DOI: https://doi.org/10.17816/ecogen112363

PhaseAll: a simple tool for read-based allele phasing

Peter M. Zhurbenko [1, 2], Fedor N. Klimenko [3]

[1] Saint Petersburg State University, Saint Petersburg, Russia;
[2] Komarov Botanical Institute of the Russian Academy of Sciences, Saint Petersburg, Russia;
[3] Federal Scientific Center of Rehabilitation of the Disabled named after G.A. Albrecht, Saint Petersburg, Russia

The currently used genome assembly algorithms do not provide for allele phasing. This can lead to the loss of important information about the genotype of diploid and polyploid individuals. Here we introduce PhaseAll, a simple tool for allele phasing based on short reads obtained by second-generation sequencing. As input data, the tool takes paired reeds in SAM format. PhaseAll iterates sequentially through each alignment position. When a polymorphic position (SNP, insertion or deletion) is first encountered, a unique mutation is written to each allele. For each subsequent polymorphic position, a test is made to verify whether it is located on the same pair of reads (one DNA fragment) as the previous one. If two mutations are located on the same fragment, they are considered to belong to the same allele. If no fragments are found that connected at least one pair of neighboring polymorphic positions, an «X» is written in the allele sequences. This means that the alleles can swap at this position.

PhaseAll is written in python 3. SAM files are processed using the pysam library. PhaseAll is designed to separate only two alleles. To avoid possible sequencing errors, the user can set a read depth threshold below which the polymorphic position will be skipped. Some indels can cause errors in allele phasing, so PhaseAll has an option to skip indels for more accurate SNP reconstruction. The tool was tested on sequences of agrobacterial origin in the *Camellia* L. genome in more than 100 samples. PhaseAll is available for download on the GitHub: https://github.com/pzhurbenko/PhaseAll.

The research was supported by RSF (project No. 21-14-00050).

AUTHORS' INFO

Peter M. Zhurbenko, Researcher, Laboratory of Biosystematics and Cytology, Komarov Botanical Institute of the Russian Academy of Sciences; Junior Researcher, Department of Genetics and Biotechnology. Saint Petersburg State University, Saint Petersburg, Russia. E-mail: pj_28@mail.ru

Fedor N. Klimenko, Junior Researcher, Department of Innovative Technologies and Technical Means of Rehabilitation, Institute of Prosthetics and Orthotics. Federal Scientific Center of Rehabilitation of the Disabled named after G.A. Albrecht, Saint Petersburg, Russia. E-mail: tniknko@gmail.com