Building ancestral recombination graphs for whole genomes

Thao Thi Phuong Nguyen[1], Vinh Sy Le[2*], Hai Bich Ho[1], Quang Si Le[3*]

[1] Institute of Information Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam,
[2] VNU University of Technology and Engineering, 144 Xuan Thuy, Cau Giay, Ha Noi, Vietnam,
[3] School of Pharmacy and Biomedical Sciences, University of Portsmouth, Winston Churchill Avenue Portsmouth, United Kingdom, PO1 2UP

Introduction

Ancestral Recombination Graph (ARG) plays an important role in human population genetics. Nevertheless, most of current ARG inference algorithms are only applicable to small data sets due to their computational burden. Margarita by Minichiello and Durbin [1] can handle larger data sets; however, it is still not feasible at genome scale. We hereby propose a heuristic algorithm, called ARG4WG, to construct plausible ARGs from thousands of whole chromosome samples, in which the so-called *longest shared end, i.e.* the longest match between left or right ends of sequences, is used for recombination in the building process. This strategy not only allows ARG4WG to significantly reduce the computational cost, by hundreds to thousands times faster than Margarita but also leads to ARGs with fewer number of recombination events. The ARGs resulted from our algorithm also perform reasonably well in association study with 5560 haplotypes across whole chromosome 11 of Gambia dataset. These results indicate that ARG4WG is a good candidate for genome-wide association study from large data sets.

Method

ARG4WG works backward in time from a set of sequences (haplotypes) until reaching a single common ancestor to build an ARG. It includes 3 steps: *Coalescence*, *Mutation* and *Recombination*.

At first, we look for identical sequences to make coalescences. This step reduces the number of sequences until reaching a single common ancestor. In the mutation step, we search for singleton markers, i.e. minor alleles which are then converted into major alleles. This might result in identical sequences which are fed back to the coalescence step. A mutation step can be considered as removing a mutation from the ARG or moving back to the state before a mutation.

In the recombination step we seek a sequence pair (S_1, S_2) with the longest shared end. Assuming that S_1 contains less ancestral material in its shared end part than S_2, we perform a recombination event by breaking S_1 into two new subsequences. The subsequence containing the shared region will be coalesced with S_2 (Figure 1). The recombination step does not increase the number of sequences in the building process.

