AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

DATABASES AND SOFTWARE

gndDb, a Database of Partial *gnd* Sequences To Assist with Analysis of *Escherichia coli* Communities Using High-Throughput Sequencing

Adrian L. Cookson,[a,b] David W. Lacher,[c] Flemming Scheutz,[d] David A. Wilkinson,[b,f] Patrick J. Biggs,[b,e,f] Jonathan C. Marshall,[b,e] Gale Brightwell[a,f]

[a]AgResearch Limited, Hopkirk Research Institute, Palmerston North, New Zealand

[b]mEpiLab, Hopkirk Research Institute, School of Veterinary Science, Massey University, Palmerston North, New Zealand

[c]Division of Molecular Biology, Office of Applied Research and Safety Assessment, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, Laurel, Maryland, USA

[d]Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark

[e]School of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[f]New Zealand Food Science and Safety Research Centre, Massey University, Palmerston North, New Zealand

ABSTRACT The use of culture methods to detect *Escherichia coli* diversity does not provide sufficient resolution to identify strains present at low levels. Here, we target the hypervariable *gnd* gene and describe a database containing 534 distinct partial *gnd* sequences and associated O groups for use with culture-independent *E. coli* community analysis.

Current culture-dependent studies to investigate *Escherichia coli* diversity in fecal and environmental samples often fail to identify strains that are present in low numbers. Our previous work using sequencing of metabarcoded amplicons has targeted the hypervariable *gnd* gene to provide a comprehensive analysis of *E. coli* community structure from complex samples such as feces (1). The *gnd* gene encodes 6-phosphogluconate dehydrogenase, the third enzyme in the pentose phosphate pathway, and is found in most *Enterobacteriaceae* (2–4). Usually, *gnd* is found adjacent to the highly recombinatorial O-antigen biosynthesis gene cluster (O-AGC) (5, 6), a region of the *E. coli* chromosome prone to horizontal gene transfer and recombination (7), which influences the O-group structure and cell surface antigenicity as the outermost component of the lipopolysaccharide (LPS) moiety. Although having no role in O-antigen biosynthesis, *gnd* has been described as a passive hitchhiker of recombination events influencing LPS antigenic changes (4).

By targeting *gnd* polymorphisms and adopting culture-independent methods, our previous work provided an indication of intestinal *E. coli* diversity and in parallel developed a *gnd* database for cross-referencing purposes using O-AGC DNA sequence data from distinct O groups (1). However, the increasing availability and analysis of whole-genome sequencing (WGS) data from *E. coli* (and *Shigella*) isolates have provided new insights as to the range of different *E. coli* O groups according to incongruent *wzx* and *wzy* (provisional OXY designations) or *wzm* and *wzt* (provisional OMT designations) gene sequences and the identification of six novel O groups (8). By isolating and examining the *gnd* gene from *E. coli* and *Shigella* draft genome assemblies described in recent studies (9, 10) and in other studies where novel O-AGCs have been submitted to GenBank (5, 6, 8), we have identified novel *gnd* alleles that have been included in a database resource that may be used for analysis of *E. coli* communities using amplicon sequencing. Analysis of WGS data from *E. coli* (and *Shigella*) with novel O groups has

Citation Cookson AL, Lacher DW, Scheutz F, Wilkinson DA, Biggs PJ, Marshall JC, Brightwell G. 2019. gndDb, a database of partial *gnd* sequences to assist with analysis of *Escherichia coli* communities using high-throughput sequencing. Microbiol Resour Announc 8:e00476-19. https://doi.org/10.1128/MRA.00476-19.

Editor Christina A. Cuomo, Broad Institute

Address correspondence to Adrian L. Cookson, adrian.cookson@agresearch.co.nz.

Received 25 April 2019
Accepted 18 July 2019
Published 15 August 2019

provided evidence for the identification of 534 distinct *gnd* sequence types (gSTs), each 284 bp in length, forming the core of this database.

