AMERICAN SOCIETY FOR MICROBIOLOGY Journal of Clinical Microbiology

COMMENTARY

Molecular Assay Validation Using Genomic Sequence Databases

John P. Dekker

Department of Laboratory Medicine, National Institutes of Health Clinical Center, Bethesda, Maryland, USA

Whole-genome sequence databases offer new *in silico* approaches for designing and validating PCR assays in the clinical microbiology laboratory. An article in this issue of the *Journal of Clinical Microbiology* (M. J. Jansen van Rensburg, C. Swift, A. J. Cody, C. Jenkins, and M. C. J. Maiden, J Clin Microbiol, 54:2882–2890, 2016, http://dx.doi.org/10.1128/JCM.01522-16) demonstrates the use of publicly available genomic sequence data to evaluate a PCR assay for distinguishing *Campylobacter* species.

Ornithologists once classified northern flickers (*Colaptes auratus*, or beetle-eating woodpeckers) into two distinct species on the basis of wing shaft color. Yellow-shafted flickers were found in eastern North America, whereas red-shafted flickers were found in the west. Later, it was discovered that the yellow and red varieties interbred extensively over a wide range of territorial overlap, with production of various intergrades (1). In this well-cited example, a binary morphological classifier derived from a population subset appeared to distinguish two species but broke down when the larger population structure was more fully understood.

Though the classification of bacteria into distinct species may rest on less firm conceptual ground than the analogue in the world of sexually reproducing animals (2–5), the problem of optimal target sequence selection for PCR assays designed to distinguish bacterial taxa is fundamentally a population biology problem with analogies to many challenges found in other areas of biologic systematics. Of the large number of published primer sets that have been used by clinical microbiologists to make bacterial species distinctions, many have been designed on the basis of relatively small numbers of representative target sequences, and the targets themselves have often been selected for idiosyncratic or historical reasons, as opposed to an approach that employs a systematic genome-wide search strategy. In the age of Sanger methods, available sequenced targets were comparatively limited, and in general, it was not possible to examine even these limited targets (beyond ribosomal genes) at the population level. Consequently, many PCR assays have been designed without a solid understanding of the underlying wild-type sequence diversity and how the population is partitioned by this diversity.

The availability of a large number of sequenced bacterial genomes has made possible new approaches to the problem of optimal primer design for taxonomic distinctions and offers quantitative answers to questions of how diversity in a set of sequence targets is distributed in a population. Clinical microbiologists, however, have arguably not yet made optimal use of these data, probably for a number of underlying reasons, including lack of familiarity with genome assemblies and the bioinformatics tools required for analysis (6). In work published in this issue of the *Journal of Clinical Microbiology*, Jansen van Rensburg and colleagues demonstrate one approach to species-specific primer validation using whole-genome sequence (WGS) data that may be used as a blueprint by other clinical microbiology labs (7). For this study, the authors focused on the distinction between the related gastrointestinal pathogens *Campylobacter coli* and *Campylobacter jejuni* and evaluated the ability of a previously published primer-probe set targeting the *mapA* gene of *C. jejuni* and the *cueE* gene of *C. coli* to distinguish these species (8). The two genes are present in both species, and the assay works by targeting conserved species-specific sequence differences in these genes. The primers were originally designed to be used as a duplex real-time PCR assay and were extensively tested against clinical isolates, but the creators of the assay did not have the advantage in 2003 of the genomic sequence archives that are available today (8). In the intervening years, this assay and derivatives have been used by multiple labs for routine isolate identification to the species level and epidemiology (9, 10).

