Infection, Genetics and Evolution 9 (2009) 1010–1019

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Discussion

Phylogenetic understanding of clonal populations in an era of whole genome sequencing

Talima Pearson [a], Richard T. Okinaka [a], Jeffrey T. Foster [a], Paul Keim [a,b,*]

[a] Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ, USA
[b] Pathogen Genomics Division, Translational Genomics Research Institute, Phoenix, AZ, USA

ARTICLE INFO

Article history:
Received 22 December 2008
Received in revised form 15 May 2009
Accepted 20 May 2009
Available online 27 May 2009

Keywords:
Phylogeny
Discovery bias
Clonal populations
B. anthracis
Whole genome sequence

ABSTRACT

Phylogenetic hypotheses using whole genome sequences have the potential for unprecedented accuracy, yet a failure to understand issues associated with discovery bias, character sampling, and strain sampling can lead to highly erroneous conclusions. For microbial pathogens, phylogenies derived from whole genome sequences are becoming more common, as large numbers of characters distributed across entire genomes can yield extremely accurate phylogenies, particularly for strictly clonal populations. The availability of whole genomes is increasing as new sequencing technologies reduce the cost and time required for genome sequencing. Until entire sample collections can be fully sequenced, harnessing the phylogenetic power from whole genome sequences in more than a small subset of fully sequenced strains requires the integration of whole genome and partial genome genotyping data. Such integration involves discovering evolutionarily stable polymorphic characters by whole genome comparisons, then determining allelic states across a wide panel of isolates using high-throughput genotyping technologies. Here, we demonstrate how such an approach using single nucleotide polymorphisms (SNPs) yields highly accurate, but biased phylogenetic reconstructions and how the accuracy of the resulting tree is compromised by incomplete taxon and character sampling. Despite recent phylogenetic work detailing the strengths and biases of integrating whole genome and partial genome genotype data, these issues are relatively new and remain poorly understood by many researchers. Here, we revisit these biases and provide strategies for maximizing phylogenetic accuracy. Although we write this review with bacterial pathogens in mind, these concepts apply to any clonally reproducing population or indeed to any evolutionarily stable marker that is inherited in a strictly clonal manner. Understanding the ways in which current and emerging technologies can be used to maximize phylogenetic knowledge is advantageous only with a complete understanding of the strengths and weaknesses of these methods.

1. Introduction

Reconstructing the patterns of descent for a group of organisms can yield important insights into why and how members of that group have specific characteristics and how those organisms are distributed across the environment. As most characters are inherited in a vertical manner from parent to offspring, depicting patterns of descent in the form of a phylogenetic tree can serve as a map for character acquisition and loss as well as spatial dispersion of the organism. Therefore, phylogenies provide the ability to predict phenotypic and genotypic traits, allowing for a better understanding of biotic and abiotic factors influencing the ecology and distribution of organisms, and can yield more efficiently designed diagnostic assays and vaccines. Integrating population patterns with phylogenetic knowledge provides insights into epidemiological tracking of an organism at different evolutionary scales, from within a single patient (Smith et al., 2006) to across the globe (Li et al., 2007; Van Ert et al., 2007; Holt et al., 2008; Nubel et al., 2008; Moodley et al., 2009). The strength of any phylogenetic inference, however is highly reliant on the accuracy of the tree from which conclusions are drawn.

Since the conception of phylogenetic trees (Darwin, 1859), morphological comparisons have been used to determine patterns of descent. Molecular based methods using allozymes and ultimately DNA have now largely replaced morphological comparisons for phylogenetic inferences, allowing hundreds or even thousands of characters to be compared across samples. Historically, numerous DNA based approaches have been taken to discriminate, sub-type, and build phylogenies for groups of organisms. The success of these methods is dependent on selecting

* Corresponding author at: Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ, USA. Tel.: +1 928 523 1078; fax: +1 928 523 4015.
E-mail address: paul.keim@nau.edu (P. Keim).

1567-1348/$ – see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.meegid.2009.05.014

appropriate loci for different levels of evolution (Keim et al., 2004) as some loci are very conserved and may not discriminate among samples while others are highly variable and may be mutationally saturated, providing misleading phylogenetic information. For microbial pathogens phylogenetic analyses are often conducted in order to determine whether one particular outbreak may be related to another during times of an epidemic. While the clonal nature of an outbreak could be readily measured and predicted, Maynard Smith et al. (1993) pointed to the potential importance of homologous recombination as a determinant in the overall population structure of many bacterial species. These notions are now supported by several typing methods including multi-locus sequence typing (MLST), a relatively standardized approach examining sequences from 7 to 8 ‘‘housekeeping genes’’, that has been used to evaluate at least 52 different microorganisms. Publically available databases (see, for example, http://pubmlst.org and http://www.mlst.net) provide examples where clinical sub-typing has allowed epidemiological, geographical and/or evolutionary hypotheses to be established within pathogens like *Streptococcus pneumonia*, *Neisseria meningitides*, *Neisseria gonorrhoeae*, *Campylobacter*, *Borrelia*, *Vibrios* and *Staphylococcus aureus* (Maiden et al., 1998; Enright et al., 2000; Brueggemann and Spratt, 2003; Brehony et al., 2007; Saunders and Holmes, 2007; Sawabe et al., 2007; Maiden, 2008; Margos et al., 2008; Wilson et al., 2008; Sheppard et al., 2009). Interestingly, the general *B. cereus* MLST scheme is one of the few that displays a relatively conserved clonal population structure (Helgason et al., 2004; Priest et al., 2004). Many of the high-profile, genetically monomorphic pathogens with apparent clonal population structures are either recently evolved or have recently experienced a bottleneck (Achtman, 2008). For these organisms, the selected housekeeping genes often do not have sufficient numbers of SNPs to provide significant resolution, e.g., *Yersinia pestis* (Achtman et al., 1999), *Bacillus anthracis* (Helgason et al., 2004; Priest et al., 2004), *Burkholderia mallei* (Godoy et al., 2003).

