

ALPPACA - A tool for Prokaryotic Phylogeny And Clustering Analysis

Håkon Kaspersen¹ and Eve Zeyl Fiskebeck¹

¹ Norwegian Veterinary Institute, Ås, Norway

DOI: [10.21105/joss.04677](https://doi.org/10.21105/joss.04677)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Charlotte Soneson](#)

Reviewers:

- [@mberacochea](#)
- [@hseabolt](#)
- [@rcannood](#)

Submitted: 23 June 2022

Published: 22 November 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

A tool for Prokaryotic Phylogeny And Clustering Analysis (ALPPACA) is a pipeline that allows both *de-novo* and reference-based phylogenetic analysis of prokaryotic genomes. The pipeline provides a suite of analyses tailored for different scenarios, designed to allow analysis of datasets represented by three different genetic diversity levels, all in one package. These levels of similarity influence what assumptions are used to consider sequences as orthologous when reconstructing the multiple alignment required for phylogenetic inference. By selecting an appropriate track for the data at hand, the user can be confident that these assumptions are taken care of within the framework of ALPPACA.

Statement of need

Phylogenetic analysis is frequently used to unravel outbreaks and track the origins of pathogens worldwide. Phylogenetic analysis has also become commonplace in several research projects and clinical investigations, where time is often of the essence. This kind of analysis often entails running several tools consecutively, and many assumptions are made for the data used in each tool. To add to this complexity, several tools have been developed for each step in such an analysis, and sifting through these tools as a user may be time-consuming. Additionally, choosing a combination of compatible software for various analysis scenarios may require in-depth knowledge and experience in the field of microbial evolution, which often prevents non-specialists from utilizing such analyses. Here we present a solution that will help alleviate these problems in hopes of making it easier and faster to run reproducible phylogenetic analyses.

State of the field

The ability to run different datasets through different tracks within the same framework makes ALPPACA unique compared to other phylogeny pipelines such as Bactmap (<https://nf-co.re/bactmap>) and SNVPhyl (https://github.com/DHQP/SNVPhyl_Nextflow), where only mapping-based phylogeny is possible. Several research projects have provided the developmental platform of ALPPACA, and peer-reviewed papers have been published using this framework for phylogenetic analysis ([Franklin-Alming et al., 2021](#); [Kaspersen et al., 2020](#); [Kravik et al., 2022](#); [Smistad et al., 2022](#)).

Pipeline and track descriptions

Pipeline

The ALPPACA pipeline is written in Nextflow ([Di Tommaso et al., 2017](#)), and the code and documentation are publicly available on GitHub (<https://github.com/NorwegianVeterinaryInstitute/ALPPACA>). The user has the option of running the pipeline using different container handlers, such as docker, singularity, or conda. Each track generates a tidy html report summarizing the main results from each analysis.

Tracks

The pipeline consists of three separate tracks depending on the objectives and data available to the user ([Figure 1](#)). The tracks differ in the way they detect homologous regions to construct the multiple alignment needed for phylogenetic inference.

First, the core gene track is designed to be used for datasets that are expected to have a relatively high level of genetic diversity. This track is useful if you have a dataset with different but closely related species, or different sequence types (ST) of the same species. The track provides the means to generate a relatively high resolution phylogeny on diverse datasets. Here, the genomes are annotated with Prokka ([Seemann, 2014](#)), the pangenome is inferred using Panaroo ([Tonkin-Hill et al., 2020](#)), and a core gene alignment is produced and used for phylogenetic reconstruction with IQ-TREE ([Nguyen et al., 2015](#)).

The mapping track is designed for datasets that are expected to have a medium to low diversity level. This track maps reads to a reference using Snippy (<https://github.com/tseemann/snippy>). Mapping reads circumvents the need to generate assemblies, which is a time-consuming process. Time is usually of the essence in an outbreak situation, and this track allows for rapid analysis of outbreak data. Additionally, this track ensures that only genetic material from vertical descent is included in the analysis, as recombinant areas are detected by Gubbins ([Croucher et al., 2015](#)) and masked with Maskrc-svg (<https://github.com/kwongj/maskrc-svg>) before the phylogeny is inferred with IQ-TREE.

Lastly, the core genome track is designed to be used for datasets that are expected to have a low level of genetic diversity, e.g. within the same ST. This track is useful after identifying clusters of interest using the core gene track above, to increase the resolution of the phylogenetic analysis on a subset of the dataset. This track outputs the percent length of each genome included in the alignment, reported as average genome coverage by ParSNP ([Treangen et al., 2014](#)). This is an important parameter to consider when interpreting the results, as it tells the user how much of each genome the phylogenetic inference is based on. Similar to the mapping track, only data from vertical descent is included in the analysis.

All three tracks also generate SNP distance statistics with snp-dists (<https://github.com/tseemann/snp-dists>). The SNP distances are very useful when defining clusters, or if defining outbreak clades based on a SNP distance cutoff. The user also has options to deduplicate the alignment with SeqKit ([Shen et al., 2016](#)), and filter out constant sites with snp-sites ([Page et al., 2016](#)), which will reduce the runtime of the pipeline.