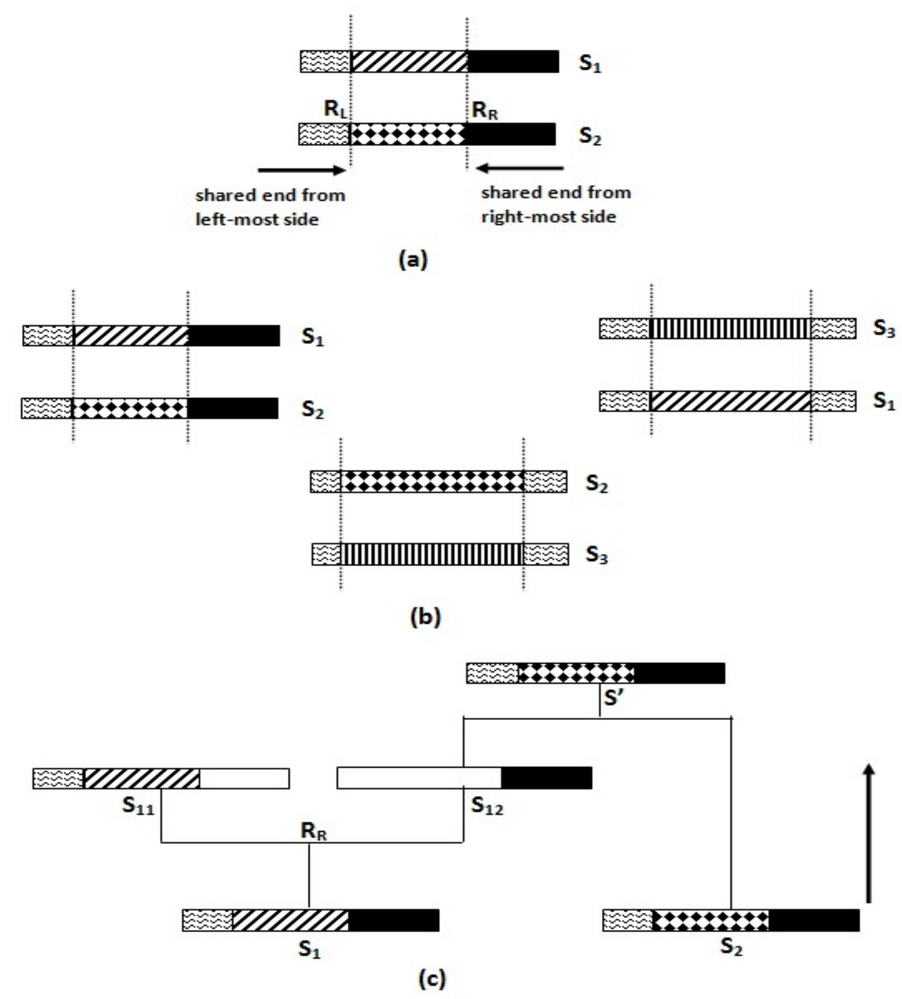

Fig. 1. Recombination event presented in the ARG4WG algorithm. (a) Consider two sequences S_1 and S_2, the shared ends of two sequences from the left-most side (wave pattern) and from the right-most side (black color) are determined. (b) For a set of three sequences S_1, S_2 and S_3, the shared ends of each sequence pair is calculated (wave pattern) and the longest shared end is determined in black color. (c) A recombination event is putted on sequence S_1 to produce two subsequences S_{11} and S_{12}. S_{12} containing the longest shared end will then be coalesced with S_2. Thus, only one recombination event is required and the number of sequences is not increased in the building process.

References

1. M. Minichiello and R. Durbin, "Mapping trait loci using inferred ancestral recombination graphs," *Am. J. Hum. Genet.*, vol. 79, pp. 910–922, 2006.
2. G. P. CONSORTIUM et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.
3. G. K. Chen, P. Marjoram, and J. D. Wall, "Fast and flexible simulation of dna sequence data," *Genome Research*, vol. 19, no. 1, pp. 136–142, 2009.
4. G. BAND *et al.*, "Imputation-based meta-analysis of severe malaria in three African populations," *PLoS genetics*, vol. 9, no. 5, p. e1003509, 2013.
5. O. Delaneau, J. Zagury, and J. Marchini, "Improved whole chromosome phasing for disease and population genetic studies," *Nature methods*, vol. 10, pp. 5–6, 2013.

Acknowledgment

Nguyen Thi Phuong Thao and Ho Bich Hai are supported by Vietnam Academy of Science and Technology (VAST.ĐLT.03/14-15). Le Sy Vinh and Le Si Quang were supported by Vietnam National Foundation for Science and Technology (102.01-2013.04). Computations were conducted in support of Centre for Informatics Computing (VAST).

Results

Results for Real Data

We compared the runtime and the number of recombination events on real data sets from the 1kGP [2] of 500, 1000, and 2000 haplotypes with 1000, 2000, 5000 and 10000 SNPs. For each condition, 3 independent tests, each corresponding to one ARG, were analysed on three different regions of Chromosome 1. The average of the running time and recombination events were then calculated for each condition. ARG4WG leads to ARGs with fewer number of recombination events in much less time when compared with ones by Margarita.

The number of recombination events from Margarita is much higher, 1.4 times that of ARG4WG in all tests (Figure 5). ARG4WG is up to thousands of times faster than Margarita (Figure 4). It took Margarita infeasibly long with 2000 sequences and 10,000 SNPs while ARG4WG only 466 seconds on average.

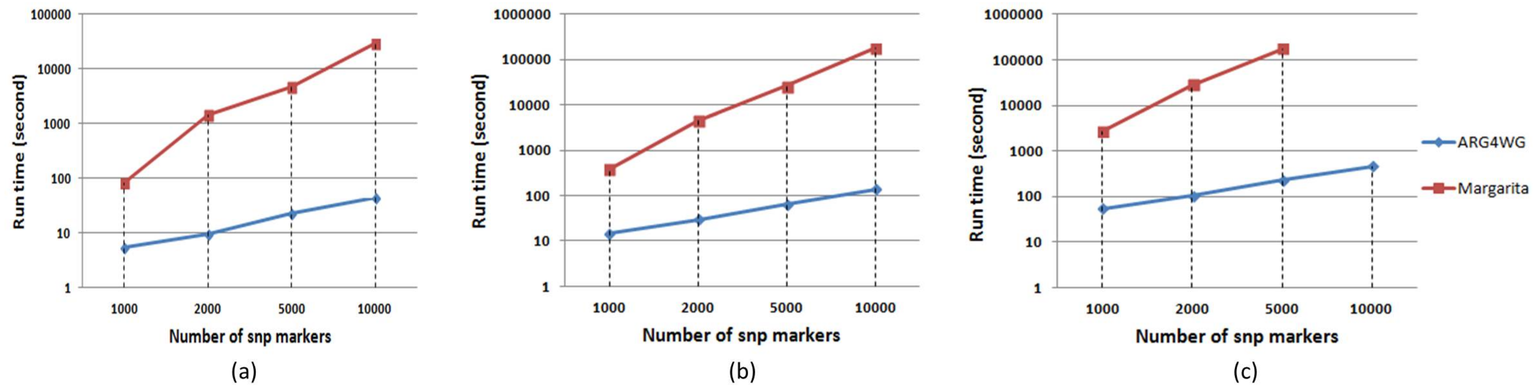

Fig. 4. Average of runtimes of Margarita and ARG4WG for: (a) 500 haplotypes; (b) 1000 haplotypes; and (c) 2000 haplotypes.