Jansen van Rensburg et al. began by selecting >1,700 *Campylobacter* genomes from the public PubMLST database (http://pubmlst.org/campylobacter/). The Bacterial Isolate Genome Sequence database (BIGSdb) software within PubMLST was used to identify the *mapA* and *cueE* gene targets, as well as multilocus sequence type (MLST) and ribosomal MLST (rMLST) loci, in each assembly (12–15). In this type of analysis, appropriately represented diversity in the input sequence collection is a critical parameter in performance, as *in silico* prediction methods work only insofar as the sequences on which they operate are representative of the biological populations under study. The authors approached the quantification of diversity and representation in two ways. First, to measure the overall allelic diversity of the selected set of genomes, they performed calculations using the bias-corrected version of Simpson's diversity index over aggregate MLST and rMLST data (16, 17). The index was calculated as 0.972 to 0.976 (95% confidence interval [CI]) for the MLST data and 0.988 to 0.991 (95% CI) for the rMLST data, indicating a high level of absolute diversity by this measure. Second, to demonstrate that the diversity in their data set was reflective of the actual diversity observed in clinical isolates, they compared the detailed distribution of MLST clonal complexes with that observed for >3,300 clinical isolates recovered from 2003 to 2009 in their geographic region and found that almost all of the clonal clusters in their set

Accepted manuscript posted online 12 October 2016

Citation Dekker JP. 2016. Molecular assay validation using genomic sequence databases. J Clin Microbiol 54:2854–2856. doi:10.1128/JCM.01797-16.

Editor: N. A. Ledeboer, Medical College of Wisconsin

Address correspondence to john.dekker@nih.gov.

For the article discussed, see doi:10.1128/JCM.01522-16.

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.

demonstrated proportions similar to those found in the clinical isolates.

To study the species structure of the genomic data set, the authors examined rMLST data to define ribosomal sequence type (rST) sets (18). The sequences from each unique rST were concatenated, aligned, and used to generate a phylogenetic tree from which species designations were inferred. Consistent with findings of previous *Campylobacter* population studies (19–22), almost all of the isolates fell cleanly into the *C. jejuni* or *C. coli* category by this analysis, and *C. coli* was best described as three clades, a feature that relates to certain results described below. It is important to note the comprehensive approach the authors took to establishing an accurate taxonomic classification of the starting sequences, which is particularly important in light of known problems with misidentified genomes deposited in public databases (23).

Sequence alignments were then constructed from unique *mapA* and *cueE* alleles. The regions corresponding to the primer and probe binding sites for each gene were extracted and concatenated, and gene phylogenies were generated. Analysis of *mapA* primer and probe sequences (designed originally for specific *C. jejuni* identification) demonstrated conservation among *C. jejuni* isolates with relatively limited variation relative to the published sequences, whereas *mapA* primer and probe binding sites in *C. coli* isolates demonstrated significant variation (up to 18 nucleotides) from the published sequences. Consistent with prior work (5, 24, 25), some *C. coli* isolates were observed that carried *C. jejuni*-specific *mapA* alleles or composites of the two, to be discussed in more detail below. Analysis of *cueE* primer and probe sequences (designed originally for specific *C. coli* identification) demonstrated conservation among *C. coli* strains and similarity to published sequences and poor matches for *C. jejuni*, as expected. In contrast to *mapA* sequences, all *cueE* sequences were species specific.

Overall, species designations based on *in silico* analysis of *mapA*/*cueE* sequences were 100% concordant for *C. jejuni* and 96.9% concordant for *C. coli*, compared with rMLST-based designations, indicating excellent *in silico* performance of the assay. For *in vivo* confirmation of their findings, the authors had access to archived isolates and extracted DNA corresponding to the assemblies they used, demonstrating one of the powerful uses of isolate biorepositories paired with a genomic sequence database. A subset of isolates representing the genetic diversity of the data set was selected, and each unique primer-probe combination was tested at least once. Results were as predicted by the *in silico* analysis for all isolates, including the *C. coli* isolates carrying complete *C. jejuni mapA* sequences, which gave mixed amplifications (*mapA* positive/*cueE* positive). These genomic findings are explained by introgression—horizontal transfer of *C. jejuni mapA* sequences into the *C. coli* genome—with whole-allele replacement. Indeed, the original 2003 study had found a small number of isolates in which both gene targets were present, but the full nature and extent of the phenomenon were difficult to study at that time (8).

