中文摘要

在生物科技快速發展的現代，許多人嘗試要透過遺傳相關的研究找出疾病發生的先天個人因素，用以解釋發病原因在類似環境因素下仍存在的人與人之間的差異。在常見的病例對照研究 (case-control study) 中，遺傳性相關研究 (genetic association study) 會透過檢視病例組 (case) 與對照組 (control) 的單一核苷酸多型性 (single nucleotide polymorphism, SNP) 的差異來探討與疾病有關的基因標記 (marker) 或致病基因 (causal SNP)，試圖找出可能與發病有關的遺傳位點。針對這類型的研究，目前統計方法著重於發展標記集之檢定方法 (marker-set analysis)，如：單一核苷酸多型性集合之分析 (SNP-set analysis)，因此類型方法的檢定力 (power) 較好，且又可同時考慮一群單一核苷酸多型性彼此之間的交互作用。而目前這類型的方法，多是聚焦於事前已經決定好之集合，或是利用滑動視窗 (sliding window) 的方式將整個染色體分成小區段進行檢定，尚無在先驗資訊未知之下先定義出集合再進行檢定之方法。因此，在本研究中，將以單一核苷酸多型性之集合的角度出發，先透過群聚分析 (cluster analysis) 定義出單一核苷酸多型性之集合，接著再利用此集合進行遺傳相關性檢定。

在研究者沒有任何先驗知識 (prior knowledge) 之下，我們提出一種利用漢明距離 (Hamming distance) 為單一核苷酸多型性之間相似度測量 (similarity measure)

的階層式分群演算法 (hierarchical clustering algorithm)，配合最大差異性樹狀圖 (dendrogram) 節點 (node) 的篩選，定義出單一核苷酸多型性之集合；之後，一樣利用漢明距離，比較病例組與對照組間在此單一核苷酸多型性之集合上的差異性，利用此差異在組內 (within group) 與組間 (between group) 分佈的不同創造出新的統計量進行統計檢定 (Hamming distance-based association test, HDAT)，針對常見變異 (common variants) 與罕見變異 (rare variants) 的特性差異，本研究中也分別提出兩種不同的檢定統計量用以進行遺傳相關性檢定，此統計量為 U-statistic 的一種，本研究中也推導出其相關之統計性質與大樣本理論。

在實際資料的應用上，所提出的分群演算法可以正確定義出恰當的分群結果，且與其他方法相較，是較有效率的演算法，可以花費較少的時間得到分群的結果，此方法不僅適用於遺傳的資料，也可應用於類別型態的資料 (categorical data)；從一些模擬實驗來看，所提出的分群演算法配合最大差異性樹狀圖節點的篩選，可以正確地將有相關性的單一核苷酸多型性 (correlated SNPs) 分成一群，而排除沒有相關的單一核苷酸多型性。針對 HDAT，在常見變異的部分，從一些模擬結果也可看出，不論訊雜比 (signal-to-noise ratio) 為多少（與疾病有關的單一核苷酸多型性和與疾病無關的單一核苷酸多型性的比例）、樣本數大小為何、集合中的單一核苷酸多型性是否對疾病具有一致的影響力 (effect) 都可以有不錯的檢定力 (power)，型一誤差 (type I error) 也可以控制在一定的範圍內；此外，也針對 WTCCC 研究中的冠狀動脈心臟病 (Coronary artery disease, CAD) 的資料進行分析，先進行群聚分析後再進行相關性檢定，可以找出一組單一核苷酸多型性之集合與疾病有關，此集合中的四個單一核苷酸多型性也曾在文獻中被提到與冠狀動脈心臟病有相關。在罕見變異的部分，從模擬的結果也可看出，不論訊雜比

(signal-to-noise ratio) 為多少、樣本數大小為何、病例組與對照組樣本數比例為何都可以有不錯的檢定力，型一誤差也可以控制在一定的範圍內。

本研究中所提出的分群演算法可以找出單一核苷酸多型性潛在的群聚特徵，並依照此分群結果進行 HDAT 相關性檢定能得到更佳的檢定力；不論從模擬研究的結果或實際資料的應用上來看，所提出的方法都可以有不錯且穩定的表現。此外，本研究同時針對 Hamming distance 的統計性質進行進一步的討論。然而，本研究中所提出的 HDAT 在分析常見變異時，若病例組與對照組樣本數不相等，表現較不如其他統計方法，針對此問題，檢定統計量需要進行一些改良；而如何將其他非類別型態的疾病相關因子納入考慮，也是另一個重要的議題。

關鍵字：遺傳相關性檢定、群聚分析、常見變異、樹狀圖、漢明距離、罕見變異、相似度、單一核苷酸多型性之集合。

Abstract

With the advance in biotechnology, many researchers try to identify disease-associated markers through genetic association studies. In recent genetic association studies, developing methods to reduce intractably large numbers of genetic variants in genomic data to more computationally manageable numbers and finding ways to increase the power of statistical tests used in association studies have been two major challenges. Tackling these problems with a marker-set study such as SNP-set analysis can be an efficient solution. Such method can also evaluate joint effect of grouped SNPs in a pre-specified genomic region. Nowadays, most association tests, however, figure out possible marker sets based on testing pre-specified SNP sets or testing through sliding window for whole genome. It seems that no combined procedure to define SNP sets in advance than to test association between SNP sets and the disease of interest.

To construct SNP sets, we first propose a clustering algorithm, which employs Hamming distance to measure the similarity between strings of SNP genotypes and evaluates whether the given SNPs should be clustered. We also recommend a rule-of-thumb to determine the number of clusters after a dendrogram is produced. With the SNP sets obtained, we next develop an association test to examine susceptibility to the disease of interest. For common variants, this proposed test assesses, based on Hamming distance, whether the similarity in genotypes between a diseased and a normal individual differs from the similarity between two individuals with the same disease status. For rare variants, the proposed test evaluates whether the similarity in genotypes within the case group differs from the similarity within the control group. These two statistics are U-statistics, and their statistical properties and limiting behaviors are also discussed. Additionally, simulation studies and real data applications were conducted to demonstrate the performance of our proposed methods.

The results showed that the Hamming distance-based clustering algorithm can identify correct clustering patterns and is also an efficient algorithm. This method can be applied not only to genetic data, but also to categorical data in general. Additionally, for common variants, the Hamming distance-based association test (HDAT) works well regardless of the sample size, effects of SNPs within the given set, and the signal-to-noise ratio (proportion of the number of disease-associated SNPs to the number of neutral SNPs). Moreover, for genotyping data of coronary artery disease (CAD) from the WTCCC, our proposed methods found one SNP set with four SNPs were associated with the disease. These four SNPs have been reported in literatures. For rare variants, the numerical results demonstrated that the HDAT works well in spite of the sample size, the case-to-control ratio, and the signal-to-noise ratio.

To conclude, the proposed clustering algorithm and association test are illustrated with simulations and a genome-wide association study, and the results indicate reliable and satisfactory performance. In our proposed methodology, no inference of haplotypes is needed, and SNPs under consideration do not need to be linked. Specifically, this test works well for a SNP-set containing both SNPs with a deleterious effect and those with a protective effect, and for a set containing many neutral SNPs. Moreover, the statistical properties of the proposed methods are discussed. However, some issues remain unsolved. First, for common variants, some extensions of the HDAT to imbalanced sizes of the case and control group need to be studied. Second, even though categorical disease-related factors can be consider as pseudo genetic markers, how to incorporate disease-related factors, such as environmental factors and personal characteristics, still need to be studies.

Keywords: association test, clustering analysis, common variants, dendrogram, Hamming distance, rare variants, similarity, SNP set.