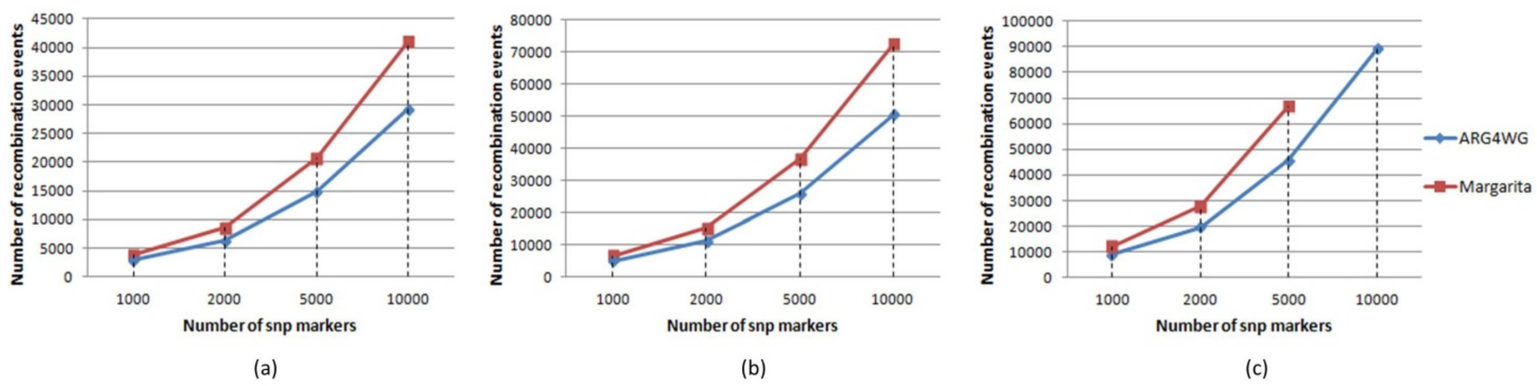

Fig. 5. Average of the number of recombination events of Margarita and ARG4WG for: (a) 500 haplotypes; (b) 1000 haplotypes; and (c) 2000 haplotypes.

We also assessed the performance of ARG4WG on 4246 haplotypes (2123 samples) across the whole Chromosome 1 (174,234 SNPs) from the 1kGP. ARG4WG can infer an ARG in a single run in 4.5 hours using a 16-thread CPU.

Results for Simulated Data

We used MacS [3] to simulate 500 sequences on a region 1Mbp long with an effective population size of 15000. The mutation rate was set to 1.2×10^{-8} per site per generation. We create four different data sets with the mutation to recombination rate ratios of 1, 2, 4 and 6, respectively. For each data set, we selected around 800 SNPs which were closest to the real SNP sites on a 1Mbp region of Chromosome 1 (167,000,000-168,000,000) in 1kGP to make the simulation more realistic and reasonably large. Twenty ARGs were inferred by Margarita and ARG4WG each. The marginal trees at SNP sites of ARGs inferred by both algorithms were extracted for comparison. We compared the tree topologies inferred by both algorithms to the true trees at corresponding positions using Robinson-Fould (RF) distance. RF distance is calculated as the number of bi-partitions in one of the two trees but not the other, divided by the number of possible bi-partitions.

The averaged RF distances across a range of mutation to recombination rate ratios are shown in Figure 6. The RF distances of Margarita are slightly better than that of ARG4WG. This difference decreases with the increase of mutation to recombination rate. At rate 6, RF distances in both cases are almost the same.

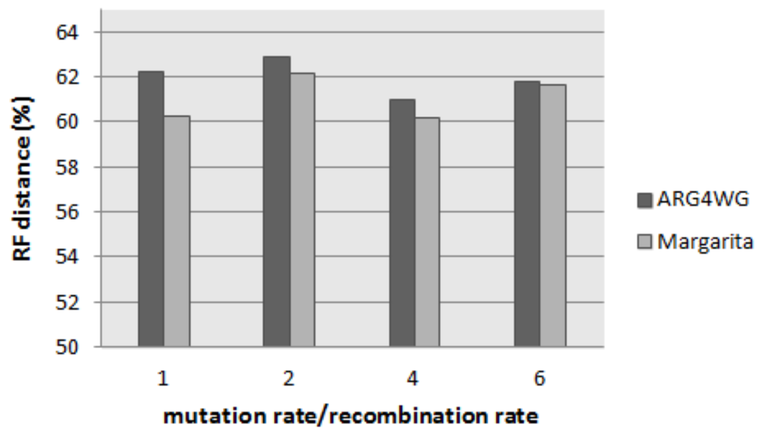

Fig. 6. The Robinson-Foulds distances of trees inferred by Margarita and ARG4WG in comparison with simulation trees over a range of ratios of mutation to recombination rates.

