The Journal of Open Source Software

# Dnaapler: A tool to reorient circular microbial genomes

**George Bouras** [1,2], **Susanna R. Grigson** [3], **Bhavya Papudeshi** [3], **Vijini Mallawaarachchi** [3], **and Michael J. Roach** [3,4]

**1** Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia **2** The Department of Surgery – Otolaryngology Head and Neck Surgery, Central Adelaide Local Health Network, Adelaide, South Australia 5000, Australia **3** Flinders Accelerator for Microbiome Exploration, College of Science and Engineering, Flinders University, Bedford Park, Adelaide, South Australia 5042, Australia **4** Adelaide Centre for Epigenetics and South Australian Immunogenomics Cancer Institute, The University of Adelaide, Adelaide, South Australia 5005, Australia

**DOI:** 10.21105/joss.05968

**Software**

- Review
- Repository
- Archive

**Editor:** Frederick Boehm

**Reviewers:**

- @mkerin
- @vinisalazar

**Submitted:** 09 August 2023
**Published:** 11 January 2024

## Summary

Microorganisms found in natural environments are fundamental components of ecosystems and play vital roles in various ecological processes. Studying their genomes can provide valuable insights into the diversity, functionality, and evolution of microbial life, as well as their impacts on human health. Once the genetic material is extracted from environmental samples, it undergoes sequencing using technologies like whole genome sequencing (WGS). The raw sequence data is then analysed, and computational methods are applied to assemble the fragmented sequences and reconstruct the complete microbial genomes (Wick et al., 2021) (Mallawaarachchi et al., 2023) (Bouras et al., 2023).

Many biological entities including Bacteria, Archaea, plasmids, bacteriophages and other viruses can have circular genomes. Once assembled, a circular genome sequence is represented as a linear character string and labelled in some way to indicate that it should be circular. The point at which the linear sequence begins is random, due to the nature of the algorithms employed in assembling genomes from sequencing reads. Such arbitrary startpoints can affect downstream genome annotation and analysis; they may occur within coding sequences (CDS), can disrupt the prediction potential of mobile genetic elements like prophages, and make pangenome analyses based on gene order difficult. Therefore, microbial sequences are often required to be reoriented to begin by convention with certain genes: the dnaA chromosomal replication initiator gene for bacterial chromosomes, the repA plasmid replication initiation gene for plasmids and the terL large terminase subunit gene for bacteriophages as shown in Figure 1. Here we present Dnaapler, a flexible microbial sequence reorientation tool that allows for rapid and consistent orientation of circular microbial genomes such as Bacteria, plasmids and bacteriophages. Dnaapler is hosted on GitHub at github.com/gbouras13/dnaapler.