Introgression, or recombination between the genomes of two bacterial species occupying the same niche and resulting partial or full gene replacement, is a well-studied phenomenon, particularly in *C. jejuni* and *C. coli* (5, 24, 25). Introgression in this data set was analyzed with the Structure software package (26), and mixed *mapA* ancestry was found in 8.9% of the *C. coli* isolates tested, including two complete *mapA* allele transfers; mixed alleles resulting from partial transfers accounted for the remaining cases of introgression. As a consequence of the low frequency of complete gene transfer and the location of recombination breakpoints for partial transfers, the assay performance was, in fact, not significantly compromised by *mapA* introgression. Interestingly, no mixed ancestry was observed in the *cueE* gene. Similar to the case with shaft color in the northern flicker, introgression in binary gene pairs chosen for discrimination of "interbreeding" bacterial species (those linked by horizontal gene transfer) may lead to incorrect assumptions about species identity that are appreciated only when the larger population structure and degree of introgression are understood. Sequence analysis also demonstrated that a small number of isolates that could not be classified because of poor amplification of *cueE* corresponded to members of the *C. coli* third clade, mentioned above, and contained a *cueE* gene that was divergent from that of other *C. coli* isolates. Again, understanding of the larger population structure allowed *in silico* prediction of this finding.

The authors conclude that the high specificity of the assay is the consequence of high interspecies diversity and intraspecies conservation of the target genes. The rare instances of lack of specificity in *C. coli* isolates are explained as a consequence of introgression, and the instances of failure of the *cueE* primers to amplify a product in *C. coli* are explained as a consequence of clade divergence. This study both demonstrates the substantial contributions that WGS data can make to the evaluation and validation of traditional lab-developed PCR assays and provides one step-by-step approach using publicly available resources that another lab may follow. The next conceptual step beyond this work involves approaches to optimal primer selection from genome-wide analyses of potential targets. Though this study dealt with the particular problem of distinguishing bacterial species, the approach would find general application in the field of outbreak epidemiology, where distinctions between strains and clones are needed, as the same considerations apply. And as WGS databases continue to grow, these methods are sure to find countless other applications in clinical microbiology.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the National Institutes of Health Clinical Center. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. **National Audubon Society.** 2016. Audubon guide to North American birds. http://www.audubon.org/field-guide/bird/northern-flicker. National Audubon Society, Washington, DC. Accessed 16 September 2016.
2. **Rosselló-Mora R, Amann R.** 2001. The species concept for prokaryotes. FEMS Microbiol Rev **25:**39–67. http://dx.doi.org/10.1111/j.1574-6976.2001.tb00571.x.
3. **Cohan FM.** 2002. What are bacterial species? Annu Rev Microbiol **56:**457–487. http://dx.doi.org/10.1146/annurev.micro.56.012302.160634.
4. **Riley MA, Lizotte-Waniewski M.** 2009. Population genomics and the bacterial species concept. Methods Mol Biol **532:**367–377. http://dx.doi.org/10.1007/978-1-60327-853-9_21.
5. **Sheppard SK, McCarthy ND, Falush D, Maiden MC.** 2008. Convergence of Campylobacter species: implications for bacterial evolution. Science **320:**237–239. http://dx.doi.org/10.1126/science.1155532.
6. **Lefterova MI, Suarez CJ, Banaei N, Pinsky BA.** 2015. Next-generation sequencing for infectious disease diagnosis and management: a report of the Association for Molecular Pathology. J Mol Diagn **17:**623–634. http://dx.doi.org/10.1016/j.jmoldx.2015.07.004.

