Molecular Phylogenetics and Evolution 107 (2017) 246–255

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

The underlying mechanisms of genetic innovation and speciation in the family *Corynebacteriaceae*: A phylogenomics approach

Xiao-Yang Zhi [a,b,*], Zhao Jiang [a], Ling-Ling Yang [a], Ying Huang [b]

[a] Yunnan Institute of Microbiology, School of Life Sciences, Yunnan University, Kunming 650091, China
[b] State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Article history:
Received 26 July 2016
Revised 4 November 2016
Accepted 14 November 2016
Available online 15 November 2016

Keywords:
Corynebacteriaceae
Genetic innovation
Speciation
Horizontal gene transfer
Gene loss
Phylogenomics

ABSTRACT

The pangenome of a bacterial species population is formed by genetic reduction and genetic expansion over the long course of evolution. Gene loss is a pervasive source of genetic reduction, and (exogenous and endogenous) gene gain is the main driver of genetic expansion. To understand the genetic innovation and speciation of the family *Corynebacteriaceae*, which cause a wide range of serious infections in humans and animals, we analyzed the pangenome of this family, and reconstructed its phylogeny using a phylogenomics approach. Genetic variations have occurred throughout the whole evolutionary history of the *Corynebacteriaceae*. Gene loss has been the primary force causing genetic changes, not only in terms of the number of protein families affected, but also because of its continuity on the time series. The variation in metabolism caused by these genetic changes mainly occurred for membrane transporters, two-component systems, and metabolism related to amino acids and carbohydrates. Interestingly, horizontal gene transfer (HGT) not only caused changes related to pathogenicity, but also triggered the acquisition of antimicrobial resistance. The Darwinian theory of evolution did not adequately explain the effects of dispersive HGT and/or gene loss in the evolution of the *Corynebacteriaceae*. These findings provide new insight into the evolution and speciation of *Corynebacteriaceae* and advance our understanding of the genetic innovation in microbial populations.

1. Introduction

The family *Corynebacteriaceae* comprises the genus *Corynebacterium*, with more than 130 species and subspecies (http://www.bacterio.net/), and the monospecific genus *Turicella*, which exhibits differences in the mycolic acid and quinone system, but does not have an independent phylogenetic position (Bernard and Funke, 2012; Busse, 2012). Members of the family *Corynebacteriaceae* cause a wide range of serious infections in humans and animals (including opportunistic infections), such as diphtheria, endocarditis and lymphadenitis. Intriguingly, there are also some non-pathogenic species, e.g. *C. glutamicum*, which is used industrially for the large-scale production of amino acids (Lee et al., 2016). Their importance in public healthy and industry have meant that the genomic sequences of many corynebacterial species have been determined. Based on these genomic data, however, previous studies mainly focused on revealing the pathogenicity and its underlying genetic mechanisms in individual species (Soares et al., 2013; Trost et al., 2012; Meinel et al., 2014). Therefore, little is known about the evolutionary history and speciation of the family *Corynebacteriaceae* from a genomics perspective. In the process of long-term evolution, the main mechanisms that have shaped corynebacterial genetic diversity remain unknown.

The definitions of microbial species have been debated and remain controversial. Microbiologists are struggling to summarize their genetic diversity, to classify them, to determine the mechanisms that lead to speciation and whether microbial species even exist (Achtman and Wagner, 2008). Microorganisms pose a particular challenge because of their genetic diversity, asexual reproduction and promiscuous horizontal gene transfer (HGT). However, combined with their advantages, such as rapid evolution and easily sequenced genomes, microbes offered an unprecedented opportunity to study and understand speciation (Shapiro et al., 2016). Our lack of knowledge about microbial speciation theoretically has led to most current definitions of microbial species still focusing on methodology. Speciation is a process of gradual accumulation of differences amongst the individuals of a population. The differences are reflected not only in the sequences of biological

[*] Corresponding author at: Yunnan Institute of Microbiology, School of Life Sciences, Yunnan University, Kunming 650091, China.
E-mail address: xyzhi@ynu.edu.cn (X.-Y. Zhi).

http://dx.doi.org/10.1016/j.ympev.2016.11.009
1055-7903/© 2016 Elsevier Inc. All rights reserved.

macromolecules, but also in the structures of their genetic material. Therefore, similar to the role of polyphasic identification in microbial taxonomy, efforts to address microbial speciation should integrate multiplex information from the research objects (Zhi et al., 2012).

The major mechanisms leading to genetic innovation are gene gain and loss, gene origination *de novo*, and gene duplication followed by divergence. The increasing wealth of genomic data has provided a new perspective on gene loss as a pervasive source of genetic change that has great potential to cause adaptive phenotypic diversity (Albalat and Cañestro, 2016). Recent exhaustive analyses comparing hundreds of genomes of bacteria and archaea revealed that loss of gene families is also pervasive. In some cases, gene family loss dominated their evolution, with frequencies up to three times higher than the rate of gene gain (Koskiniemi et al., 2012; Puigbo et al., 2014). Even at the species level, gene loss was completely dominant as a source of genetic variation within pathogenic bacterial species, while nonclonal species diversify through a combination of changes to gene sequences, gene loss and gene gain (Bolotin and Hershberg, 2015). Gene loss is the main impetus of genetic reduction. By contrast, gene gain, including exogenous HGT and endogenous gene genesis *de novo*, leads to genetic expansion. The genetic diversity of a population is the balanced result of genetic expansion and reduction. For a bacterial population like the *Corynebacteriaceae*, determining the underlying mechanisms of genetic innovation and their role in the variation of phenotypic characteristics will benefit our understanding of prokaryotic evolution and improve the concept of a species.

