The Journal of Open Source Software

chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data

Markus J Ankenbrand[1, a], **Simon Pfaff**[2, a], **Niklas Terhoeven**[2, 3], **Musga Qureischi**[3, 4], **Maik Gündel**[3], **Clemens L. Weiß**[5], **Thomas Hackl**[6], and **Frank Förster**[2, 3, 7]

1 Department of Animal Ecology and Tropical Biology (Zoology III), University of Würzburg, Germany **2** Center for Computational and Theoretical Biology, University of Würzburg **3** Department of Bioinformatics, University of Würzburg **4** Centre for Experimental Molecular Medicine, University Clinics Würzburg, Germany **5** Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany **6** Department of Civil and Environmental Engineering, Massachusetts Institute of Technology **7** Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Applied Ecology and Bioresources, Gießen, Germany **a** These authors contributed equally to this work

DOI: 10.21105/joss.00464

Software

- Review
- Repository
- Archive

Submitted: 29 September 2017
Published: 16 January 2018

Summary

This is an automated pipeline that extracts and reconstructs chloroplast genomes from whole genome shotgun data. It is capable to assemble the incidentally sequenced chloropast DNA, which is present in almost all plant sequencing projects, due to the extraction of whole cellular DNA. It works by analyzing the k-mer distribution (determined with Jellyfish, (Marçais and Kingsford 2011), peak detection with R (R Core Team 2008)) of the raw sequencing reads. Usually the coverage of the chloroplast genome is much higher than that of the nuclear genome. Using mapping to reference chloroplast sequences (using bowtie2 (Langmead and Salzberg 2012), samtools ("The Sequence Alignment/Map Format and Samtools" 2009), and bedtools (Quinland and Hall 2010)) and the k-mer distribution candidate chloroplast reads are extracted from the complete set (Figure 1). Afterwards, the targeted assembly of those sequences is much faster and yields less contigs compared to an assembly of all reads. Assemblers usually fail to assemble chloroplast genomes as a single contig due to their structure, consisting of two single copy regions and an inverted repeat. The size of the inverted repeat is in most cases multiple kilobasepairs in size, therefore it can not be resolved using short reads only. However SPAdes (Nurk et al. 2013) returns the assembly graph where the typical chloroplast structure can be recognized and reconstructed using the knowledge of its structure. Using our demo set, one can achieve a single contig assembly of the chloroplast of *Spinacia oleracea* . If the assembly process does not finish with a single chloroplast sequence all remaining sequences are BLASTed (Camacho et al. 2009) against a database of reference chloroplasts to retain all partial sequences of interest. The final chloroplast sequence can be further annotated with tools like DOGMA (Wyman, Jansen, and Boore 2004), cpGAVAS (Liu et al. 2012) and VERDANT (McKain et al. 2017). Such assemblies, can be used to remove chloroplast reads before a genomic assembly of the remaining nuclear DNA. Moreover, chloroplast genomes are useful in phylogenetic reconstruction (Huang et al. 2016) or barcoding applications (Coissac et al. 2016). A similar tool, aiming the assembly of whole chloroplast genomes is the Python program org.ASM, but it is not production ready, yet. Also plasmid SPAdes (Antipov et al. 2016) could possibly be used for this purpose although it is not intended for it. In the future, we plan to use our chloroExtractor to screen NCBI's Sequence Read Archive (Leinonen, Sugawara, and Shumway 2011) for chloroplast genomes in public sequencing datasets