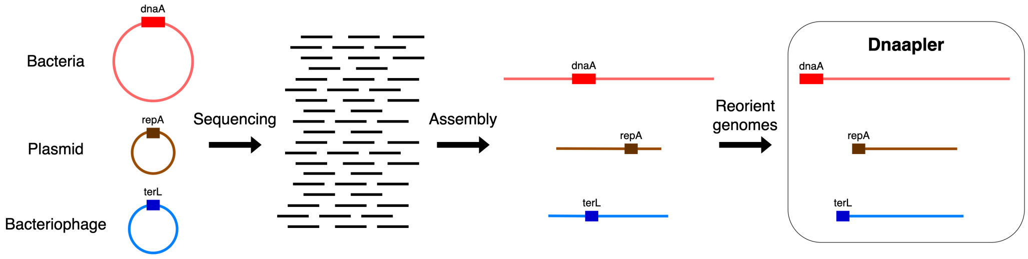

**Figure 1:** Example microbial genome assembly workflow.

## Statement of need

Circlator (Hunt et al., 2015) is the most commonly used dedicated tool for reorienting bacterial genomes. However, Circlator was designed for bacterial chromosomes and plasmids only, is no longer supported by its developers, has several burdensome external dependencies, and requires the corrected reads in FASTA or FASTQ format along with the FASTA genome assembly as input. Alternatively, genome reorientation is often performed manually or with custom scripts on a genome-by-genome and project-by-project basis, making integration into assembly workflows difficult, and creating inconsistencies between different projects and researchers. We propose Dnaapler, a light-weight command-line tool written in Python that can be easily integrated into assembly workflows. Dnaapler takes only a FASTA formatted genome file as input. It uses BLAST (Altschul et al., 1990) (Mount, 2007) — its only external dependency — or Pyrodigal (Larralde, 2022) (Hyatt et al., 2010) depending on the chosen subcommand for reorientation. A list of the subcommands provided in Dnaapler are as follows:

| Subcommand | Database used | Gene used to reorient |
|---|---|---|
| chromosome | Custom database downloaded from Swissprot | dnaA chromosomal replication initiator gene |
| plasmid | repA database curated from Unicycler (Wick et al., 2017) | repA plasmid replication initiation gene |
| phage | Prokaryotic Virus Remote Homologous Groups database (PHROGs) (Terzian et al., 2021) | terL large terminase subunit gene |
| all | Chromosome, plasmid and phage databases combined | dnaA, repA and terL |
| custom | User specified | Custom gene |
| mystery | Pyrodigal predicted coding sequences | Random CDS |
| nearest | Pyrodigal predicted coding sequences | First CDS (nearest to the start) |
| largest | Pyrodigal predicted coding sequences | Largest CDS |
| bulk | Either chromosome, plasmid, phage or custom. Requires multiple input contigs. | dnaA, repA, terL or a custom gene |

Specifically, Dnaapler 'chromosome', 'phage' and 'plasmid' subcommands use blastx (protein databases are searched using a translated nucleotide query) to search for the dnaA, terL or repA gene respectively in the input genomes, using built-in amino acid databases for each gene. Dnaapler 'all' will run a blastx search against all three databases simultaneously, prioritising dnaA hits then repA and finally terL if multiple genes have hits. Taking the top blastx hit, Dnaapler will check that the first amino acid of the BLAST alignment begins with Methionine, Valine, or Leucine (the 3 most used gene start codons in bacteria and bacteriophages). If it does, then it will then reorient the genome to begin at that position in the forward direction. If it does not, then Pyrodigal will be used to predict all coding sequences. Dnaapler will calculate the CDS with the most overlap to the top blastx hit, and reorient the genome to begin with the start codon of that CDS in the forward direction.

If the 'custom' subcommand is selected, the same process will be conducted but with a user specified amino acid FASTA formatted input database. If the 'mystery', 'nearest' or 'largest' subcommands are selected, Pyrodigal will be used to predict all coding sequences, and the genome will be reoriented to begin with either a random (mystery), the first (nearest) CDS, or the largest CDS respectively. Dnaapler returns an output directory containing a log file and the genome reoriented as a FASTA formatted file. Finally, the 'bulk' subcommand can be used to reorient multiple input contigs (in a mulitFASTA format file) using either the chromosome, plasmid, phage or custom reorientation strategies.

Examples of Dnaapler's functionality on the C333 *Staphylococcus aureus* chromosome and the C333 Sa3int prophage (GenBank accession GCA_030288915.1, Sample Number SAMN32360890 from BioProject PRJNA914892 from (Houtak et al., 2023)) are shown below using the plotting functionalities of Bakta v1.8.2 (Schwengers et al., 2021) and Pharokka v1.5.1 (Bouras et al., 2022).