7. **Jansen van Rensburg MJ, Swift C, Cody AJ, Jenkins C, Maiden MCJ.** 2016. Exploiting bacterial whole-genome sequencing data for evaluation of diagnostic assays: *Campylobacter* species identification as a case study. J Clin Microbiol **54:**2882–2890. http://dx.doi.org/10.1128/JCM.01522-16.
8. **Best EL, Powell EJ, Swift C, Grant KA, Frost JA.** 2003. Applicability of a rapid duplex real-time PCR assay for speciation of Campylobacter jejuni and Campylobacter coli directly from culture plates. FEMS Microbiol Lett **229:**237–241. http://dx.doi.org/10.1016/S0378-1097(03)00845-0.
9. **Schuurman T, de Boer RF, van Zanten E, van Slochteren KR, Scheper HR, Dijk-Alberts BG, Moller AV, Kooistra-Smid AM.** 2007. Feasibility of a molecular screening method for detection of *Salmonella enterica* and *Campylobacter jejuni* in a routine community-based clinical microbiology laboratory. J Clin Microbiol **45:**3692–3700. http://dx.doi.org/10.1128/JCM.00896-07.
10. **Van Lint P, De Witte E, De Henau H, De Muynck A, Verstraeten L, Van Herendael B, Weekx S.** 2015. Evaluation of a real-time multiplex PCR for the simultaneous detection of Campylobacter jejuni, Salmonella spp., Shigella spp./EIEC, and Yersinia enterocolitica in fecal samples. Eur J Clin Microbiol Infect Dis **34:**535–542. http://dx.doi.org/10.1007/s10096-014-2257-x.
11. Reference deleted.
12. **Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND.** 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol **11:**728–736. http://dx.doi.org/10.1038/nrmicro3093.
13. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11:**595. http://dx.doi.org/10.1186/1471-2105-11-595.
14. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A **95:**3140–3145. http://dx.doi.org/10.1073/pnas.95.6.3140.
15. **Jolley KA, Maiden MC.** 2014. Using multilocus sequence typing to study bacterial variation: prospects in the genomic era. Future Microbiol **9:**623–630. http://dx.doi.org/10.2217/fmb.14.24.
16. **Simpson EH.** 1949. Measurement of diversity. Nature **163:**688. http://dx.doi.org/10.1038/163688a0.
17. **Hunter PR, Gaston MA.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol **26:**2465–2466.
18. **Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MC.** 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology **158:**1005–1015. http://dx.doi.org/10.1099/mic.0.055459-0.
19. **Revez J, Zhang J, Schott T, Kivisto R, Rossi M, Hanninen ML.** 2014. Genomic variation between *Campylobacter jejuni* isolates associated with milk-borne-disease outbreaks. J Clin Microbiol **52:**2782–2786. http://dx.doi.org/10.1128/JCM.00931-14.
20. **Kovanen SM, Kivisto RI, Rossi M, Schott T, Karkkainen UM, Tuuminen T, Uksila J, Rautelin H, Hanninen ML.** 2014. Multilocus sequence typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human infections in three districts during a seasonal peak in Finland. J Clin Microbiol **52:**4147–4154. http://dx.doi.org/10.1128/JCM.01959-14.
21. **Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC.** 2013. Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol **51:**2526–2534. http://dx.doi.org/10.1128/JCM.00066-13.
22. **Cody AJ, McCarthy NM, Wimalarathna HL, Colles FM, Clark L, Bowler IC, Maiden MC, Dingle KE.** 2012. A longitudinal 6-year study of the molecular epidemiology of clinical *Campylobacter* isolates in Oxfordshire, United Kingdom. J Clin Microbiol **50:**3193–3201. http://dx.doi.org/10.1128/JCM.01086-12.
23. **Beaz-Hidalgo R, Hossain MJ, Liles MR, Figueras MJ.** 2015. Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for Aeromonas genomes in the GenBank database. PLoS One **10:**e0115813. http://dx.doi.org/10.1371/journal.pone.0115813.
24. **Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJ, Ogden ID, Forbes K, French NP, Carter P, Miller WG, McCarthy ND, Owen R, Litrup E, Egholm M, Affourtit JP, Bentley SD, Parkhill J, Maiden MC, Falush D.** 2013. Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol **22:**1051–1064. http://dx.doi.org/10.1111/mec.12162.
25. **Sheppard SK, McCarthy ND, Jolley KA, Maiden MC.** 2011. Introgression in the genus Campylobacter: generation and spread of mosaic alleles. Microbiology **157:**1066–1074. http://dx.doi.org/10.1099/mic.0.045153-0.
26. **Pritchard JK, Stephens M, Donnelly P.** 2000. Inference of population structure using multilocus genotype data. Genetics **155:**945–959.