Contents

中文摘要 i

Abstract v

1 Introduction 1

1.1 Genetic Association Study 1

1.2 Cluster Analysis for Genetic Sequencing Data 7

1.3 Similarity Measure 10

1.3.1 Distance Metrics for Categorical Data 10

1.3.2 Distance Metrics for Genetic Data 13

1.4 Motivation and Aim of the Study 18

1.5 Data Structure 22

1.6 Real Data Applications 22

2 Statistical Properties of Hamming Distance Statistic 25

2.1 HD Statistic for Clustering 26

2.2 HD statistic for Paired Individuals 29

2.3 Asymptotic Distribution of HDAT 33

3 Clustering SNP-sets with Hamming Distance 41

3.1 Hamming Distance-based Clustering Algorithm 42

3.2 Modified Merging Procedure and Ties 43

3.3 Select SNP Sets for Association Studies 45

3.4 Results 48

3.4.1 Indices for Cluster Validation 48

3.4.2 Simulation Studies 51

3.4.3 Real Data Applications 55

4 SNP-set Association Test via Hamming Distance for CV 71

4.1 Hamming Distance-based Association Test 72

4.2 Results 74

4.2.1 Simulation Studies for Association Tests 74

4.2.2 Simulation Studies for Combined Procedures 79

4.2.3 Real Data Applications 85

5 SNP-set Association Test via Hamming Distance for RV 93

5.1 HDAT for Rare Variants 97

5.2 Numerical Results 98

6 Discussion and Future Work 103

6.1 Other Applications of HD-Cluster 105

6.2 Comparison with LD Pattern and Haplotype Analysis 110

6.3 Future Work 112

Bibliography 114

A Illustration of Selecting SNP Sets 127

B Theory of U-statistic 129

C Proof of Statistical Properties of HDAT 135