BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 23 2007, pages 3254–3255
doi:10.1093/bioinformatics/btm426

Genetics and population analysis

TAGster: efficient selection of LD tag SNPs in single or multiple populations

Zongli Xu[1], Norman L. Kaplan[2] and Jack A. Taylor[1,3,*]

[1]Epidemiology Branch, [2]Biostatistics Branch and [3]Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA

Received on June 5, 2007; revised on July 23, 2007; accepted on August 15, 2007

Advance Access publication September 7, 2007

Associate Editor: Martin Bishop

ABSTRACT

Summary: Genetic association studies increasingly rely on the use of linkage disequilibrium (LD) tag SNPs to reduce genotyping costs. We developed a software package *TAGster* to select, evaluate and visualize LD tag SNPs both for single and multiple populations. We implement several strategies to improve the efficiency of current LD tag SNP selection algorithms: (1) we modify the tag SNP selection procedure of Carlson *et al*. to improve selection efficiency and further generalize it to multiple populations. (2) We propose a redundant SNP elimination step to speed up the exhaustive tag SNP search algorithm proposed by Qin *et al*. (3) We present an additional multiple population tag SNP selection algorithm based on the framework of Howie *et al*., but using our modified exhaustive search procedure. We evaluate these methods using resequenced candidate gene data from the Environmental Genome Project and show improvements in both computational and tagging efficiency.

Availability: The software Package *TAGster* is freely available at http://www.niehs.nih.gov/research/resources/software/tagster/

Contact: taylor@niehs.nih.gov

Supplementary information: Additional information, including a tutorial, detailed algorithm and detailed evaluation results, is also available from *TAGster* web site (see above).

1 INTRODUCTION

Genotype data are now available for millions of SNPs from the International HapMap project (The International HapMap Consortium, 2005) and from many gene resequencing projects. Although genotyping technology is rapidly advancing, it is not yet cost effective for genetic association studies to genotype all available SNPs. Use of linkage disequilibrium (LD) tag SNPs can dramatically reduce genotyping costs, but the selection of a minimal set of tag SNPs can be challenging, particularly when studying multiple populations that have different LD structure. Here we describe a new software tool *TAGster* that selects, evaluates and visualizes LD tag SNPs both for single and multiple populations.

*To whom correspondence should be addressed.

2 METHODS AND RESULTS

2.1 Genotype data

We evaluated the software using Environmental Genome Project Panel 2 data for 207 genes that were resequenced in 95 DNA samples from 4 populations (27 Africans, 24 Asians, 22 Europeans and 22 Hispanics) (http://egp.gs.washington.edu/). There were a total of 16153 SNPs with minor allele frequency (MAF) ≥ 0.05 in at least one population. Within each population we calculate r^2 for all possible pairs of SNPs within each gene. Two SNPs are said to be in high LD if r^2 exceeds a specified threshold (e.g. $r^2 \geq 0.8$).

2.2. Algorithm 1: a greedy algorithm for single or multiple populations

We refined the greedy algorithm proposed by Carlson *et al*. (2004). In the original algorithm, a tag SNP is identified and the subset (bin) of SNPs that are in high LD with the tag are removed from further consideration. Instead, in *TAGster*, the binned SNPs are retained as potential tag SNP candidates for subsequent iterations. Specifically, our modified procedure has the following steps (see Supplementary Material for details).

(1) For each SNP that is not already selected as a tag, we count the number of as yet unbinned SNPs that are in high LD with the SNP.

(2) The SNP with the largest count is selected as a tag SNP.

(3) Unbinned SNPs in high LD with the tag SNP are placed into a bin.

The three steps are iterated until the maximal count in step (2) is 1. All the remaining unbinned SNPs are declared as singleton tag SNPs.

Evaluation in EGP Panel 2 data at r^2 threshold of 0.8 showed that the modified greedy algorithm selected l42 fewer tag SNPs than the greedy algorithm as implemented in *ldSelect* (Carlson *et al*., 2004). For 62 genes the modified greedy algorithm selected fewer tags in at least one of the four populations, whereas the original greedy algorithm selected fewer tag SNPs in only two genes in one population.

Similar to Xu *et al*. (2007), we further generalized the modified greedy algorithm to select a single set of tag SNPs for multiple populations by performing step one in each population-specific group independently, summing the SNP counts across populations and selecting as a tag, the SNP with the maximum sum. This algorithm does not require that the different population groups start with the same set of SNPs. Furthermore, LD patterns may vary between populations so that a multi-population tag SNP may capture different sets of SNPs in different populations.

