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 recombination events. 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 algorithms are almost the same.

ARG4WG works backward in time from a set of sequences (haplotypes) until reaching a unique most recent 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 (S1, S2) with the longest shared end. Assuming that S contains less ancestral material in its shared end part than S2, we perform a recombination event by breaking S into two new subsequences. The con董事长ing sequence coalesces with S2 (Figure 1). The recombination step does not increase the number of sequences in the building process.

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 averaged run 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 hundreds of times faster than Margarita (Figure 4). It took Margarita infeasibly long with 2000 sequences and 10,000 SNPs whilst ARG4WG only 466 seconds on average.

ARG4WG works backward in time from a set of sequences (haplotypes) until reaching a unique most recent 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 (S1, S2) with the longest shared end. Assuming that S contains less ancestral material in its shared end part than S2, we perform a recombination event by breaking S into two new subsequences. The con董事长ing sequence coalesces with S2 (Figure 1). The recombination step does not increase the number of sequences in the building process.

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

We used Macs [3] to simulate 500 sequences on a region 1Mb 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 1Mb region of Chromosome 1 (167,000-168,000) in the 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-Foulds (RF) distance.

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.

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 dataset. It took ARG4WG about 3.1 hours to build ARGs from the Gambia dataset.

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 permutations in three scenarios of different numbers of SNPs (Figure 7). All three settings detected the same strong disease association within the region from 4.43Mb to 6.22Mb on Chromosome 11 with p-values ≤ 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.