DOI: 10.25205/978-5-4437-1691-6-2

VALIDATION OF *A. COLUZZII* AND *A. GAMBIAE* HYBRIDS HAPLOTYPE PHASING BY TRIO BINNING

S. D. Dubovskova[1], A. A. Zamyatin[2]

[1] ITMO University, Saint Petersburg
[2] Pennsylvania State University, USA

✉ dubovskova.sofya@mail.ru

Abstract

The majority of nowadays genome assemblies are mixed haplotypes. An assembly of such haplotypes introduces DNA variants not present in any of true haplotypes, which negatively affects downstream genome analysis. Therefore, we assembled phased haplotypes of major malaria vectors in Sub-Saharan Africa using trio binning method. To validate haplotype phasing, we performed p-distance and k-mer quotient analyses. The results indicate correct haplotypes phasing.

The majority of nowadays genome assemblies is a mixture of true haplotypes. An assembly of a mixed haplotype introduces DNA variants not present in any of true haplotypes, which negatively affects downstream genome analysis [1]. Therefore, we assembled phased haplotypes of major malaria vectors in Sub-Saharan Africa, *A. coluzzii* and *A. gambiae* [2], using trio binning method [1].

Female and male F_1 offspring of *A. coluzzii* female and *A. gambiae* male were sequenced using Oxford Nanopore Technology [3], and assembled using TrioCanu [1]. As a result, we obtained *A. coluzzii* female and male haplotypes named AcoMOPhfm and AcolMOPhmm respectively, and *A. gambiae* female and male haplotypes named AgamZANUhfp and AgamZANUhmp respectively (unpublished). To validate haplotype phasing, we performed p-distance analysis, and KQ (k-mer quotient).

The p-distance is a proportion of nucleotides in which two sequences differ [4]. To assess haplotypes phasing we calculated p-distance of parental Illumina reads to phased haplotypes. To calculate p-distance we aligned parental Illumina reads on phased haplotypes using BWA-MEM2 [5]. The data from the alignment files was then used to calculate average p-distance per 10 Kb.

P-distance analysis demonstrated that p-distance of parental reads to the same species haplotypes is lower than 0.01 in general, while p-distance of parental reads to different species haplotypes lays between 0.02 and 0.04 in general (Fig. 1–4). This result indicates that parental reads possess higher similarity to the same species haplotypes, therefore, haplotypes are phased correctly.

