VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name: Nguyen Thi Phuong Thao 2. Sex: Female
3. Date of birth: 07 November 1983 4. Place of birth: Hanoi
5. Admission decision number: 1138/QĐ-CTSV Dated 18th December 2013 by the Rector of UET.
6. Changes in academic process:

- Decision No. 46/QĐ-ĐT, dated 28/01/2016 of the Rector of the University of Engineering and Technology on changing the title of doctoral thesis.

7. Official thesis title: Building ancestral recombination graphs for genome-scale data.
8. Major: Computer Science 9. Code: 9480101.01
10. Supervisors:

Assoc. Prof. Le Sy Vinh
Assoc. Prof. Luong Chi Mai

11. Summary of the **new findings** of the thesis:

i. This thesis proposes a heuristic-based algorithm, called ARG4WG, to build plausible ARG from large whole-genome data sets by using the longest shared end for recombination inference. We implemented and published the source code of ARG4WG at https://github.com/thaontp711/arg4wg. The proposed algorithm has the following notable advantages: (1) ARG4WG is the only algorithm currently able to build a full ARG for thousands of human genomes; (2) ARG4WG is hundreds to thousands times faster than Margarita, one of the most efficient ARG inference methods; (3) The experimental results in applying ARG4WG to an association study between the HBB gene and severe malaria in the Gambia dataset show the ability of the proposed algorithm in dissecting association signals for whole-genome association study from large data sets.

ii. We proposed two additional improvements to ARG4WG algorithm to minimize the ARG: (1) The REARG algorithm combines the longest shared end strategy with different criteria, i.e., the maximum similarity between sequences and the length of the sequence; (2) The GAMARG algorithm combines 4-gamete test with the longest shared end strategy in recombination step to reduce the number of recombination events in the ARG building process. Experiments on both simulated and real datasets show that REARG could find out ARGs with smaller number of recombination events than ARG4WG for medium and large datasets. The GAMARG is more general and help find ARGs closer to the optimal solutions.

12. Practical applicability: The proposed methods and software serve as new approaches and tools to help scientists (both theoretical and experimental researchers) analyze and apply ARG efficiently to many practical problems such as association mapping, imputing/phasing/calling SNPs from genome-wide data.

13. Further research directions:

i. Combining combinatorial optimization techniques with GAMARG to build minimal ARGs.

ii. Developing new applications of ARG4WG and GAMARG for imputing SNPs from genome-wide data.

14. Thesis-related publications:

1. Nguyen, T. T. P., Le, V. S., Ho, H. B., & Le, Q. S. (2016), "Building ancestral recombination graphs for whole genomes", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *14*(2), 478-483. (SCIE, IF=2.428).
2. Nguyen, T. T. P., Le, V. S. (2017), "Building minimum recombination ancestral recombination graphs for whole genomes", *The 4th NAFOSTED Conference on Information and Computer Science 2017*, pp. 248-253. (IEEE conference).
3. Nguyen, T. T. P., Le, V. S. (2019), "A Hybrid Approach to Optimize the Number of Recombinations in Ancestral Recombination Graphs", *In Proceedings of the* 2019 9th *International Conference on Bioscience, Biochemistry and Bioinformatics, pp. 36-42*. (ACM conference).