

# EFFICIENT ALGORITHMS FOR GENOME-WIDE TAGSNP SELECTION ACROSS POPULATIONS VIA THE LINKAGE DISEQUILIBRIUM CRITERION

Lan Liu, Yonghui Wu, Stefano Lonardi and Tao Jiang\*

Department of Computer Science and Engineering, University of California, Riverside, CA 92507, USA

\*Email:jiang@cs.ucr.edu

In this paper, we study the tagSNP selection problem on multiple populations using the pairwise  $r^2$  linkage disequilibrium criterion. We propose a novel combinatorial optimization model for the tagSNP selection problem, called the *minimum common tagSNP selection* (MCTS) problem, and present efficient solutions for MCTS. Our approach consists of three main steps including (i) partitioning the SNP markers into small disjoint components, (ii) applying some data reduction rules to simplify the problem, and (iii) applying either a fast greedy algorithm or a Lagrangian relaxation algorithm to solve the remaining (general) MCTS. These algorithms also provide lower bounds on tagging (*i.e.* the minimum number of tagSNPs needed). The lower bounds allow us to evaluate how far our solution is from the optimum. To the best of our knowledge, it is the first time tagging lower bounds are discussed in the literature. We assess the performance of our algorithms on real HapMap data for genome-wide tagging. The experiments demonstrate that our algorithms run 3 to 4 orders of magnitude faster than the existing single-population tagging programs like FESTA, LD-Select and the multiple-population tagging method MultiPop-TagSelect. Our method also greatly reduces the required tagSNPs compared to LD-Select on a single population and MultiPop-TagSelect on multiple populations. Moreover, the numbers of tagSNPs selected by our algorithms are almost optimal since they are very close to the corresponding lower bounds obtained by our method.

## 1. INTRODUCTION

The rapid development of high-throughput genotyping technologies has recently enabled genome-wide association studies to detect connections between genetic variants and human diseases. *Single-nucleotide polymorphism* (SNP) is the most frequent form of polymorphism in the human genome. Common SNPs with *minor-allele frequency* (MAF) of 5% have been estimated to occur once every  $\sim 600$  bps<sup>18</sup>, and there are more than 10 million verified SNPs in dbSNP<sup>11</sup>. Given these numbers, it is currently infeasible to consider all the available SNPs to carry out association studies. This motivates the selection of a *subset* of informative SNPs, called *tagSNPs*.

The selection of tagSNPs *in silico* is a well-studied research topic. Existing computational methods for tagSNP selection can be classified into two categories: *haplotype-based* methods<sup>1, 12, 17, 19, 24, 28, 31, 32, 34</sup> and *haplotype-independent* methods<sup>5, 15, 16, 20–22, 25, 27, 26</sup>. The haplotype-based methods require phased multi-locus haplotypes, whereas the haplotype-independent methods do not require haplotype information. The main shortcoming of haplotype-based methods is that the pre-processing step (*i.e.* the inference of haplotypes from genotypes) is computationally demanding. In addition, since there is not an authoritative inference method, the haplotypes generated by the existing haplotype inference methods are often quite different<sup>7, 32, 35</sup>. Consequently, the tagSNPs selected by the haplotype-based methods would be quite different. Recently, Carlson *et al.*<sup>5</sup> proposed a haplotype-independent method that employs the  $r^2$  linkage disequilibrium (LD) statistical criterion to

measure the association between SNPs. The tagSNPs selected by this method are shown to be effective in disease association mapping studies, because the measure  $r^2$  is directly related to the statistical power of association mapping. Because this method has comparable performance at a lower computational cost than many other methods<sup>33, 27</sup>, tagging approaches based on  $r^2$  LD statistics have gained popularity among researchers in the SNP community<sup>2, 5, 8, 22, 26, 33</sup>.

Most approaches using the  $r^2$  criterion require that tagSNPs be defined within a single population, because LD patterns (see the caption of Figure 1(A) for a definition) are quite susceptible to population stratification<sup>5</sup>. In two populations with different evolutionary histories, a pair of SNPs having remarkably different allele frequencies and very weak LD may show strong LD in the admixed population (see such an example in Table 1). Recent study<sup>6</sup> shows that the LD patterns and allele frequencies across populations are very different<sup>6, 29</sup> in fact. For example, among the populations collected in the HapMap project (*i.e.* YRI, CEU, CHB and JPT), 81% of the SNPs in YRI population have a near perfect proxy (*i.e.* SNPs that have  $r^2 \geq 0.8$  with other SNPs), while in the other three populations, 91% of the SNPs have a near perfect proxy<sup>9</sup>. Therefore, tagSNPs picked from the combined populations or one of the populations might not be sufficient to capture the variations in all populations. In order to maintain the power of association mapping, we need generate a common (or universal) tagSNP set to type all the populations with sufficient accuracy.

A simple approach to select a universal tagSNP set is to tag one population first and then select a supplementary set for each of the other populations one by one<sup>2, 23, 22</sup>.

\*Corresponding author.





















