Infection, Genetics and Evolution 21 (2014) 54–57

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Short communication

A lack of genetic basis for biovar differentiation in clinically important *Corynebacterium diphtheriae* from whole genome sequencing

Vartul Sangal [a], Andreas Burkovski [b], Alison C. Hunt [c], Becky Edwards [d], Jochen Blom [e], Paul A. Hoskisson [a,*]

[a] Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow G4 0RE, UK
[b] Lehrstuhl für Mikrobiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[c] Department of Medical Microbiology, Aberdeen Royal Infirmary, Foresterhill, Aberdeen AB25 2ZN, UK
[d] Microbiology Department, Royal Infirmary of Edinburgh, 51 Little France Cresent, Old Dalkeith Road, Edinburgh EH16 4SA, UK
[e] Bioinformatics Resource Facility, Center for Biotechnology (CeBiTec), Bielefeld University, Germany

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form 22 October 2013
Accepted 23 October 2013
Available online 5 November 2013

Keywords:
Corynebacterium diphtheriae
Biovar
Genomic comparison
Gene contents
SNPs
Pseudogenes

ABSTRACT

The differentiation of clinically important *Corynebacterium diphtheriae* into specific biovars is complex and phylogenetically unclear. Comparative genomic analyses of 17 strains indicate that the division of *C. diphtheriae* into different biovars does not correlate with the variation in the gene content in the relevant metabolic categories that are potentially involved in the biovar discrimination. The biochemical separation is also not supported by phylogenetic analyses, suggesting molecular methods of typing *C. diphtheriae* strains should be adopted much more widely.

Corynebacterium diphtheriae, the causative agent of diphtheria, remains a significant cause of global morbidity and mortality (http://www.who.int/immunization_monitoring/diseases/diphteria/) and has traditionally been subdivided into the four biovars gravis, intermedius, mitis and belfanti based on biochemical characteristics (Funke et al., 1997; Goodfellow et al., 2012). This biochemical separation has been prominent in the clinic, yet the differentiation of strains is complex and unreliable. The overall basis is as follows, only intermedius is lipophilic that forms small grey or translucent colonies, whereas other biovars form large white or opaque colonies (Funke et al., 1997). The strains of biovar gravis can ferment dextrin, glycogen and starch while intermedius strains are dextrin positive and may utilise glycogen and starch (Funke et al., 1997; Goodfellow et al., 2012). Biovar mitis strains can rarely use starch but not glycogen or dextrin. Only intermedius strains are non-haemolytic and only belfanti can not reduce nitrate or use glycogen and starch as carbon sources (Table 1; 1, 2). The API Coryne system is largely used for biochemical characterization of *C. diphtheriae* isolates into biovars, yet it cannot discriminate between biovars mitis and intermedius (Funke et al., 1997). Rapid-strip API-Coryne is usually glycogen negative for intermedius, yet this biovar can be positive when tested by the conventional tube-based assay methods (Goodfellow et al., 2012). The quality assurance scheme for diphtheria diagnostics by the European diphtheria surveillance network (EDSN) highlighted the problems with the correct identification of biovars mitis and intermedius by several participating laboratories (Neal and Efstratiou, 2009).

The population structure of *C. diphtheriae* strains has been investigated by MLST studies that showed an overall lack of correlation between the sequence types (STs) and biovars, with genetically diverse isolates falling within the same biovar designation (Bolt et al., 2010; Farfour et al., 2012). There was no correlation between the severity of the disease with the biovar designation (Bolt et al., 2010). In this study, we have sequenced two *C. diphtheriae* genomes and performed comparative analyses with an additional 15 published genome sequences to try to understand the genetic basis of biovar differentiation.

The genomes of two clinical non-toxigenic *C. diphtheriae* (designated biovar gravis) strains 'Aberdeen' (Edwards et al., 2011) and DSM43988, isolated from throat swabs (Table S1), were sequenced using GS-Junior 454 apparatus as previously detailed (Sangal et al., 2012a,b). The reads were assembled using GS *de novo* assembler and the resulting contigs were reordered onto *C. diphtheriae* C7(β) or NCTC 13129 genomes. The genomes were annotated using the Prokaryotic Genomes Automatic Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html). The 'Aberdeen' genome was assembled into 38 contigs with a total assembly

* Corresponding author. Tel.: +44 (0) 141 548 2819; fax: +44 (0) 141 548 4124.
E-mail address: paul.hoskisson@strath.ac.uk (P.A. Hoskisson).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.10.019

Table 1
Biochemical differentiation of *C. diphtheriae* biovars.

Biovar	Lipophilism	Haemolysis[a]	Nitrate reduction	Ability to utilise[b]	
				Starch	Glycogen
Gravis	–	±	+	+	+
Intermedius	+	–	+	±	±
Mitis	–	±	+	±	–
Belfanti	–	?	–	–	–

[a] Biovar gravis is haemolytic but some strains may be weakly haemolytic. Biovar mitis is weakly haemolytic. The haemolytic properties of biovar belfanti are not clear in the literature.
[b] Strains of biovar intermedius may utilise glycogen and starch while mitis strains can rarely use starch but not glycogen.

size of 2,367,453 bp and 2207 protein coding sequences (CDS) (Accession No. AUZO00000000). The genome of DSM43988 was assembled into 140 contigs with 2,141,743 bp assembly size and 2088 CDS (Accession No. AUZN00000000).

