An Unsupervised Diplotype Clustering Method to Improve Race-Based Medicine

Junji Tanaka\(^1\) \quad Masato Inoue\(^2\) \quad Naoyuki Kamatani\(^3\)

\footnotesize{jtanaka@genesys-tech.co.jp \quad masato.inoue@eb.waseda.ac.jp}

\footnotesize{\(^1\) Genesys Technologies, Inc., 5th fl., Nihonbashi Center Bldg., 1-4-12, Nihonbashi Hon-cho, Chuoh-ku, 103-0023 Tokyo, Japan
\(^2\) Department of Electrical Engineering and Bioscience, School of Science and Engineering, Waseda University, 169-8555 Tokyo, Japan
\(^3\) Division of Genomic Medicine, Department of Applied Biomedical Engineering and Science and Institute of Rheumatology, Tokyo Women’s Medical University, 162-8666 Tokyo, Japan

Keywords: race-based medicine, haplotype inference, population structure, multilocus SNP

1 Introduction

In recent years, race-based medicine has become increasingly important [1]. We consider the categorization of patients so that each group can be provided with proper medication a reasonable approach, and a thorough application of this approach would ultimately lead to personalized medicine. However, we question whether race is the best criterion for such categorization.

In this paper, we propose a method for clustering subjects based on their diplotype of interest. We use multilocus single nucleotide polymorphism (SNP) data, because SNP data are rather common and less expensive to obtain than genealogical data or phased data. This availability is part of a trade-off, though, since we have to infer the diplotype (both haplotypes) of each subject [2]. We have extended the haplotype inference method, which uses the expectation-maximization (EM) algorithm [3], to perform diplotype clustering. The Markov-chain Monte Carlo method also provides good clustering [4], but the calculation cost and amount of required memory are usually huge.

2 Method

We focus on completely linked \( N \) polymorphic sites of SNPs and unphased non-genealogical data. The data of \( i \)-th subject, \( x_i \), consist of all the genotypes of \( N \) polymorphic sites, where \( i = 1, ..., I \). Using this dataset, we infer most probable number of genetic populations, \( K \), and corresponding haplotype frequencies \( \{ b_k \}_{k=1,...,K} \), where \( b_k \triangleq [b_{k,1}, ..., b_{k,H}] \), and \( \sum_k b_{k,H} = 1 \) and \( H \) denotes the number of possible haplotypes. Because this inference problem includes two types of unobservable random variables, the true diplotype information of each subject \( d_i \triangleq [d_{i,1}, d_{i,2}] \) and the true population identification \( k_i \) each subject originated from, we solve this problem by the EM algorithm, roughly as

\[
\{ \hat{K}, \{ \hat{b}_k \} \} = \arg \max_{K, \{ b_k \}} \left\langle \ln P \left( \{ x_i \}, \{ d_i \}, \{ k_i \} \mid \{ b_k \} \right) \right\rangle_{P \left( \{ d_i \}, \{ k_i \} \mid \{ x_i \}, \{ b_k \} \right)}
\]

where the hat denotes inferred variables, and \( \langle \bullet \rangle_o \) denotes the expectation of \( \bullet \) with respect to \( o \). \( \hat{k}_i \) is also given using \( \hat{K} \) and \( \{ \hat{b}_k \} \). Moreover, through approximation, we reduce the memory required for this procedure from \( \mathcal{O}(KIH) \) to \( \mathcal{O}(IH) \). This method can be applied recursively to each sub-group.
Figure 1: Validation results of the clustering ability. Each boxed number denotes the size of each original group. Solid circles denote the application of the method, and the solid lines denote the generated sub-groups. Dashed lines show where the method could not further divide a sub-group.

Figure 2: We reused the dataset from Fig. 1(a)-(c), but slightly modified it so that several original groups formed population A and the rest of the original groups formed population B by mixing one of the haplotypes of each subject randomly within each population.

3 Validation

First, we validated the clustering ability of the method. Specifically, we generated artificial data to ensure every subject had no heterogeneous alleles. When the group sizes were almost equal (Fig. 1(a)), the first application of this method successfully separated the whole population into proper groups, and no original group was separated into sub-groups by the method. If the group sizes were moderately different, this method determined that some hierarchical structures were more probable than a flat structure (Fig. 1(b) and (c)). If the group sizes were considerably different, the proposed method did not separate subjects even if their haplotypes were completely different (Fig. 1(d)). We welcome this result because it shows that the method can provide rather robust results which are not affected by small rare populations. Next, we validated the proposed method including diplotype inference. Specifically, we modified the dataset used in Fig. 1(a)-(c) so that the dataset had two genetic populations (A and B). The method successfully clustered the subjects into the two populations, and did not divide any single population into sub-groups (Fig. 2).

References