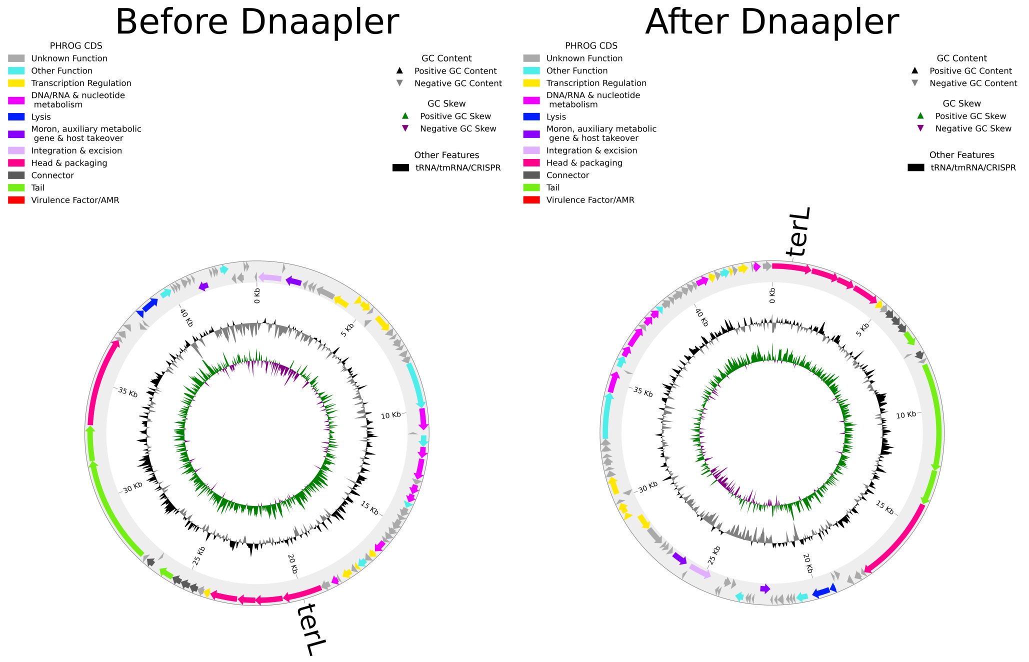

**Figure 2:** Example Dnaapler phage reorientation of the c333 Sa3int prophage as a circular genomic map from Pharokka beginning at the top of the circle. Each coloured arrow represents a CDS. The large terminase subunit gene is labelled as terL. Dnaapler identified the terL gene as beginning with coordinate 19146 on the forward strand.