C. diphtheriae core and pangenomes were calculated using EDGAR (Blom et al., 2009), including the additional 15 published genomes (Table S1). The core and pan genomes consisted of 1331 and 4918 genes respectively, which does not fit with the previous core genome development predictions (~1611 genes) and the pangenome increment rate (Heaps' law growth exponent alpha = 0.69; Trost et al., 2012). The inclusion of four draft genomes in this analysis is mainly responsible for the differences observed between the two studies. The draft genome of strain DSM43988, with 140 contigs, has contributed to most of the genomic gaps, resulting in the absence of approximately 150 genes when compared to other genomes. The exclusion of this strain increased the core genome to 1544 genes and reduced the pan genome to 4609 genes (data not shown). Most of the gaps are present in the mobile genetic elements including phage associated regions and putative transposases due to repetitive nature of these regions. These differences are also partly contributed by the additional global strain diversity (Table S1) as the previous study was more focused on Brazilian *C. diphtheriae* isolates (9/13 isolates) whereas we included two British isolates from the pre-vaccination era (NCTC03529 and NCTC05011) and two non-toxigenic isolates (one recently isolated in the UK), increasing the overall diversity of the strain collection analysed. A CDS blast map using CGView (Grant et al., 2012) revealed a high degree of synteny between all *C. diphtheriae* genomes (Fig. 1A). All 17 genomes were aligned using Mauve (Darling et al., 2004) and genome wide SNPs were extracted. SNPs in putative transposases and missing alleles were removed from the analysis. SNPs in other mobile genetic elements (putative repeats, insertion sequences and bacteriophage regions) were identified using Tandem Repeat Finder (Benson, 1999), ISfinder (Siguier et al., 2006) and PHAST (Zhou et al., 2011) and were excluded. A maximum parsimony phylogeny from the resulted 88,733 SNPs using MEGA 5.03 was constructed (Tamura et al., 2011; Close Neighbour Interchange method with 100 initial random trees and a MP search level of 10) and distinctly separated most isolates (Fig. 1A). The analysis showed that strain C7(β) was closely related to DSM43988 and HC01 grouped closely with strain 241. A phylogenetic tree constructed from 400 conserved protein sequences (Segata et al., 2013) also showed a lack of biovar specific groupings (Fig. S1). These results are consistent with the phylogenetic relatedness between *C. diphtheriae* strains from the core genome (Sangal et al., 2013) and previous multilocus sequence typing (MLST) studies (Bolt et al., 2010; Farfour et al., 2012), suggesting that biotyping does not necessarily group genetically related isolates together.

Reciprocal blast searches through EDGAR (Blom et al., 2009) were used to identify potential biovar specific genes between four genomes using a single representative of each biovar (NCTC13129, NCTC03529, NCTC05011, and INCA402). A total of 1719 genes were common between these strains and 173–263 genes were found to be strain-specific (Fig. 1B). The majority of differences between different biovars were the mobile genetic elements or genes encoding hypothetical proteins (Table 2). Some of the specific genes belonged to restriction-modification and CRIPSR-Cas systems (Tables 2; S2) and these differences may represent barriers to recombination among different strains (Sangal et al., 2013) but show no correlation with biovar designations.

Genes involved in carbohydrate, lipid, iron and nitrogen metabolisms were extracted from the accessory gene pool (i.e. those not shared by all four representative biovar strains and thus likely to be associated with the biovars discrimination tests) and were searched for homologs in the remaining 13 genomes using EDGAR comparative viewer (Blom et al., 2009). Most of these genes have homologs in other strains except for four genes that are involved in carbohydrate metabolism (DIP0660 – a putative propionyl-CoA carboxylase beta-subunit; DIP1011 – putative aldose 1-epimerase; DIP1302 - putative ribose-5-phosphate isomerase and DIP1639 – putative dihydrolipoamide acetyltransferase) were absent in intermedius and one gene encoding a putative 5,10-methylenetetrahydrofolate reductase (DIP1611) was missing in befanti (Table S2). These genes, based on similarity searches do not appear to have a role in the metabolic pathways associated with the biochemical properties used to distinguish these biovars (Table 1). These data suggest that the differences may be strain-specific rather than biovar-specific. Examining the presence of pseudogenes and miscellaneous features such as potential frameshift mutations in four representative strains again showed no correlation with the biovar groupings (Table S3).

Of the total 88,733 genomic SNPs between all seventeen strains, 5856 SNPs appeared to be biovar-specific: 14 to gravis, one to mitis, 2236 to belfanti and 3604 to intermedius (Table S4). The relatively high numbers of belfanti and intermedius specific SNPs may indicate a sample bias as only one strain has been sequenced that is assigned to these biovars. C7(β) was derived from a mitis strain (Barksdale and Pappenheimer, 1954; Freeman, 1951) and PW8, the vaccine producing strain, was also designated as biovar mitis (Lampidis and Barksdale, 1971). These strains were used for testing whether the SNP pattern can reliably predict biovar for these strains or not. Interstingly, C7(β) is biovar gravis, according to the SNP pattern wheras PW8 showed mosaic SNP patterns and no biovar can be assigned (Table S4). Therefore, these SNPs are not biovar specific. From the three strains with undetermined biovars, 241 could be assigned to biovars mitis and 31A and BH8 showed mosaic SNPs patterns similar to those across multiple biovar designations (Table S4). The properties of these SNPs were investigated using TRAMS (Reumerman et al., 2013) with strain INCA 402 (biovar belfanti) as the reference genome. A total of 1651 SNPs were non-synonymous, including six nonsense mutations. All nonsense SNPs were only present in biovar intermedius (NCTC 05011) and appears to have resulted in the inactivation of the genes encoding an Na^+/H^+ antiporter-like protein (DIP0794), a putative aldose 1-epimerase (DIP1011; potentially involved in starch/glycogen metabolism), anthranilate synthase component I