The Journal of Open Source Software

Acanthophis: a comprehensive plant hologenomics pipeline

Kevin D. Murray [1¶], **Justin O. Borevitz** [2], **Detlef Weigel** [1], and **Norman Warthmann** [2,3¶]

1 Max Planck Institute for Biology, Tübingen, Deutschland **2** Research School of Biology, Australian National University, Canberra, Australia **3** FAO/IAEA Joint Centre of Nuclear Techniques in Food and Agriculture, Plant Breeding and Genetics Laboratory, Seibersdorf, Austria ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Marcos Vital

Reviewers:

- @bricoletc
- @gbouras13
- @abhishektiwari

Submitted: 15 October 2023
Published: unpublished

Summary

Acanthophis is a comprehensive pipeline for the joint discovery and analysis of both plant genetic variation and variation in the composition and abundance of plant-associated microbiomes. Implemented in Snakemake (Köster & Rahmann, 2012), Acanthophis handles data from raw FASTQ read files through quality control, alignment of the reads to a plant reference, variant calling, taxonomic classification and quantification of microbes, and metagenome analysis. The workflow contains numerous practical optimisations, both to reduce disk space usage and maximise utilisation of computational resources. Acanthophis is available under the Mozilla Public Licence v2 at https://github.com/kdm9/Acanthophis as a python package installable from conda or PyPI (`pip install acanthophis`).

Statement of Need

Understanding plant biology benefits from ecosystem-scale analysis of genetic variation, and increasingly demands the characterisation of not only plant genomes but also the genomes of their associated microbes. Such analyses are often data intensive, particularly at the scale required for quantitative analyses, i.e. thousands of host individuals (Karasov et al., 2022; Regalado et al., 2020). They demand computationally-efficient pipelines that perform both host genotyping and host-associated microbiome characterisation in a consistent, flexible, and reproducible fashion.

Currently, no such unified pipelines exist. Previous pipelines perform only a subset of these tasks (e.g. Snakemake's variant calling pipeline; Köster et al. (2021)). In addition, most host-aware microbiome analysis pipelines do not allow for host genotyping and/or assume an animal host (e.g. Taxprofiler; Yates et al. (2023)). Acanthophis has attracted many users, and has been referred to in peer-reviewed journal articles and preprints (e.g. Murray et al. (2019); Ahrens et al. (2021)).

Components and Features

Acanthophis is a pipeline for the analysis of plant population resequencing data. It expects short-read shotgun whole (meta-)genome sequencing data, typically of plants collected in the field. A typical dataset might be 10s-1000s of samples from one or multiple closely related species, sequenced with 2x150bp paired-end short read sequencing. In a plant-microbe interaction genomics study, these plants and therefore sequencing libraries can contain microbes (a "hologenome"), however datasets focusing only on host genome variation are also catered for. Acanthophis can be configured to do any of the following analyses: mapping reads to a reference, calling variants, annotating variant effects, estimating genetic distances *de novo*, and profiling and/or assembling metagenomes. While we developed Acanthophis to handle plant data, there is no reason why it cannot be applied to other taxa, however some parameters may need adjustment (see below).

Across the entire pipeline, Acanthophis operates on 'sample sets', named groups of one or more samples, and each sample can be in any number of sample sets. The pipeline is configured via a global `config.yaml` file, in which one can configure the pipeline per sample-set. This way, one can configure the analyses to be run (most can be disabled if not needed), as well as tool-specific settings or thresholds. We provide a documented template as well as a reproducible workflow to simulate test data, which can be used as a basis for customisation.

Stage 1: Raw reads to per-sample reads

Input data consists of FASTQ files per **run** of each **library** corresponding to a **sample**. For each **runlib** (one run of one library), Acanthophis uses `AdapterRemoval` (Schubert et al., 2016) to remove low quality and adaptor sequences, and optionally to merge overlapping read pairs. It then uses `FastQC` to summarise sequence QC before and after `AdaptorRemoval`.

Stage 2: Alignment to reference(s)

To align reads to reference genomes, Acanthophis can use any of `BWA MEM` (Li, 2013), `NGM` (Sedlazeck et al., 2013), and `minimap2` (Li, 2018, 2021). Then, Acanthophis merges per-runlib BAMs to per-sample BAMs, and uses `samtools markdup` (Danecek et al., 2021; Li et al., 2009) to mark duplicate reads. Input reference genomes should be uncompressed, `samtools faidx`ed FASTA files.

