Infection, Genetics and Evolution 12 (2012) 811–818

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in *Mycobacterium tuberculosis*

Josephine F. Reyes [1], Carmen H.S. Chan, Mark M. Tanaka *

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia
Evolution & Ecology Research Centre, University of New South Wales, Sydney, NSW, Australia

ARTICLE INFO

Article history:
Received 21 December 2010
Received in revised form 21 May 2011
Accepted 22 May 2011
Available online 30 May 2011

Keywords:
Tuberculosis
Homoplasy
Spoligotyping
DR locus
MIRU
Minisatellites

ABSTRACT

Homoplasy is the occurrence of genotypes that are identical by state but not by descent. It arises through a number of means including convergent and reverse evolution, and horizontal gene transfer. When using molecular markers that are based on sequences possessing a finite number of character states, such as VNTR or spoligotypes, this is an unavoidable phenomenon. Here we discuss the extent of homoplasy and its impact on inferences drawn from spoligotypes and VNTR in epidemiological studies of tuberculosis. To further explore this problem, we developed a computer simulation model combining the processes of mutation and transmission. Our results show that while the extent of homoplasy is not negligible, its effect on the proportion of isolates clustered ("$n-1$ method") is likely to be relatively low for spoligotyping. For VNTR-typing, homoplasy occurs at a low rate provided the number of loci used is high and the mutation rate is relatively high. However, deep phylogenetic inferences using spoligotypes or VNTRs with a small number of loci are likely to be unreliable.

1. Introduction

Strain differentiation of monomorphic bacterial pathogens largely relies on molecular markers based on repetitive DNA, rather than sequence data. For *Mycobacterium tuberculosis*, a common genotyping method is restriction fragment length polymorphism (RFLP) using the transposable element IS*6110*. However IS*6110* RFLP typing is by comparison slow and labour-intensive, and requires subjective analysis to determine the number of bands and gel mobility differences (Blackwood et al., 2004). This somewhat limits the utility of this method for rapid response to outbreak and contact tracing. PCR-based methods such as spoligotyping and variable tandem repeat (VNTR) typing conveniently complement IS*6110*-typing. Spoligotypes are based on the direct repeat sequences in the CRISPR locus which is composed of numerous identical 36-bp direct repeats, interspersed by non-repetitive short sequences or direct variable repeats (DVR) called spacers. VNTR typing is based on short-sequence repeats that are located throughout bacterial genomes.

Both of these methods allow representation by simple sequences that can easily be stored in databases and compared between laboratories. Spoligotypes consist of binary characters representing the presence or absence of variable spacer sequences in the direct repeat region; a VNTR profile is a sequence of numbers representing the number of repeats at specific VNTR loci. However, due to this discreteness, spoligotype and VNTR profiles are restricted to a relatively small space of possible patterns, making it possible for a particular pattern to evolve independently in different lineages. This phenomenon, known as homoplasy, has the potential to undermine analyses of genotype samples that rely on marker-evolution being a divergent process. In contrast, large sequence polymorphisms (Mostowy et al., 2002) and multilocus DNA sequences (Hershberg et al., 2008), used to study tuberculosis, should exhibit little homoplasy.

Homoplasy has been identified as an issue affecting inference using microsatellites in eukaryotes (Anmarkrud et al., 2008; Garza et al., 1996; Navascues et al., 2005). Estoup et al. (2002) studied homoplasy occurring in microsatellites of sexual diploids using mathematical and computational models. They showed that the probability of homoplasy is higher under larger effective population sizes, higher mutation rates and longer divergence times between subpopulations.

From a phylogenetic perspective, because trees represent only divergent evolution, the occurrence of homoplasy interferes with reconstructing topologies accurately. This interference can be quantified. For example, the consistency index (CI) (Givnish and Sytsma, 1997) is the ratio of the minimum number of changes possible to the minimum number of changes observed in the tree, and the homoplasy excess ratio (HER) (Archie, 1989) is a similar measure taking

* Corresponding authors.
E-mail address: m.tanaka@unsw.edu.au (M.M. Tanaka).
[1] Present address: The Kirby Institute, University of New South Wales, Sydney, NSW, Australia.

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.05.018

into account the expected increase in overall homoplasy levels with increasing numbers of taxa. These indices are used as goodness of fit measures for phylogenies; higher values indicate that the presence of homoplasy is more likely to have caused confounding effects, thus rendering a phylogeny less reliable. When the major source of homoplasy is recombination, using information about phylogenetic inconsistencies is a useful way to detect recombination (Maynard Smith and Smith, 1998). In the case of *M. tuberculosis*, recombination occurs at a negligibly low rate (Cole et al., 1998; Liu et al., 2006; Smith et al., 2006). Phylogenetic inconsistencies here are therefore likely due to convergent or reverse evolution. Comas et al. (2009) found that phylogenies constructed using spoligotypes and VNTR loci contained inconsistencies compared to trees built from multilocus DNA sequences. To infer deep phylogenetic relationships using VNTRs, a large number of loci, namely 24, should be used to minimise error from homoplasy (Comas et al., 2009).