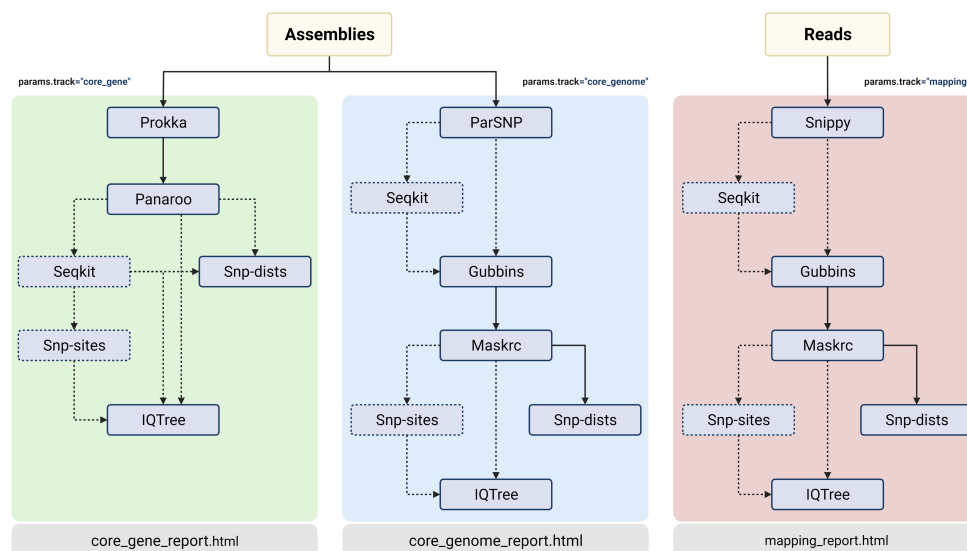


Figure 1: Overview of the three tracks in ALPPACA.

Conclusion

The ALPPACA pipeline provides a suite of phylogenetic analyses for different scenarios, all in one package. This enables a variety of uses without having to download several tools and programs, and the Nextflow framework allows for user-friendly and reproducible use of the pipeline. Additional tracks may be added to ALPPACA in the future, such as clustering based on core/whole genome multi locus sequence typing, or additions to existing tracks, such as recombination detection in the core gene analysis. Clustering analysis using FastANI (<https://github.com/ParBLiSS/FastANI>) will be added as a separate track, to assist users in selecting the correct track by evaluating genetic diversity in their dataset.

Acknowledgements

The projects QREC-MaP (Research Funding for Agriculture and the Food Industry, Norwegian Research Council, project number 255383), KLEB-GAP (Trond Mohn Foundation, project number TMS2019TMT03), Yersiniosis at Sea (Norwegian Seafood Research Fund grant, project number 901505), and Increasing sustainability of Norwegian food production by tackling streptococcal infections in modern livestock systems (FFL/JA, Norwegian Agricultural Agreement Research Fund, project number 280364) are acknowledged for providing the research platform for this work. The computations were performed on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway.

References

- Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., & Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, 43(3), e15–e15. <https://doi.org/10.1093/nar/gku1196>
- Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319. <https://doi.org/10.1038/nbt.3820>
- Franklin-Alming, F., Kaspersen, H., Hetland, M. A. K., Bakksjø, R., Nesse, L. L., Leangapichart, T., Löhr, I. H., Telke, A. A., & Sunde, M. (2021). Exploring *Klebsiella pneumoniae* in healthy poultry reveals high genetic diversity, good biofilm-forming abilities and higher prevalence in turkeys than broilers. *Frontiers in Microbiology*, 12(725414), 11. <https://doi.org/10.3389/fmicb.2021.725414>
- Kaspersen, H., Sekse, C., Fiskebeck, E. Z., Slettemeås, J. S., Simm, R., Norström, M., Urdahl, A. M., & Lagesen, K. (2020). Dissemination of quinolone-resistant *Escherichia coli* in the Norwegian broiler and pig production chains and possible persistence in the broiler production environment. *Applied and Environmental Microbiology*, 86(7), e02769–19. <https://doi.org/10.1128/AEM.02769-19>
- Kravik, I. H., Kaspersen, H., Sjurseth, S. K., Jonsson, M., David, B., Aspholm, M., & Sekse, C. (2022). High sequence similarity between avian pathogenic *E. coli* isolates from individual birds and within broiler chicken flocks during colibacillosis outbreaks. *Veterinary Microbiology*, 267(109378), 109378. <https://doi.org/10.1016/j.vetmic.2022.109378>
- Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von, & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274. <https://doi.org/10.1093/molbev/msu300>
- Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, A., & Harris, S. R. (2016). SNP-sites: Rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, 5. <https://doi.org/10.1099/mgen.0.000056>
- Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
- Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE*, 11(10), e0163962. <https://doi.org/10.1371/journal.pone.0163962>
- Smistad, M., Kaspersen, H., Franklin-Alming, F. V., Wolff, C., Sølverød, L., Porcellato, D., Trettenes, E., & Jørgensen, H. J. (2022). *Streptococcus dysgalactiae* ssp. *dysgalactiae* in Norwegian bovine dairy herds: Risk factors, sources, and genomic diversity. *Journal of Dairy Science*, 105(4), 3574–3587. <https://doi.org/10.3168/jds.2021-21471>
- 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(1), 180. <https://doi.org/10.1186/s13059-020-02090-4>
- Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11), 524. <https://doi.org/10.1186/s13059-014-0524-x>