The Journal of Open Source Software

Corekaburra: pan-genome post-processing using core gene synteny

Magnus G. Jespersen [1], **Andrew Hayes** [1], **and Mark R. Davies** [1]

1 Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

DOI: 10.21105/joss.04910

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @iferres
- @asafpr

Submitted: 18 October 2022
Published: 30 November 2022

Summary

Pan-genome analysis enables an assessment of the total gene content from a set of genome sequences. Since the 'first' defined pan-genome (Tettelin et al., 2005), many tools have been developed which improve the methodological process used to construct pan-genomes (Gautreau et al., 2020; Inman et al., 2019; Page et al., 2015; Thorpe et al., 2018; Tonkin-Hill et al., 2020). However, current limitations lay in extracting meaningful interpretations from downstream analyses of pan-genomes. Few tools have been made to address this problem, and often focus on a specific analysis, such as pan-genome association studies (Brynildsrud et al., 2016; Lees et al., 2018; Whelan et al., 2020). Here we present Corekaburra, a tool to define gene regions based on core gene synteny within a pan-genome. Defining regions by flanking core genes provide context to genes and allow for easier systematic comparisons of genomic features, such as genomic inversions and gene insertions and deletions.

Statement of need

Bacterial genomes from the same species can vary considerably in their genetic content (Welch et al., 2002). In population genomics and other studies of bacterial genomes an important piece of information is shared genetic information. Due to this, pan-genomes have become a standard method for investigating variation in genetic content of bacteria (Medini et al., 2020). The analysis of pan-genomes is critical to basic research, industrial strain development, and public health surveillance systems. Despite this, methods to systematically dissect pan-genomes are sparse.

We propose Corekaburra, a program designed to reduce the complexity of outputs from pan-genome pipelines, easing the discovery of regions of variation within a pan-genome. The input for Corekaburra is annotated genomes (Gff3 format with appended genome), similar to those used by existing pan-genome pipelines, and the output folder from a pan-genome tool (currently Roary or Panaroo)(Page et al., 2015; Tonkin-Hill et al., 2020). Corekaburra introduces core gene synteny by scanning and summarizing input Gff files based on the genetic distance of nucleotides and any coding sequences between core genes of the pan-genome. Using gene synteny is not a novel concept and is used by many pan-genome tools and comparable methods to facilitate pan-genome accuracy and analysis (Bayliss et al., 2019; Bazin et al., 2020; Beier & Thomson, 2022; Page et al., 2015; Tonkin-Hill et al., 2020). The novelty of Corekaburra is its focus on core genes and defining regions using these. Core genes are 'stable' in occurrence across genomes of a pan-genome, making them good reference points for comparisons across genomes. Additionally, Corekaburra is not associated with a single pan-genome pipeline, but takes a general input defined by presence and absence of genes plus the genome annotations in a standard format used by popular pan-genome pipelines. As long as the two above input formats can be supplied, Corekaburra is agnostic to the specifics of the

pan-genome tools used. This allows for easy adaptation of Corekaburra to current, future, or custom pan-genome pipelines.

Acknowledgements

MGJ is supported by The Melbourne Research Scholarship from The University of Melbourne. MRD is supported by a University of Melbourne CR Roper Fellowship.

References

Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K., & Feil, E. J. (2019). PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience*, *8*(10), giz119. https://doi.org/10.1093/gigascience/giz119

Bazin, A., Gautreau, G., Médigue, C., Vallenet, D., & Calteau, A. (2020). panRGP: A pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, *36*(Supplement_2), i651–i658. https://doi.org/10.1093/bioinformatics/btaa792

Beier, S., & Thomson, N. R. (2022). Panakeia-a universal tool for bacterial pangenome analysis. *BMC Genomics*, *23*(1), 1–8. https://doi.org/10.1186/s12864-022-08303-3

Brynildsrud, O., Bohlin, J., Scheffer, L., & Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with scoary. *Genome Biology*, *17*(1), 1–9. https://doi.org/10.1186/s13059-016-1108-8

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P. C., & Vallenet, D. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Computational Biology*, *16*(3), e1007732. https://doi.org/10.1371/journal.pcbi.1007732

Inman, J. M., Sutton, G. G., Beck, E., Brinkac, L. M., Clarke, T. H., & Fouts, D. E. (2019). Large-scale comparative analysis of microbial pan-genomes using PanOCT. *Bioinformatics*, *35*(6), 1049–1050. https://doi.org/10.1093/bioinformatics/bty744

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). Pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310–4312. https://doi.org/10.1093/bioinformatics/bty539

Medini, D., Donati, C., Rappuoli, R., & Tettelin, H. (2020). *The Pangenome: A Data-Driven Discovery in Biology*. https://doi.org/10.1007/978-3-030-38281-0_1

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

Thorpe, H. A., Bayliss, S. C., Sheppard, S. K., & Feil, E. J. (2018). Piggy: A rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience*, *7*(4), giy015. https://doi.org/10.1093/gigascience/giy015

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., &

Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*. https://doi.org/10.1186/s13059-020-02090-4

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L. T., Donnenberg, M. S., & Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *99*(26), 17020–17024. https://doi.org/10.1073/pnas.252529799

Whelan, F. J., Rusilowicz, M., & McInerney, J. O. (2020). Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics*, *6*(3). https://doi.org/10.1099/mgen.0.000338