In the context of molecular epidemiology, a statistic measuring variation in a sample, the Hunter–Gaston index (HGI), has been used to compare the discriminatory power of different genotyping methods (Hunter et al., 1988). Molecular samples with lower values of HGI may indicate higher levels of homoplasy if identical genotypes of distinct origin are grouped together. However, low values of HGI may also be due to low mutation rates, which are expected to lead to lower homoplasy levels (Estoup et al., 2002). There is a need to clarify the relationship between the evolution of molecular markers and the configuration of genetic variation observed in samples. In particular, how do homoplasy events generated by marker evolution affect statistics used to characterise the extent of transmission? To study ongoing transmission of tuberculosis, molecular epidemiologists measure the proportion of cases clustered in the sense of having identical molecular fingerprints. This method assumes that active cases are epidemiologically related if their isolates have the same fingerprint. Ongoing transmission of active tuberculosis is contrasted here with latent infection. A reduction in the proportion clustered in an area over time is considered to be a sign of improved TB control (Houben and Glynn, 2009).

Here, we consider homoplasy in tuberculosis spoligotypes and VNTR types using a model of evolution of these molecular markers within a model of tuberculosis infection. By exploring how marker evolution influences population and sample statistics we clarify the impact of homoplasy in the molecular epidemiology of tuberculosis.

2. Methods

We have developed a computer simulation model of the mutation process of spoligotypes and VNTR types, nested within an epidemiological model. Throughout the course of the simulation, a population statistic reflecting the amount of homoplasy was computed, and at fixed time points samples were taken to assess the effect of homoplasy on data analysis.

2.1. Disease transmission

We first describe the stochastic model of disease transmission that we use in conjunction with models of the mutation processes, described below. For simplicity we assume a constant closed population size N. Let the total number of infectious individuals be X, and $X = \sum_{i=1}^{G} x_i$, where G is the number of genotypes in the population and x_i is the number of infectious individuals with bacteria of genotype i. The number of susceptible individuals is given by $S = N - X$. Let β be the transmission rate and δ be the rate of death or recovery from the disease, both per capita per year. In the deterministic analogue of our model, X is described by the differential equation $dX/dt = \beta XS/N - \delta X$. The number of new infections produced by a single infectious case in a completely susceptible population (the basic reproductive number) in this model is $R_0 = \beta/\delta$, and the non-zero equilibrium number of infectious cases is $N(1 - 1/R_0)$.

We simulate this stochastic model using the exact Gillespie algorithm (Gillespie, 1977). For our model, the time (in years) between events is an exponentially distributed random variable with parameter λ, given by

$$\lambda = \frac{\beta XS}{N} + \delta X + \sum_{i=1}^{G} \mu_i x_i, \tag{1}$$

where μ_i is the mutation rate of genotype i. The probabilities of birth, death and mutation, conditional on the occurrence of an event, are given by the transition rates (provided in Table 1) divided by λ. We set $\beta = 0.8$ and $\delta = 0.2$ so that $R_0 = 4$ and set $N = 5000$ (Luciani et al., 2009; Tanaka et al., 2006). The population is initialised with a single infectious case representing a bottleneck followed by an epidemic which then leads to endemic disease.

2.2. Evolution of spoligotypes and VNTR loci

The mutation model is embedded within the transmission model described above. We define mutation in this context as the replacement of a strain by a different one in an infected individual due to fixation. A mutation event can be classified as divergent, parallel, convergent or reverse. *Divergent* mutation (Fig. 1A) results in the creation of a new genotype, distinct from all other existing genotypes. A mutation event is also considered to be divergent if it results in the "resurrection" of genotypes that previously became extinct, as illustrated in Fig. 1B–D. We say that a mutation of a genotype is *convergent* when it results in a pre-existing genotype that arose from a different parent (Fig. 1F). On the other hand, we reserve the term *parallel* for a mutation that gives rise to a pre-existing genotype from the same parent, resulting in patterns that are identical by descent and by a recent mutation event (Fig. 1E). "Parallel" and "convergent" are sometimes used interchangeably to refer to the independent evolution of identical states. However, although parallel mutation involves two separate mutation events, genotypes arising from two mutations are also identical by descent from the same ancestral type. As they do not interfere with inferences on relationships among genotypes, for the purposes of this study, parallel mutations are not considered to be instances of homoplasy. Finally, *reverse* mutation occurs when the mutation of a genotype leads it back to its ancestral state (Fig. 1G). In the deletion model for spoligotypes, this is assumed not to occur (Warren et al., 2002). However, reverse mutation is likely to occur in VNTR loci which can be described by a stepwise model (Kimura et al., 1978). The stepwise mutation model has previously been applied to studying VNTRs in *M. tuberculosis* (Grant et al., 2008; Wirth et al., 2008). The mutation processes for spoligotypes and VNTR loci are modelled separately, taking into account their biology and mechanism of evolution. Fig. 2 summarises the procedure for classifying new mutants.

Table 1
Transition rates for genotype i in the stochastic frequency-dependent infection model.

Event	Transition	Rate
Infection	$x_i \to x_i + 1$	$\beta S x_i/N$
Death or recovery	$x_i \to x_i - 1$	δx_i
Mutation	$\left\{\begin{array}{l} x_i \to x_i - 1, \\ G \to G + 1, \\ x_G \to 1 \end{array}\right\}$	$\mu_i x_i$