Application of ARG4WG in association mapping

We examined the quality of ARGs constructed from ARG4WG in association mapping on large GWAS data sets. ARG4WG was used to build ARGs from the Gambia dataset consisting of 5560 haplotypes (2780 samples, 1533 controls, 1247 cases) across the whole Chromosome 11 [4]. SHAPEIT [5] was used to phase genotyped SNPs of the Gambia data. It took AGR4WG about 3.1 hours to build an ARG using our 16-thread CPU.

We investigated the association between polymorphisms in the region of the HBB gene (4.5 Mb to 5.5 Mb of Chromosome 11) and severe malaria in the above dataset. The ARGs and extracted marginal trees from ARG4WG were input to Margarita for association mapping. We performed the mapping test with 10^6 permutations in three scenarios of different numbers of SNPs (Figure 7). All three settings detect the same strong disease association within the region from 4.43Mb to 6.28Mb on Chromosome 11 with p-values $\leq 10^{-7}$. This result agrees with the analysis of Garvin Band et al. [4], that the HBB region shows the strong evidence of association with severe malaria.

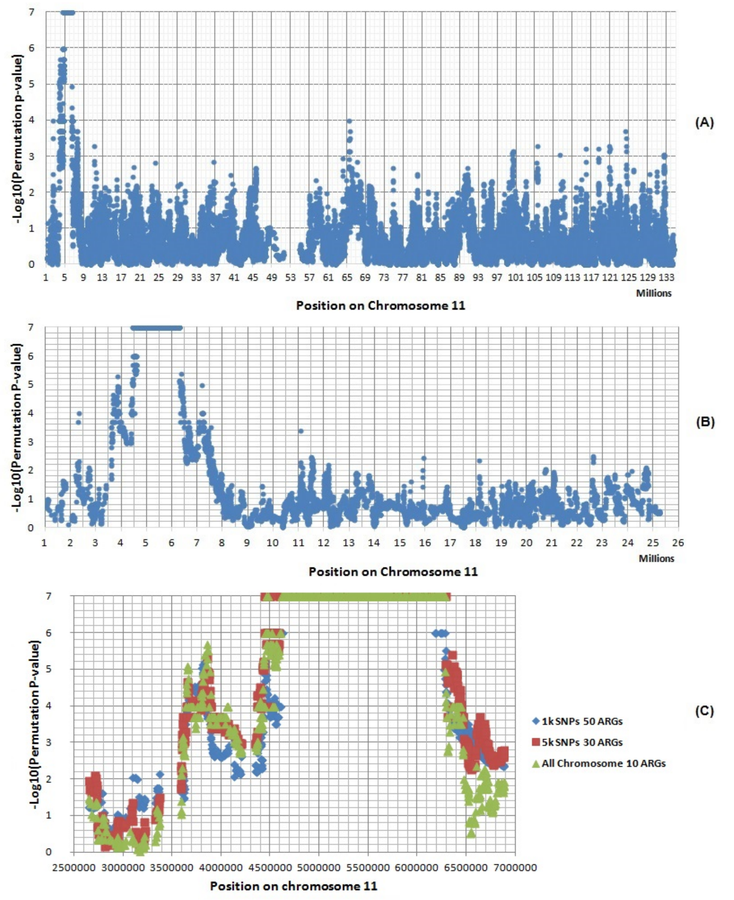

Fig. 7: Association structure of regions of the Gambia data from 10^6 permutation tests on: (A) 10 ARGs inferred across whole chromosome 11; (B) 30 ARGs inferred on the region of 5000 SNPs around HBB gene; and (C) A summary of association structures for all experiments on the region of 1000 SNPs around HBB gene.

Conclusion

We propose a heuristic algorithm, called ARG4WG, to build plausible ancestral recombination graphs (ARGs) from thousands of whole genome samples. By using the *longest shared end* for recombination inference, ARG4WG constructs ARGs with small numbers of recombination events.

The results from the Gambia dataset demonstrate that ARGs constructed by ARG4WG are good for whole-genome association studies. From those promising results, we believe that ARG4WG could be applied to other practical problems such as imputing, phasing, SNP calling and so forth from genome-wide data.