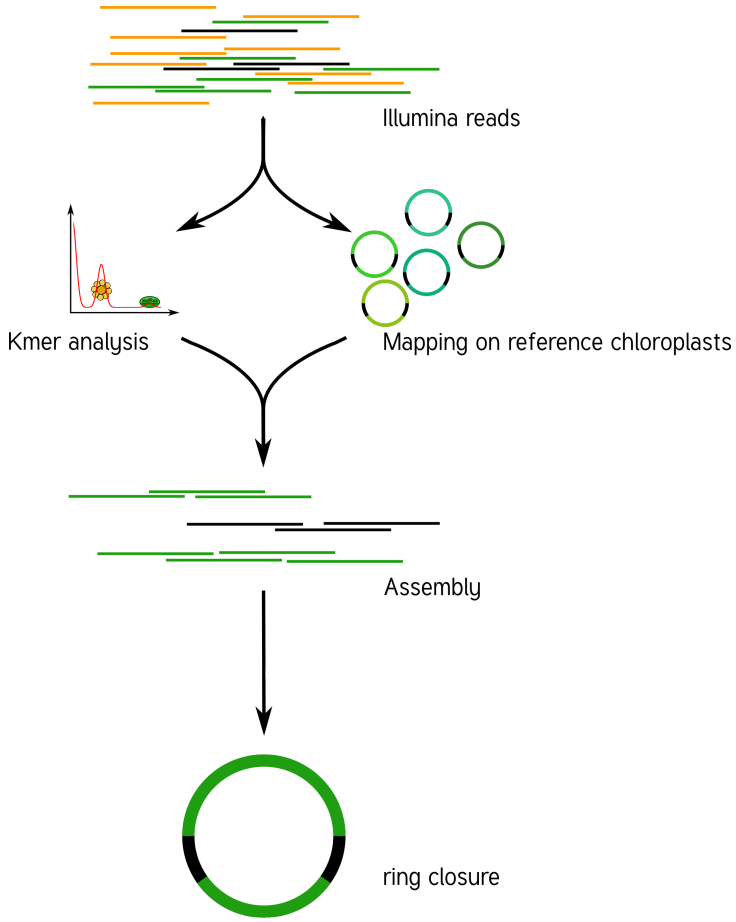

Figure 1: Schematic workflow of chloroExtractor.

that are not yet available in chloroplast databases, eg. chloroDB (Cui et al. 2006) to broaden our knowledge about chloroplasts.

In addition to the components cited above the chloroExtractor uses Ghostscript, Phyton, and Perl. Further the following Perl modules are used: Moose, Log::Log4Perl, Graph, Term::ProgressBar, IPC::Run, and File::Which.

Acknowledgements

MJA was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. We thank Daniel Amsel for his help testing and solving the Mac-based issues with our docker container.

References

Antipov, Dmitry, Nolan Hartwick, Max Shen, Mikhail Raiko, Alla Lapidus, and Pavel Pevzner. 2016. "PlasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data." *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/048942.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications."

BMC Bioinformatics 10 (1). https://doi.org/10.1186/1471-2105-10-421.

Coissac, Eric, Peter M. Hollingsworth, Sébastien Lavergne, and Pierre Taberlet. 2016. "From Barcodes to Genomes: Extending the Concept of Dna Barcoding." *Molecular Ecology* 25 (7):1423–8. https://doi.org/10.1111/mec.13549.

Cui, Liying, Narayanan Veeraraghavan, Alexander Richter, Kerr Wall, Robert K. Jansen, Jim Leebens-Mack, Izabela Makalowska, and Claude W. dePamphilis. 2006. "ChloroplastDB: The Chloroplast Genome Database." *Nucleic Acids Research* 34:D692–D696. https://doi.org/10.1093/nar/gkj055.

Huang, Yuling, Xiaojuan Li, Zhenyan Yang, Chengjin Yang, Junbo Yang, and Yunheng Ji. 2016. "Analysis of Complete Chloroplast Genome Sequences Improves Phylogenetic Resolution in Paris (Melanthiaceae)." *Frontiers in Plant Science* 7:1797. https://doi.org/10.3389/fpls.2016.01797.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9. https://doi.org/10.1038/nmeth.1923.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. 2011. "The Sequence Read Archive." *Nucleic Acids Research* 39:D19–D21. https://doi.org/10.1093/nar/gkq1019.

Liu, Chang, Linchun Shi, Yingjie Zhu, Haimei Chen, Jianhui Zhang, Xiaohan Lin, and Xiaojun Guan. 2012. "CpGAVAS, an Integrated Web Server for the Annotation, Visualization, Analysis, and Genbank Submission of Completely Sequenced Chloroplast Genome Sequences." *BMC Genomics* 13 (1):715. https://doi.org/10.1186/1471-2164-13-715.

Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers." *Bioinformatics* 27 (6):764–70. https://doi.org/10.1093/bioinformatics/btr011.

McKain, Michael R., Ryan H. Hartsock, Molly M. Wohl, and Elizabeth A. Kellogg. 2017. "Verdant: Automated Annotation, Alignment and Phylogenetic Analysis of Whole Chloroplast Genomes." *Bioinformatics* 33 (1):130–32. https://doi.org/10.1093/bioinformatics/btw583.

Nurk, Sergey, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey Prjibelsky, et al. 2013. "Assembling Genomes and Mini-Metagenomes from Highly Chimeric Reads." In *Research in Computational Molecular Biology: 17th Annual International Conference, Recomb 2013, Beijing, China, April 7-10, 2013. Proceedings*, edited by Minghua Deng, Rui Jiang, Fengzhu Sun, and Xuegong Zhang, 158–70. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37195-0_13.

Quinland, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (1). https://doi.org/10.1093/bioinformatics/btq033.

R Core Team. 2008. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

"The Sequence Alignment/Map Format and Samtools." 2009. *Bioinformatics* 25 (1). https://doi.org/10.1093/bioinformatics/btp352.

Wyman, Stacia K., Robert K. Jansen, and Jeffrey L. Boore. 2004. "Automatic Annotation of Organellar Genomes with Dogma." *Bioinformatics* 20 (17):3252–5. https://doi.org/10.1093/bioinformatics/bth352.