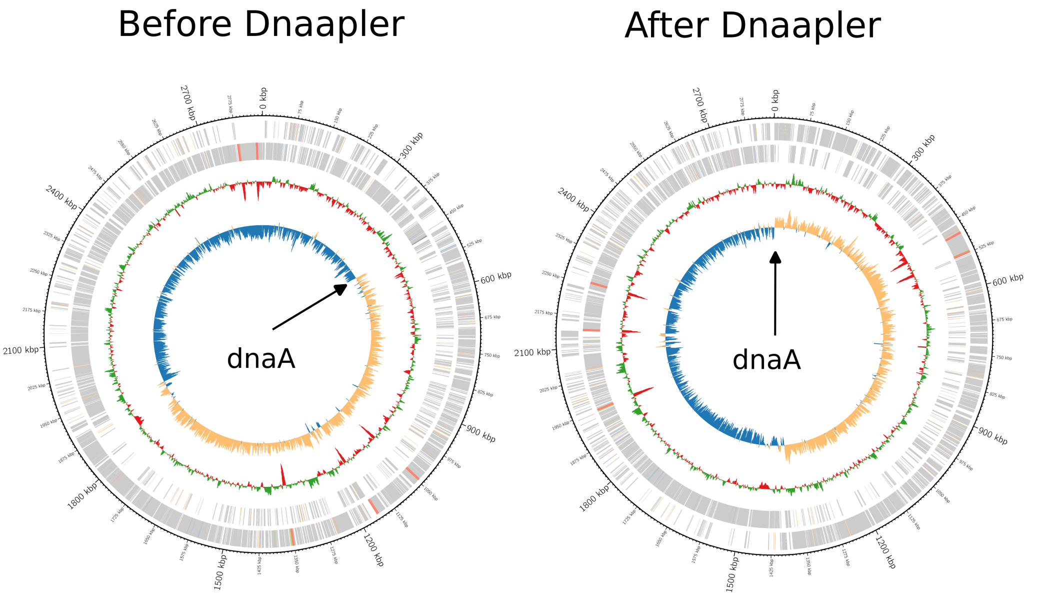

**Figure 3:** Example Dnaapler chromosome reorientation of the C333 chromosome as a circular genomic map from Bakta beginning at the top of the circle. Each thin line represents a CDS, with the positive stranded CDSs denoted in the outer ring and the negative stranded CDSs in the inner ring. The position of the chromosomal replication initiator gene is labelled as dnaA. The red and green ring denotes the GC content while the blue and yellow ring denotes the GC skew. Dnaapler identified the dnaA gene as beginning with coordinate 466140 on the reverse strand.

## Availability

Dnaapler is distributed on PyPI. A Conda package is also available in the Bioconda channel (Grüning et al., 2018). The source code is available on GitHub, and features continuous integration tests and test coverage, and continuous deployment using GitHub actions. Dnaapler has already been integrated into the United States of America StaPH-B (State Public Health Lab Bioinformatics) consortium Docker image collection.

## Acknowledgements

We would like to thank Michael B. Hall for providing some code snippets particularly the external tool class from his tool tbpore, Ryan Wick for curating a repA database from Unicycler and Sarah Vreugde and Robert A. Edwards for their supervision.

## References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bouras, G., Nepal, R., Houtak, G., Psaltis, A. J., Wormald, P.-J., & Vreugde, S. (2022). Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics*, *39*(1), btac776. https://doi.org/10.1093/bioinformatics/btac776

Bouras, G., Sheppard, A. E., Mallawaarachchi, V., & Vreugde, S. (2023). Plassembler: an automated bacterial plasmid assembly tool. *Bioinformatics*, *39*(7), btad409. https://doi.org/10.1093/bioinformatics/btad409

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Houtak, G., Bouras, G., Nepal, R., Shaghayegh, G., Cooksley, C., Psaltis, A. J., Wormald, P.-J., & Vreugde, S. (2023). The intra-host evolutionary landscape and pathoadaptation of persistent staphylococcus aureus in chronic rhinosinusitis [Journal Article]. *Microbial Genomics*, *9*(11). https://doi.org/10.1099/mgen.0.001128

Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., & Harris, S. R. (2015). Circlator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biology*, *16*(1), 294. https://doi.org/10.1186/s13059-015-0849-0

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Larralde, M. (2022). Pyrodigal: Python bindings and interface to prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, *7*(72), 4296. https://doi.org/10.21105/joss.04296

Mallawaarachchi, V., Roach, M. J., Decewicz, P., Papudeshi, B., Giles, S. K., Grigson, S. R., Bouras, G., Hesse, R. D., Inglis, L. K., Hutton, A. L. K., Dinsdale, E. A., & Edwards, R. A. (2023). Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics*, *39*(10), btad586. https://doi.org/10.1093/bioinformatics/btad586

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harbor Protocols*, *2007*(7), pdb–top17. https://doi.org/10.1101/pdb.top17

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification [Journal Article]. *Microbial Genomics*, *7*(11). https://doi.org/10.1099/mgen.0.000685

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M.-A., & Enault, F. (2021). PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics*, *3*(3), lqab067. https://doi.org/10.1093/nargab/lqab067

Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Méric, G., Vezina, B., Wyres, K. L., & Holt, K. E. (2021). Trycycler: Consensus long-read assemblies for bacterial genomes. *Genome Biology*, *22*(1), 266. https://doi.org/10.1186/s13059-021-02483-z

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), 1–22. https://doi.org/10.1371/journal.pcbi.1005595