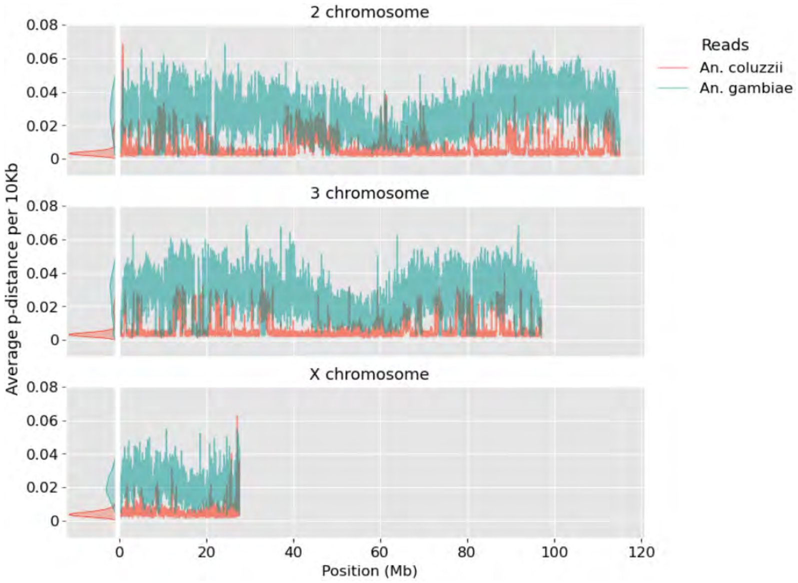

Fig. 1. P-distance of parental reads to *A. coluzzii* AcolMOPhfm haplotype

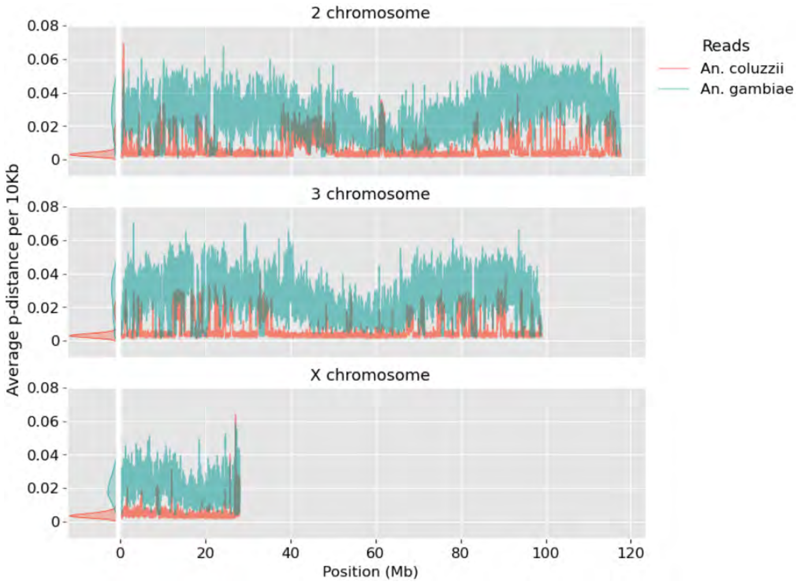

Fig. 2. P-distance of parental reads to *A. coluzzii* AcolMOPhmm haplotype

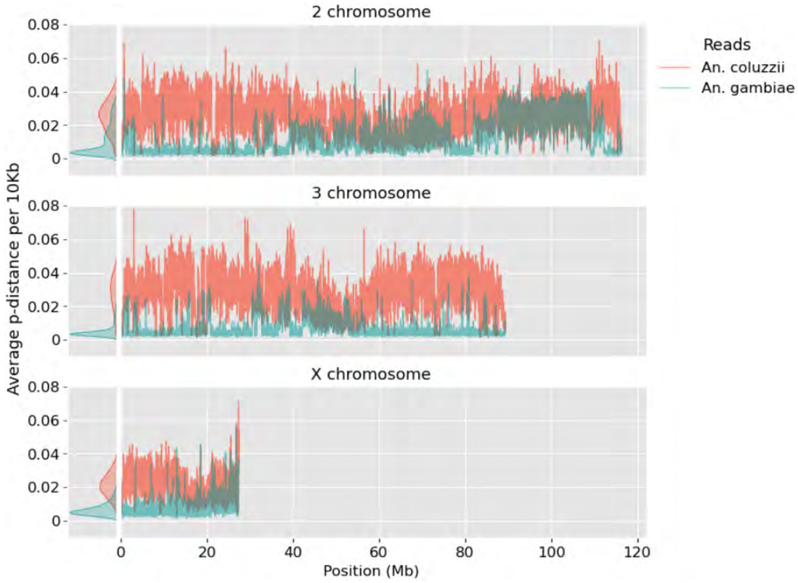

Fig. 3. P-distance of parental reads to *A. gambiae* AgamZANUhfp haplotype

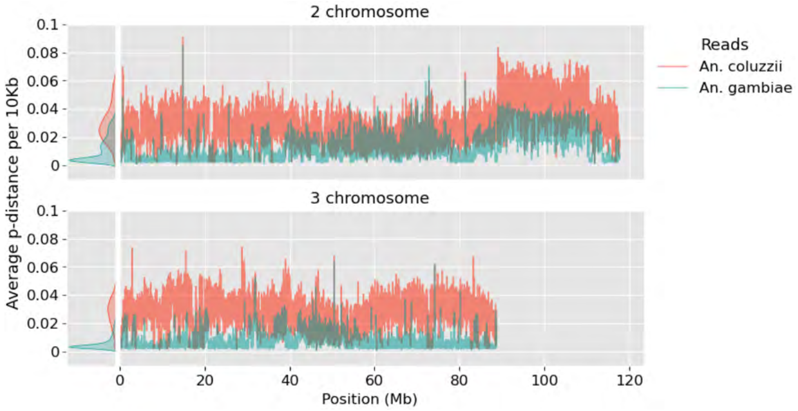

Fig. 4. P-distance of parental reads to *A. gambiae* AgamZANUhmp haplotype

The KQ analysis was initially developed for heterogametic chromosome contigs identification in the genome, and was tested on *B. mori*, females of which have W heterogametic chromosome [6]. In this method W chromosome contigs are identified through number of female-specific k-mers mapped to contigs. According to KQ, k-mer is female-specific if it is not found in male genome, and contig is W chromosome-derived if it has more than 10 female-specific k-mers per 1 Kb mapped to it. But, to validate haplotype phasing, we searched for contigs derived from *A. gambiae* instead of heterogametic chromosome contigs. We extracted k-mers from parental reads using Jellyfish [7] and used them as an input for the script developed by the authors of this method along with phased haplotypes.

