005 Lisboa, Portugal*
[f] *Instituto de Doenças do Torax, Universidade Federal do Rio de Janeiro, 21.941-913 Rio de Janeiro, Brazil*
[g] *Instituto de Ciencias da Saude, Universidade Federal da Bahia, 40.150-510 Salvador, Brazil*
[h] *Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz, 40.296-710 Salvador, Brazil*
[i] *Centro de Pesquisa em Tuberculose, Universidade Federal do Rio de Janeiro, 21.941-913 Rio de Janeiro, Brazil*

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form 23 April 2013
Accepted 24 April 2013
Available online 3 May 2013

Keywords:
Mycobacterium tuberculosis complex
Portugal
Brazil
SNP-typing
Spoligotyping
Phylogeny

ABSTRACT

Human tuberculosis is an infectious disease caused by bacteria from the *Mycobacterium tuberculosis* complex (MTBC). Although spoligotyping and MIRU-VNTR are standard methodologies in MTBC genetic epidemiology, recent studies suggest that Single Nucleotide Polymorphisms (SNP) are advantageous in phylogenetics and strain group/lineages identification. In this work we use a set of 79 SNPs to characterize 1987 MTBC isolates from Portugal and 141 from Northeast Brazil. All Brazilian samples were further characterized using spolygotyping. Phylogenetic analysis against a reference set revealed that about 95% of the isolates in both populations are singly attributed to bacterial lineage 4. Within this lineage, the most frequent strain groups in both Portugal and Brazil are LAM, followed by Haarlem and X. Contrary to these groups, strain group T showed a very different prevalence between Portugal (10%) and Brazil (1.5%). Spoligotype identification shows about 10% of mis-matches compared to the use of SNPs and a little more than 1% of strains unidentifiability. The mis-matches are observed in the most represented groups of our sample set (i.e., LAM and Haarlem) in almost the same proportion. Besides being more accurate in identifying strain groups/lineages, SNP-typing can also provide phylogenetic relationships between strain groups/lineages and, thus, indicate cases showing phylogenetic incongruence.

Overall, the use of SNP-typing revealed striking similarities between MTBC populations from Portugal and Brazil.

1. Introduction

Human tuberculosis (TB) is an airborne bacterial disease caused by the *Mycobacterium tuberculosis* complex (MTBC). Currently, WHO estimates that one third of the world's population is infected with this pathogen. From these, a minority progresses to disease, accounting for about 10 million new cases and 2 million deaths per year (WHO, 2011). Recent studies suggest that an increase in prevalence of immunosuppressive diseases (e.g. HIV), population ageing and changes in social patterns are leading to increasing rates of disease activation (Lönnroth et al., 2009). Furthermore, drug-resistance acquisition is also a concern, and reports of bacteria resistant to first and second-lines drugs are growing

* Corresponding author. Address: Instituto Gulbenkian de Ciencia, Apartado 14, 2781-901 Oeiras, Portugal. Tel.: +351 214407900; fax: +351 214407970.
E-mail addresses: j.sollari.lopes@gmail.com (J.S. Lopes), imarques@igc.gulbenkian.pt (I. Marques), psoares@igc.gulbenkian.pt (P. Soares), hanna.guimaraes@gmail.com (H. Nebenzahl-Guimaraes), jcosta@igc.gulbenkian.pt (J. Costa), amiranda@ibmc.up.pt (A. Miranda), raquelafduarte@gmail.com (R. Duarte), aalves@ibmc.up.pt (A. Alves), rita.macedo@dgs.pt (R. Macedo), tonya.duarte@gmail.com (T.A. Duarte), theolis@bahia.fiocruz.br (T. Barbosa), martholiveira@yahoo.com.br (M. Oliveira), joilda_nery@yahoo.com.br (J.S. Nery), n_boechat@yahoo.com (N. Boechat), susanmp@ufba.br (S.M. Pereira), mauricio@ufba.br (M.L. Barreto), jleal@igc.gulbenkian.pt (J. Pereira-Leal), ggomes@igc.gulbenkian.pt (M.G.M. Gomes), cpenha@igc.gulbenkian.pt (C. Penha-Goncalves).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.04.028

considerably (Gandhi et al., 2010). Thus, detailed knowledge of MTBC genetic diversity and geographical distribution is becoming of increasing importance.

MTBC genome is characterized by low substitution rates and, consequently, low DNA sequence diversity, while having marked population subdivisions (Hershberg et al., 2008). These pathogens are generally believed to be highly clonal (Achtman, 2008) with rare horizontal gene transfer (Liu et al., 2006; Namouchi et al., 2012), further decreasing the chances of diversity. Several explanations for this were put forward including the isolated life-style inside mammalian cells, the long generation time and the latent stage with little activity (Smith et al., 2003). The lack of genetic diversity in tuberculosis (TB) makes the study of short-term epidemic networks and long-term evolutionary histories particularly difficult with commonly used markers, such as spoligotype patterns (Kamerbeek et al., 1997) and MIRU-VNTR (Supply et al., 2000, 2001). However, these same traits are ideal for phylogenetic studies using vast single nucleotide polymorphism (SNP) data. In fact, the absence of horizontal gene transfer, which can derange phylogenetic trees by attracting far related branches, and the observed slow substitution rates greatly reduce the problem of convergent evolution when constructing phylogenetic trees.

