bioRxiv preprint doi: https://doi.org/10.1101/317610; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Easily phylotyping *E. coli* via the EzClermont web app and command-line tool.

Nicholas R. Waters,[1,2] Florence Abram,[1] Fiona Brennan,[1,3] Ashleigh Holmes,[4] and Leighton Pritchard[2*]

[1] *Department of Microbiology, School of Natural Sciences, National University of Ireland, Galway, Ireland*
[2] *Information and Computational Sciences, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland*
[3] *Soil and Environmental Microbiology, Environmental Research Centre, Teagasc, Johnstown Castle, Wexford, Ireland*
[4] *Cell and Molecular Sciences, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland*

*To whom correspondence should be addressed: leighton.pritchard@hutton.ac.uk

Compiled: 2018/05/08 18:30:00

Summary

The Clermont PCR method of phylotyping Escherichia coli has remained a useful classification scheme despite the proliferation of higher-resolution sequence typing schemes. We have implemented an in silico Clermont PCR method as both a web app and as a command-line tool to allow researchers to easily apply this phylotyping scheme to genome assemblies easily.

Availability and Implementation

EzClermont is available as a web app at `https://nickp60.pythonanywhere.com`. For local use, EzClermont can be installed with pip or installed from the source code at `https://github.com/nickp60/ezclermont`. All analysis was done with version 0.4.0.

Contact

n.waters4@nuigalway.ie

Supplementary information

Table S1: test dataset; S2: validation dataset; S3: results.

Keywords: genome assembly, ribosome, benchmarking, scaffolding, *de fere novo*

bioRxiv preprint doi: https://doi.org/10.1101/317610; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Escherichia coli is among the most widely studied organisms, and the species is very diverse [10, 12]. Because of this diversity, many methods have been developed to differentiate the different *E. coli* lineages. In 1987, Selandar and colleagues used electrophoretic analysis of a 35 enzyme digest to classify the *E. coli* Reference Collection (ECOR) in 6 phylogenetic groups (A-F) [12]. Clermont and colleagues published their triplex PCR method [3] of phylotyping, which proved to be an extremely valuable tool to differentiate groups A, B1, B2, and D, being cited over 625 times as of April 2018. In 2013, Clermont and colleagues published an update to this work [4], in which they showed that by adding a 4th set of primers (with additional primers to differentiate the subgroups and the cryptic clades), higher resolution could be achieved, as this expanded the method to detect groups E, F, and differentiate the cryptic clades. This approach has been widely adopted, as the method is reliable, easy to interpret, and can be performed rapidly.

Other sequence typing schemes have been developed to classify *E. coli* strains. These include the Achtman 2012 7 gene Multi Locus Sequence Typing (MLST) [1, 2], Michigan EcMLST [11], whole-genome MLST (`http://www.applied-maths.com/applications/wgmlst`), core-genome MLST [6], two-locus MLST [16], and ribosomal MLST [9]. All these methods classify *E. coli* with greater accuracy and granularity than the phylotyping, but at the cost of interpretability. The Clermont 2013 phylotyping scheme remains a regularly utilised tool in classifying *E. coli*.

We developed EzClermont to provide a simple implementation of the Clermont phylotyping algorithm to genome assemblies. For researchers unfamiliar with command-line tools, we have implemented the software as a web application; for those needing to process large numbers of assemblies, a command-line interface can be installed via pip.

In short, the software uses constrained string matching as an in silico PCR to determine the presence or absence of the alleles used to determine the phylotype. As assemblies may contain alleles interrupted by breaks between contigs, we give the user the option to allow partial matches (ie, if one of the two primers matched, but the expected position of the other primer fell beyond the sequence end).

As PCR primers do not necessarily need 100% sequence identity to function, we determined the variability at the priming sites in 523 strains. To do this, we downloaded the genome assemblies from NCBI Bioprojects PRJNA218110, PRJNA231221, and PRJNA352562. From each assembly, we extracted the 7 regions matching the theoretical amplicons of the quadriplex, E-specific, C-specific, and E/C control primer sets from Clermont 2013. Any differences between a sequence and the primer sequence reported in Clermont 2013 were incorporated into the search query, except for differences in the last 5 nucleotides on the 3' regions (as those can be used to differentiate alleles) [14].

To assess the performance of EzClermont, we selected a test dataset and a validation dataset. Additionally, the strains from Clermont, 2013 Figure 1 are used as unit tests in the package.