KQ analysis identified no *A. gambiae* derived contigs in *A. coluzzii* phased haplotypes AcolMOPhfm and AcolMOPhmm (Fig. 5). On the contrary, this method assigned to *A. gambiae* all chromosome scaffolds and some unscaffolded contigs of *A. gambiae* phased haplotypes AgamZANUhfp and AgamZANUhmp (see Fig. 5). Therefore, KQ validates haplotype phasing.

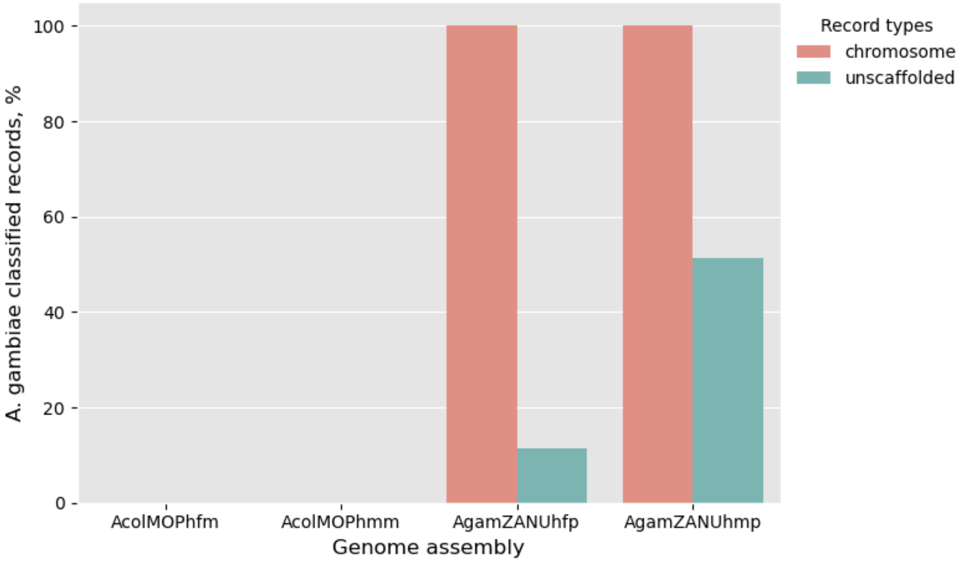

Fig. 5. Genome assembly scaffolds and contigs classified as *A. gambiae* derived

In conclusion, p-distance and KQ analyses validate *A. coluzzii* and *A. gambiae* hybrids haplotype phasing by trio binning.

References

1. Koren S., Rhie A., Walenz B. P. et al. De novo assembly of haplotype-resolved genomes with trio binning // Nat. Biotechnol. 2018. Vol. 36 (12). P. 1174–1182.

2. Zamyatin A., Avdeyev P., Liang J. et al. Chromosome-level genome assemblies of the malaria vectors *Anopheles coluzzii* and *Anopheles arabiensis* // GigaScience. 2021. Vol. 10 (3).

3. Astier Y., Braha O., Bayley H. Toward Single Molecule DNA Sequencing: Direct Identification of Ribonucleoside and Deoxyribonucleoside 5‘-Monophosphates by Using an Engineered Protein Nanopore Equipped with a Molecular Adapter // J. Am. Chem. Soc. 2006. Vol. 128 (5). P. 1705–1710.

4. Bredemeyer K. R., Harris A. J., Li G. et al. Ultracontinuous Single Haplotype Genome Assemblies for the Domestic Cat (*Felis catus*) and Asian Leopard Cat (*Prionailurus bengalensis*) // Journal of Heredity. 2021. Vol. 112 (2). P. 165–173.

5. Vasimuddin Md., Misra S., Li H. et al. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. // 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Rio de Janeiro, Brazil, 2019. IEEE, 2019. P. 314–324.

6. Li S., Ajimura M., Chen Z. et al. A new approach for comprehensively describing heterogametic sex chromosomes // DNA Research. 2018. Vol. 25 (4). P. 375–382.

7. Marçais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers // Bioinformatics. 2011. Vol. 27 (6). P. 764–770.