Phylogenomics aims to reconstruct the evolutionary history of organisms based on an analysis of their genomes (Delsuc et al., 2005). Along with the dramatic increase in genomic data, many studies have demonstrated the power of this approach. In this study, to understand the mechanisms of speciation of the *Corynebacteriaceae*, we analyzed its pan-genome and reconstructed the phylogenomic tree. The homologs of different protein families that represent the genetic innovation of *Corynebacteriaceae* were analyzed systematically. The results indicated that compared with other mechanisms, gene loss has been the primary force causing genetic changes. Putative HGT has also played an important role in the speciation of the *Corynebacteriaceae*, especially for certain physiological traits, e.g. pathogenicity and antimicrobial resistance. The phylogenomic analysis revealed that dispersive HGT and/or gene loss that caused genetic variations in the middle stage of the evolutionary history of *Corynebacteriaceae* were inconsistent with the molecular phylogeny. These findings indicated that these bacteria underwent a complex evolution and speciation process, which included both vertical and lateral modes.

2. Materials and methods

2.1. Genome sequences

The genomic sequences of 83 *Corynebacterium* species, and *Turicella otitidis* DSM 8821, were retrieved from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/) and Integrated Microbial Genomes server (http://img.jgi.doe.gov/, further referenced as the JGI_IMG database, Markowitz et al., 2014). The detailed information of these 84 genomes is shown in Table S1. The protein coding sequences (CDS) were predicted by Prodigal (Hyatt et al., 2010) using the normal mode, to avoid differences caused by different programs and/or under different parameters. All CDSs shorter than 150 bp (corresponding to a protein of <50 amino acids) and that had >30 ambiguous sites (unknown bases and/or gaps), were removed from the transcriptional sequences dataset, which was translated to protein sequences directly.

2.2. Homologous protein family analysis and function annotation

Homologous protein families (PFs) were determined by the program OrthoMCL (version 2.0, Li et al., 2003). In our analyses, all translated protein sequences were adjusted to a prescribed format and were grouped into homologous clusters using OrthoMCL, based on sequence similarity. The BLAST reciprocal best-hit algorithm (Moreno-Hagelsieb and Latimer, 2008) was employed, and Markov Cluster Algorithms (MCL, Enright et al., 2002) were applied with an inflation index of 1.5.

All PFs were divided into two categories according to their distribution among 84 genomes: core PFs, including paralogs and orthologs, and group-specific (GS) PFs that involved all PFs that only appeared in partial species. The GS PFs were further divided into 10 groups according to the proportion of their distribution in the 84 species. For instance, group GS9 included all PFs that could be found in ⩾90% (76) of the 84 genomes; group GS0 included all PFs that could be found in ⩾one genome, but <9 genomes. There were at least two sequences if the PF that were only distributed in one genome. In addition, in the pan-genome of *Corynebacteriaceae*, there were many genes present as a single copy (singletons). Therefore, the whole pan-genome could be divided into 12 portions: core, 10 GSs, and singletons. In this work, all singleton proteins whose lengths were less than 100 amino acids (aa) were ignored because of their functional uncertainty and the potential inaccuracy of gene predication. The functions of PFs and singletons were determined using BLAST (Camacho et al., 2009) against the clusters of orthologous groups COGs (Galperin et al., 2015) and the Kyoto encyclopedia of genes and genomes (KEGG) (KAAS, Moriya et al., 2007) databases. All PF members were subjected to functional annotation, and the majority of the annotation results were used to represent the functions of the PFs.

2.3. Homologous proteins searching in prokaryotes

In addition to functional annotation, for each PF, all members were used as queries to BLAST search against a local genomes database that included 2756 prokaryotic predicted proteomes, which excluded the proteomes belonging to the family *Corynebacteriaceae*. The thresholds of sequence identity and match coverage were set to 30% and 50%, respectively. For each single protein, the best hit was analyzed based on its taxonomic sources. A majority of taxonomic sources of all best hits was then chosen to represent the closest homologous counterpart of the protein family.

2.4. Phylogenomic analysis

To determine the phylogenetic relationships amongst the species of *Corynebacteriaceae*, based on genomic data, both supermatrix and gene content methods were applied to infer phylogenetic trees. For the supermatrix method, we selected a set of orthologous PFs shared by all 84 genomes. The protein sequences of each orthologous PF were aligned using Clustal W (Larkin et al., 2007), and the resulting alignments of individual proteins were concatenated to infer the organismal phylogeny using the maximum likelihood algorithm (ML) in a message passing interface-parallelized version of RAxML version 7.3.0 (Stamatakis, 2006). The ambiguous alignments were removed by the Gblocks method in the SEAVIEW software (Gouy et al., 2010), with options for a less stringent selection. The LG model (Le and Gascuel, 2008) with a proportion of invariable sites (+I), a gamma-shaped distribution of rates across sites (+G) and observed amino acid frequencies (+F), was used to infer the phylogeny. The topologies of the phylogenetic trees were evaluated using the bootstrap resampling method of Felsenstein (1985) with 100 replicates. The phylogenies of all individual orthologous proteins were also inferred using the