The full-length *gnd* gene of 1,407 bp precludes its use in its entirety as a tool for culture-independent amplicon sequencing studies using the Illumina platform; therefore, the 284-bp region spanning nucleotide positions 443 to 727 is targeted for use in *E. coli* community analysis studies. The hypervariable nature of *gnd* also restricts suitable PCR primer sites for the generation of amplicons of a suitable size for routine high-throughput sequencing using the Illumina platform. This *gnd* database may also provide some assistance in the identification of novel O groups and offer a resource for *E. coli* subtyping using conventional dideoxy Sanger sequencing methods as a primary screen for subsequent WGS analysis (11).

Data availability. The DNA sequences of the 534 distinct gSTs are available in FASTA format from GitHub (https://github.com/mEpiLab/gnd) and are accompanied by a spreadsheet which provides matching O groups for each gST and a representative accession number of matching draft genome assemblies or submitted nucleotide sequences. The database and list of matching O groups will be curated and updated as further WGS data are made available.

ACKNOWLEDGMENTS

This work is a result of the ongoing efforts of the *E. coli/Shigella* Molecular Serotyping Working Group and was accelerated by their meeting that convened at Penn State University in November 2017.

Part of this work was supported by Strategic Science Investment Funding (Food Provenance & Assurance) and the Our Land & Water National Science Challenge.

REFERENCES

1. Cookson AL, Biggs P, Marshall JC, Reynolds A, Collis RM, French NP, Brightwell G. 2017. Culture independent analysis using *gnd* as a target gene to assess *Escherichia coli* diversity and community structure. Sci Rep 7:841. https://doi.org/10.1038/s41598-017-00890-6.
2. Selander R, Levin BR. 1980. Genetic diversity and structure of *Escherichia coli* populations. Science 210:545–547. https://doi.org/10.1126/science.6999623.
3. Biserčić M, Feutrier JY, Reeves PR. 1991. Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J Bacteriol 173:3894–3900. https://doi.org/10.1128/jb.173.12.3894-3900.1991.
4. Nelson K, Selander R. 1994. Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc Natl Acad Sci U S A 91:10227–10231. https://doi.org/10.1073/pnas.91.21.10227.
5. Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson N. 2015. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. DNA Res 22:101–107. https://doi.org/10.1093/dnares/dsu043.
6. DebRoy C, Fratamico PM, Yan X, Baranzoni G, Liu Y, Needleman DS, Tebbs R, O'Connell CD, Allred A, Swimley M, Mwangi M, Kapur V, Raygoza Garay JA, Roberts EL, Katani R. 2016. Comparison of O-antigen gene clusters of all O-serogroups of *Escherichia coli* and proposal for adopting a new nomenclature for O-typing. PLoS One 11:e0147434. https://doi.org/10.1371/journal.pone.0147434.
7. Milkman R, Jaeger E, McBride R. 2003. Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. Genetics 163:475–483.
8. Iguchi A, Iyoda S, Seto K, Nishii H, Ohnishi M, Mekata H, Ogura Y, Hayashi T. 2016. Six novel O genotypes from Shiga toxin-producing *Escherichia coli*. Front Microbiol 7:765. https://doi.org/10.3389/fmicb.2016.00765.
9. Gangiredla J, Mammel MK, Barnaba TJ, Tartera C, Gebru ST, Patel IR, Leonard SR, Kotewicz ML, Lampel KA, Elkins CA, Lacher DW. 2017. Species-wide collection of *Escherichia coli* isolates for examination of genomic diversity. Genome Announc 5:e01321-17. https://doi.org/10.1128/genomeA.01321-17.
10. Trees E, Strockbine N, Changayll S, Ranganathan S, Zhao K, Weil R, MacCannell D, Sabol A, Schmidtke A, Martin H, Stripling D, Ribot EM, Gerner-Smidt P. 2014. Genome sequences of 228 Shiga toxin-producing *Escherichia coli* isolates and 12 isolates representing other diarrheagenic *E. coli* pathotypes. Genome Announc 2:e00718-14. https://doi.org/10.1128/genomeA.00718-14.
11. Gilmour MW, Olson A, Andrysiak A, Ng L, Chui L. 2007. Sequence-based typing of genetic targets encoded outside of the O-antigen gene cluster is indicative of Shiga toxin-producing *Escherichia coli* serogroup lineages. J Med Microbiol 56:620–628. https://doi.org/10.1099/jmm.0.47053-0.