Stage 3: Variant Calling

Acanthophis uses `bcftools mpileup` and/or `freebayes` to call raw variants, using priors and thresholds configurable for each sample set. It then normalises variants with `bcftools norm`, splits multi-allelic variants, filters each allele with per-sample set filters, and combines filter-passing alleles back into unique sites, merges region-level VCFs, indexes, and calculates statistics on these final VCF files. Acanthophis provides two alternative approaches to parallelize variant calling: either a static list of non-overlapping genome windows (supplied in a BED file), or genome bins with approximately equal amounts of data, which are automatically generated using mosdepth (Pedersen & Quinlan, 2018).

Stage 4: Taxon profiling

Acanthophis uses any of Kraken 2 (Wood et al., 2019), Bracken (Lu et al., 2017), Kaiju (Menzel et al., 2016), Centrifuge (Kim et al., 2016), and Diamond (Buchfink et al., 2015) to create taxonomic profiles for each sample against any number of taxon identification databases (e.g. those provided with aforementioned methods, or from public sequence databases). We then use taxpasta (Beber et al., 2023) to combine multiple profiles into tables for easy downstream use.

Stage 5: *De novo* Estimates of Genetic Dissimilarity

Acanthophis can use either `kWIP` (Murray et al., 2017) or Mash (Ondov et al., 2016) to estimate genetic distances between samples without alignment to a reference genome. These features first sketch reads into k-mer sketches, and then calculate pairwise distances among samples.

Stage 6: Reporting and Statistics

Throughout all pipeline stages, various tools output summaries of their actions and/or outputs. We optionally combine these into unified reports by pipeline stage and sample set using MultiQC (Ewels et al., 2016).

Acknowledgements

We thank Luisa Teasdale, Anne-Cecile Colin, Rose Andrew, Johannes Köster, and Scott Ferguson for comments or advice on Acanthophis and/or on this manuscript. KDM is supported by a Marie Skłodowska-Curie Actions fellowship.

References

Ahrens, C. W., Murray, K. D., Mazanec, R. A., Ferguson, S., Bragg, J., Jones, A., Tissue, D. T., Byrne, M., Borevitz, J. O., & Rymer, P. D. (2021, August 8). *Genomic constraints to drought adaptation*. https://doi.org/10.1101/2021.08.07.455511

Beber, M. E., Borry, M., Stamouli, S., & Yates, J. A. F. (2023). TAXPASTA: TAXonomic Profile Aggregation and STAndardisation. *Journal of Open Source Software*, *8*(87), 5627. https://doi.org/10.21105/joss.05627

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Karasov, T. L., Neumann, M., Shirsekar, G., Monroe, G., Team, P., Weigel, D., & Schwab, R. (2022, April 10). *Drought selection on Arabidopsis populations and their microbiomes*. https://doi.org/10.1101/2022.04.08.487684

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*. https://doi.org/10.1101/gr.210641.116

Köster, J., Micwessolly, Kuthe, E., & De Coster, W. (2021). *Snakemake-workflows/dna-seq-gatk-variant-calling*. https://doi.org/10.5281/ZENODO.4677629

Köster, J., & Rahmann, S. (2012). Snakemake — a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://arxiv.org/abs/1303.3997v2

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. (2021). New strategies to improve Minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, *3*, e104. https://doi.org/10.7717/peerj-cs.104

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*, 11257. https://doi.org/10.1038/ncomms11257

Murray, K. D., Janes, J. K., Jones, A., Bothwell, H. M., Andrew, R. L., & Borevitz, J. O. (2019). Landscape drivers of genomic diversity and divergence in woodland Eucalyptus. *Molecular Ecology*, *28*(24), 5232–5247. https://doi.org/10.1111/mec.15287

Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., & Warthmann, N. (2017). kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLOS Computational Biology*, *13*(9), e1005727. https://doi.org/10.1371/journal.pcbi.1005727

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*, 132. https://doi.org/10.1186/s13059-016-0997-x

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Regalado, J., Lundberg, D. S., Deusch, O., Kersten, S., Karasov, T., Poersch, K., Shirsekar, G., & Weigel, D. (2020). Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. *The ISME Journal*, *14*(8, 8), 2116–2130. https://doi.org/10.1038/s41396-020-0665-8

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*, 88. https://doi.org/10.1186/s13104-016-1900-2

Sedlazeck, F. J., Rescheneder, P., & Von Haeseler, A. (2013). NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790–2791. https://doi.org/10.1093/bioinformatics/btt468

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1, 1), 1–13. https://doi.org/10.1186/s13059-019-1891-0

Yates, J. A. F., Stamouli, S., Andersson-Li, L., Beber, M. E., Mesilaakso, L., Nf-Core Bot, Christensen, T. A., Mahwash Jamy, JIANHONG OU, Stepien, R., Borry, M., Husen M. Umer, Syme, R., Hübner, A., & Zandra Fagernäs. (2023). *Nf-core/taxprofiler*. https://doi.org/10.5281/ZENODO.7728364