As a test set, we used strains listed in Sims and Kim 2011 [13] (Table S1), and the validation set of 95 strains was the genomes from Clermont 2015 [5] (Table S2)[1]. Comparing the reported phylogroup and the EzClermont phylogroup for the 19 strains in Sims and Kim (excluding strains reported in both Clermont 2015 and Sims and Kim), 3 of the 19 did not agree, but two of those (IAI39, SMS-3-5) were shown by other works to have the phylotype that EzClermont predicted (see Table 1). The one strain that typed differently (APEC01) was examined and was found to have the

[1] 6 of the 101 total strains were omitted as no genome assembly was available.

bioRxiv preprint doi: https://doi.org/10.1101/317610; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ArpA allele that is not normally detected in B2 strains.

Table 1: Comparing EzClermont to phylotypes reported by Sims and Kim 2011 [13]

Strain	Assembly	Sims and Kim	EzClermont	Notes
APEC01	GCA_003028815.1	B2	A	found arpA fragment
IAI39	GCA_000026345.1	D	F	See Hazen 2017 [8]; reported as phylogroup F
SMS-3-5	GCA_000019645.1	D	F	See Vangchhia 2016 [15]; reported as phylogroup F

We ran EzClermont on the 95 strains from Clermont 2015 and compared the results to the reported phylotype; 89 of the 95 strains classifications matched. To determine whether the inconsistent phylogroup assignments matched phylogeny, we then generated a parsimony tree using kSNP3 [7], and plotted with ggtree [17]. This revealed that the EzClermont classification of ECOR46 (similar IAI39 and SMS-3-5) appears to match the true phylogeny, as opposed to the phylogroup reported in the literature (Figure 1). Of those that didn't match, all detected at least one theoretical amplicon that was not reported to be there (Table S3).

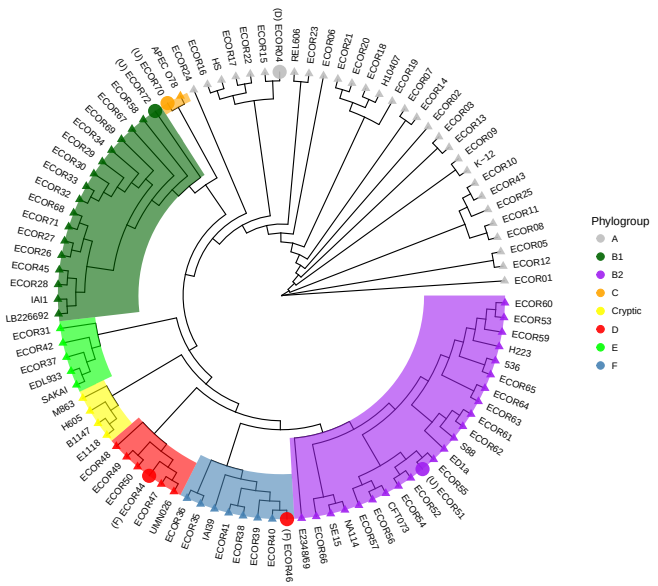

Figure 1: Parsimony cladogram of strains from Clermont et al 2015. Tree was generated with kSNP3 (k=19). Enlarged circular tips show where EzClermont differed from reported phylogroup (EzClermont type show in brackets).

Considering both the testing and validation datasets (114 strains), EzClermont has an accuracy of 94%. Given the ease of use of the web app for simple queries, and the speed of execution for larger batches, we hope that EzClermont will be of use to the community.

bioRxiv preprint doi: https://doi.org/10.1101/317610; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Competing interests

The authors declare that they have no competing interests.

Funding

The work was funded through a joint studentship between The James Hutton Institute, Dundee, Scotland, and the National University of Ireland, Galway, Ireland.

Authors' contributions

NRW wrote all the bugs.

Acknowledgements

Many thanks to Stephen Nolan, Dr. Corine Nzeteu, and Dr. Alma Siggins for their comments on the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/317610; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Mark Achtman, John Wain, François-Xavier Weill, Satheesh Nair, Zhemin Zhou, Vartul Sangal, Mary G. Krauland, James L. Hale, Heather Harbottle, Alexandra Uesbeck, Gordon Dougan, Lee H. Harrison, Sylvain Brisse, and the S. enterica MLST study Group. Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella enterica. *PLoS Pathogens*, 8(6):e1002776, jun 2012.