Defining meaningful boundaries between groups in bacteria is complicated, yet this grouping is necessary for strain classification. Various MTBC classification schemes have been proposed in the past, but none reached a clear consensus (Gagneux and Small, 2007). Recently, Comas and co-workers (2009) defined a classification based on whole-genome data that considered the global diversity of MTBC and was phylogenetically robust (Coscolla and Gagneux, 2010). This classification consisted of six main lineages of human-adapted MTBC and one that mostly infects animals. Within the six main lineages, the authors further classified the strains with a second order grouping according to previous spolygotyping classification: lineage 1, 3, 5 and 6 were defined by single comprehensive groups called EAI, CAS, AFRI1 and AFRI2, respectively; lineage 2 was defined by a non-comprehensive group called Beijing; and lineage 4 was composed by 6 groups called Cameroon, Haarlem, LAM, T, Uganda and X.

In this work we used the two-level MTBC classification to identify samples collected from patients from Portugal and Northeast of Brazil. Previous studies on the global population structure of MTBC (Brudey et al., 2006; Gagneux et al., 2006) observed that the most frequent strain in Europe and South America is lineage 4 (comprised mostly by LAM, Haarlem, T and X). Lineage 1 is also found in both regions, although in much lower frequency. Lineage 2, on the other hand, is typically absent from South America [but see (Iwamoto et al., 2012)]. There has also been previous local-scale studies examining MTBC diversity in Portugal (David et al., 2007) and Southern regions of Brazil [Rio Grande do Sul (Borsuk et al., 2005; Scholante Silva et al., 2009), Parana (Malaghini et al., 2009) and São Paulo (Mendes et al., 2011)]. These, however, have been performed using typically less than 100 samples and genotyped only by spoligotypes. In this paper we present and extensive study using more than 2000 samples from Portugal and Northeast of Brazil genotyped using SNP-typing methods. The comparison between TB populations from these two regions can be of great importance given their possible recent shared ancestry. A demographic study on understanding major past population demographic dynamics, such as admixture, ancient population splitting or migratory trends between TB populations, is out of the scope of this work. Nevertheless, a description of the genetic diversity of the two populations may help to shed some light on the question whether one population results from a recent direct expansion of the other or if they have been evolving separately long before the large human population influx between the two continents of the last five to six centuries.

We present a novel methodology to identify MTBC samples using a reference set composed by previously studied MTBC strains, which, as far as we know, are representative of this group's global diversity. This identification is performed via SNP-based phylogenetic trees, using information on monophyletic groups and their ancestry. The construction of these phylogenetic trees allowed us to further characterize the SNPs in respect to their usefulness in identifying MTBC samples. The goal of this characterization was two-folded: to obtain information on these SNPs for future SNP-typing studies; and further exploit strain ancestry and phylogenetic incongruence in our datasets. In addition to the SNP-based classification, we also analyzed the spoligotype patterns of the Brazilian samples and compared their use in MTBC identification in terms of consistency between markers and successfulness of identification.

2. Material and methods

2.1. Sample collections and molecular typing

The dataset from Portugal consisted of 2112 MTBC samples collected between 2001 and 2011 from patients diagnosed with TB in public hospitals in four major Portuguese regions (North, Center, Lisbon and Tagus Valley and South). The dataset from Northeast Brazil consisted of 147 MTBC samples collected between 2008 and 2009 from patients diagnosed with TB in a reference hospital in Salvador, Bahia. The datasets used in this study consisted solely on sequence data and no personal data was disclosed at any point, thus, there was no need to obtain ethical approval for the analysis presented here.

Genotyping was performed from 2009 to 2012 and the SNPs used were selected from a pool of polymorphisms described until then (Dos Vultos et al., 2008; Filliol et al., 2006; Hershberg et al., 2008; Kasai and Ezaki, 2000). From this pool, 80 SNPs located outside genome regions known to be related to resistance to antibiotics were chosen for phylogenetic analyses (see Table S1 for details). SNP genotyping was performed using primer extension chemistry and mass spectrometric analysis on a Sequenom MassArray platform (Gabriel et al., 2009). The genomic sequence was amplified by multiplex polymerase chain reaction (PCR) and amplified product was treated with shrimp alkaline phosphatase and used for allele specific primer extension reaction according to the MassEXTEND protocol. The reaction mixture was then spotted onto a SpectroCHIP microarray and subjected to the MALDI-TOF mass spectrometry. The genotype calls were assigned using SpectroTYPER software from the SNP-specific peaks. Quality control of the genotyping process used *Mycobacterium tuberculosis* strains EAS054, H37Rv, Haarlem, F11, C and CDC1551, which have curated and publicly available genomes.

Microbead-based spoligotyping was performed according to Cowan et al. (2004). In brief, the direct repeat (DR) region was amplified by PCR using previously described primers (Kamerbeek et al., 1997). The amplified DNA was then incubated with a mixture containing microspheres coupled to a set of 43 oligonucleotide probes (corresponding to the spacer sequences of the DR locus). The hybridization of the PCR product to each specific spacer sequence was quantified using a solid phase fluorometer (Luminex, Austin, TX, USA). A spacer was considered to be present in the genome of a given isolate when the ratio between the average median number of relative fluorescence units (MRFU) in the isolate and the MRFU of the negative control (distilled water) exceeded 5.0.

2.2. Construction of a reference set

In order to use SNP data for identification of the collected samples we constructed a reference set by selecting 31 bacterial strains