In populations where enough polymorphisms can be found, MLST and similar typing methods (e.g., multi-locus VNTR analysis or MLVA) often demonstrate examples of phylogenetic incongruence as a result of convergent evolution and/or lateral gene transfer in population structures. Even the relatively clonal *B. cereus* sub-group population shows evidence for a limited amount of homologous recombination and/or homoplasy (Didelot et al., 2009). MLST and similar methods are well suited to examining specific outbreaks and populations for many diseases and phylogenetic inference could be used to accurately portray the clonal expansion of a specific outbreak. But when an accurate overall phylogeny and evolutionary tree is needed for a particular species and its relatives these methods are often confounded by homoplasy, homologous recombination and lateral gene transfer (Achtman and Wagner, 2008) and would likely benefit from the inclusion of more data (Turner et al., 2007).

1.1. Importance of phylogenetic accuracy

An accurate phylogeny contains patterns of relatedness vis-à-vis how samples are related to each other and indicates the degree of divergence between samples. The former is determined by branching patterns, while the latter is dependent on branch lengths. Accurate branching patterns are important for defining an order of relatedness. The hypothetical most recent common ancestor between two samples lies at the bifurcation point for the two samples. The closer this ancestor lies to the terminal ends of the branches, the more closely related the samples. Any bifurcation point that is closer to either of the samples is indicative of an even more closely related lineage. Accurate branch lengths indicate the amount of divergence along a lineage; the actual number and characteristics of mutations can be determined, and if the mutation rate and generation time is known, the time between bifurcation points can be estimated (Zuckerkandl and Pauling, 1965). Thus the radiation of a group over space and time can be determined by phylogenetic analyses. Highly accurate phylogenies will lead to more informed conclusions at all evolutionary levels.

1.2. Homoplasy, confidence and accuracy

Almost all phylogenetic data sets contain significant amounts of homoplasy (character state similarity due to independent evolution), complicating the ability to trace patterns of descent. Homoplasy occurs when a character mutates to an ancestral form (reversal) or to a form found in another lineage (convergence or parallelism). Recombination among lineages (lateral gene transfer) can also occur and is common in many bacterial species, causing different regions of the genome to have different evolutionary histories. As long as trees are drawn by selecting large numbers of characters that are distributed across the genome, the influence of recombined single genomic regions in dictating tree topology will be diminished, resulting in a tree that reflects the evolutionary history of the majority of the genome. Many phylogenetic methods have been developed specifically to deal with evolutionary reversals, convergences and parallelisms. Under the premise that the simplest hypotheses are preferable, maximum parsimony methods estimate the evolutionary history while invoking a minimum number of mutational steps. Character state conflict (homoplasy) is incorporated into the resulting trees by adding extra steps. Often, but not always, there is more than one equally parsimonious way of adding these extra steps, resulting in different branching patterns and multiple trees. Evolutionary models that incorporate rates and patterns of mutation are used by maximum likelihood (ML) phylogenetic methods to calculate the probability that a proposed hypothesis gave rise to the observed data. The manner in which homoplasies are incorporated into ML trees is dependent on the selected model of evolution but, as with parsimony based methods, homoplastic characters can lead to multiple similarly likely trees which differ in the way that samples are assigned to groups.

The most popular statistical assessment of confidence for evaluating group membership in a phylogenetic tree is through bootstrap analyses (Felsenstein, 1985). Interestingly, high levels of confidence can be gained in trees even with high levels of homoplasy (Sanderson and Donoghue, 1989). Accuracy, on the other hand is how well a tree resembles the ‘‘true’’ phylogeny and can only be directly measured in simulation studies where the ‘‘true’’ tree is known. Such studies show that high levels of phylogenetic accuracy can only be achieved as homoplasy levels approach zero, enabling homoplasy to be used as a direct indicator of accuracy (Archie, 1996). While there are many indices of homoplasy, they all will be similar in value when homoplasy is low (Archie, 1996). In data sets with no homoplasy, we can have complete confidence in a group that is supported by only a single character, however bootstrap analyses will underestimate support for such a group, requiring three characters to be part of the 95% confidence interval and 6 for 100% confidence (Felsenstein, 1985).

Molecular data that are free from homoplasy must measure allelic states for characters that are inherited in a strictly clonal manner to avoid the confounding effects of lateral gene transfer, and must be evolutionarily stable to reduce the likelihood of mutational reversals or convergence. Markers that are more evolutionarily stable (mutate relatively infrequently) provide less resolution among closely related samples than markers that change quickly. However, slowly evolving markers are preferable for determining deeper levels of relatedness, because they are less prone to evolutionary reversals or convergent evolution that can obscure patterns of descent (Keim et al., 2004). Single nucleotide