[2] Nabil-Fareed Alikhan, Zhemin Zhou, Martin J. Sergeant, and Mark Achtman. A genomic overview of the population structure of Salmonella. *PLOS Genetics*, 14(4):e1007261, apr 2018.

[3] Olivier Clermont, Stéphane Bonacorsi, and Edouard Bingen. Rapid and simple determination of the Escherichia coli phylogenetic group. *Applied and environmental microbiology*, 66(10):4555–8, oct 2000.

[4] Olivier Clermont, Julia K. Christenson, Erick Denamur, and David M. Gordon. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports*, 5(1):58–65, feb 2013.

[5] Olivier Clermont, David Gordon, and Erick Denamur. Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. *Microbiology*, 161(5):980–988, may 2015.

[6] Mark de Been, Mette Pinholt, Janetta Top, Stefan Bletz, Alexander Mellmann, Willem van Schaik, Ellen Brouwer, Malbert Rogers, Yvette Kraat, Marc Bonten, Jukka Corander, Henrik Westh, Dag Harmsen, and Rob J. L. Willems. Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of Enterococcus faecium. *Journal of Clinical Microbiology*, 53(12):3788–3797, dec 2015.

[7] Shea N Gardner, Tom Slezak, and Barry G. Hall. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome: Table 1. *Bioinformatics*, 31(17):2877–2878, sep 2015.

[8] Tracy H Hazen, Jane Michalski, Qingwei Luo, Amol C Shetty, Sean C Daugherty, James M Fleckenstein, and David A Rasko. Comparative genomics and transcriptomics of Escherichia coli isolates carrying virulence factors of both enteropathogenic and enterotoxigenic E. coli. *Scientific reports*, 7(1):3513, jun 2017.

[9] K. A. Jolley, C. M. Bliss, J. S. Bennett, H. B. Bratcher, C. Brehony, F. M. Colles, H. Wimalarathna, O. B. Harrison, S. K. Sheppard, A. J. Cody, and M. C. J. Maiden. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, 158(Pt_4):1005–1015, apr 2012.

[10] Oksana Lukjancenko, Trudy M. Wassenaar, and David W. Ussery. Comparison of 61 Sequenced Escherichia coli Genomes. *Microbial Ecology*, 60, 2010.

[11] Weihong Qi, David W Lacher, Alyssa C Bumbaugh, Katie E Hyma, Lindsey M Ouellette, Teresa M Large, Cheryl L Tarr, and Thomas S Whittam. EcMLST: an Online Database for Multi Locus Sequence Typing of Pathogenic Escherichia coli. *IEEE Computational Systems Bioinformatics Conference*, 2004.

[12] R K Selander, D A Caugant, and T S Whittam. Genetic structure and variation in natural populations of Escherichia coli. In Frederick C. Neidhardt, editor, *Escherichia coli and Salmonella : cellular and molecular biology*, volume 2, pages 1625–1648. American Society for Microbiology Press, Washington, D.C. :, 1987.

bioRxiv preprint doi: https://doi.org/10.1101/317610; this version posted May 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[13] Gregory E Sims and Sung-Hou Kim. Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences of the United States of America*, 108(20):8329–34, may 2011.

[14] Ralph Stadhouders, Suzan D Pas, Jeer Anber, Jolanda Voermans, Ted H M Mes, and Martin Schutten. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *The Journal of molecular diagnostics : JMD*, 12(1):109–17, jan 2010.

[15] Belinda Vangchhia, Sam Abraham, Jan M. Bell, Peter Collignon, Justine S. Gibson, Paul R. Ingram, James R. Johnson, Karina Kennedy, Darren J. Trott, John D. Turnidge, and David M. Gordon. Phylogenetic diversity, antimicrobial susceptibility and virulence characteristics of phylogroup F Escherichia coli in Australia. *Microbiology*, 162(11):1904–1912, nov 2016.

[16] Scott J. Weissman, James R. Johnson, Veronika Tchesnokova, Mariya Billig, Daniel Dykhuizen, Kim Riddell, Peggy Rogers, Xuan Qin, Susan Butler-Wu, Brad T. Cookson, Ferric C. Fang, Delia Scholes, Sujay Chattopadhyay, and Evgeni Sokurenko. High-Resolution Two-Locus Clonal Typing of Extraintestinal Pathogenic Escherichia coli. *Applied and Environmental Microbiology*, 78(5):1353–1360, mar 2012.

[17] Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, jan 2017.