2.3. Algorithm 2: an optimal solution for single population tag SNPs

Instead of using a greedy search algorithm, one may exhaustively search for the minimum number of tag SNPs. Qin *et al.* (2006) proposed a comprehensive search algorithm by partitioning all SNPs within a genome region into disjoint precincts such that SNPs in one precinct are not in high LD with SNPs in any other precinct. An exhaustive search can then be carried out in each precinct. We further modified the algorithm as outlined in the following steps (see Supplementary Material for details).

(1) If two SNPs in a precinct have the same high/low LD relationship with all other SNPs in the precinct, we retain only one of the SNPs.

(2) We exhaustively search for the minimum number of tag SNPs in each precinct. If the search in a precinct exceeds a specified number of steps without finding a solution, then Algorithm 1 is used to find tag SNPs for the precinct.

Depending on the complexity of LD structure, this modification can substantially speed up the search algorithm. For example, we compared our algorithm to the comprehensive search algorithm implemented in *FESTA* (Qin *et al.*, 2006) using a LD threshold of 0.8 and an exhaustive search limit of 1 000 000 (default setting in *FESTA*) for both algorithms. Using African data on 207 genes from EGP Panel 2 with a 2.8 GHz Pentium personal computer, our algorithm took 498 s, and required the use of the greedy algorithm once. Conversely *FESTA* took 9307 s (19-fold more time) for the computation, and required the use of the greedy algorithm six times. Even larger differences in computational speed were seen for other populations (see Supplementary Material for detail).

2.4 Algorithm 3: a two-stage solution for multiple populations

We implemented a two-stage solution for the selection of a single set of tag SNPs for multiple populations. Exhaustive searches were employed to select a minimal number of tag SNPs within each stage. At the first stage, we employed Algorithm 2 to select a minimal number of tag SNPs for each ethnic group and for each of these tag SNPs we list those SNPs within the associated LD bin that could function as alternative tag SNPs. In the second stage, we execute the following steps (see Supplementary Material for details):

(1) Similar to Howie *et al.* (2006) we cluster the listed SNPs (see details in Supplementary Material).

(2) For each cluster we select the SNP that tags bins in the largest number of populations.

(3) We then group the selected SNPs if they tag the same bin in at least one of the populations.

(4) We perform an exhaustive search within each group to find the minimum number of tag SNPs.

We applied both Algorithms 1 and 3 to select multi-population tag SNPs in 207 genes for four populations from the EGP. As a benchmark measure, we used the total number of tag SNPs found using *ldSelect* followed by *MultiPop-TagSelect* (Howie *et al.*, 2006). Using Algorithm 1, the benchmark number is reduced by 183, whereas using Algorithm 3, the number is reduced by 159. For each gene, *TAGster* selects the smaller number of tag SNPs of these two algorithms, thereby reducing the number of tag SNPs by 233.

DISCUSSION

We implemented three improved methods of tag SNP selection into the software package *TAGster*. For EGP Panel 2 data, these methods show improvements in both computational and tagging efficiency over the alternatives. We also found gains in efficiency when we applied these methods to HapMap ENCODE data (http://www.hapmap.org, see Supplementary Material).

The program provides a number of selectable features and graphical output to assist investigators in tag SNP selection. For phase-unknown data, *TAGster* can calculate the measure of composite linkage disequilibrium proposed by Weir (1979), which unlike r^2, does not require an assumption of random mating. *TAGster* allows investigators to specify high-interest SNPs (e.g. nsSNPs) as a set of *a priori* tag SNPs. Moreover, investigators have an option of including their own user-provided scores for tag SNP preference, e.g. SNP design scores, which can be used in tag SNP selection. *TAGster* can utilize both HapMap and gene resequencing data directly for tag SNP selection. The graphical output has tracks showing LD bins, tag SNPs, nsSNPs, SNP tagging ability and allele frequency information along with LD structure or genotype data for both single and multiple populations.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Conflict of Interest: none declared.

REFERENCES

Carlson,C.S. *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Howie,B.N *et al.* (2006) Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum. Genet.*, **120**, 58–68.

Qin,Z.S. *et al.* (2006) An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Bioinformatics*, **22**, 220–225.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Weir,B.S. (1979) Inferences about linkage disequilibrium. *Biometrics*, **35**, 235–254.

Xu,Z. *et al.* (2007) LD tag SNP selection for candidate gene association studies using HapMap and gene resequencing data. *Eur. J. Hum. Genet.*, **15**, 1063–1070.