MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Effects of Genotype Error and Sample Error on Genome-wide SNP Analysis Using Pooled DNA

Siddharth Pratap[a], Scott M. Williams[c], Shawn E. Levy[a,b]

[a]*Department of Biomedical Informatics, Vanderbilt University, USA*
[b]*Department of Molecular Physiology and Biophysics, Vanderbilt University, USA*
[c]*Center for Human Genetics Research, Vanderbilt University, USA*

Abstract

A major obstacle in whole-genome association analysis is that the cost of individually genotyping hundreds of samples remains prohibitively expensive. One solution which addresses this is to pool genomic DNA of the cases and controls. However, pooling introduces an additional source of error, sampling error. In this study, we evaluate the effects of sampling error as well as genotyping error in pooled versus individual genotyping. Our results show that both genotype and sample error affect the significance of the results. Yet, sampling error has a larger overall effect.

Keywords:

single nucleotide polymorphism (SNP); pooled DNA; genome wide association

Introduction

In this study, we examine the feasibility of a pooling approach for genome wide association studies in the context of two types of experimental errors; genotyping error and sampling error. Genotyping error is present in both individual and pooled genotyping. Sampling error occurs in pooled genotyping when unequal amounts of genomic DNA are added to the pool, which can lead to incorrect allele frequency estimates for the pooled population.

Materials and methods

GenomeSIM simulation software was used to create a case/control population of one thousand individuals typed at ten thousand SNPs each. A three locus disease model with additive penetrance values was used. The final population consisted of 203 diseased individuals and 797 controls. Haploview was used for association testing of individuals. The Pooled DNA Analyzer (PDA) was employed for pooled association testing. Three levels of stringency were chosen with respect to the association testing chi square P-values. A low stringency cutoff (P-value <= $5E^{-2}$), a high stringency cutoff using Bonferonni correction (P-value <= $5E^{-6}$), and a middle stringency cutoff (P-value <= $5E^{-4}$).

Results and discussion

Although both genotype error and sample error had a significant effect, sampling error is much more influential. Overall, genotype error from 1% to 5% lowered the significance of P-values by a maximum of 1.6 fold across the entire range of individual or pooled genotyping. In contrast, sampling error from 1% to 5% in the pooled genotypes had a much more pronounced impact and lowered the significance of P-values by at least one order of magnitude across its range, with minimum of a 66 fold reduction in P-value significance (table 1). We conclude that if pooling is employed as a strategy for genome wide association analysis, great care must be taken to minimize the sampling error introduced while assembling the pools.

Table 1 – Association analysis of Individual and Pooled genotyping: P-values of causal SNPs

Low stringency (P-value <= $5E^{-2}$) in red / Intermediate stringency (P-value <= $5E^{-4}$) in *yellow italics* / High stringency (P-value <= $5E^{-6}$) in **green bold**

			Pooled + sample error				
		individual	**1%**	**2%**	**3%**	**4%**	**5%**
Genotype Error	**0%**	**7.1E-10**	**2.2E-07**	*1.7E-04*	4.4E-03	2.2E-02	5.6E-02
	1%	*4.4E-04*	8.8E-04	3.7E-03	1.4E-02	3.7E-02	7.2E-02
	2%	*3.7E-04*	7.3E-04	3.3E-03	1.3E-02	3.4E-02	6.7E-02
	3%	5.5E-04	9.9E-04	3.7E-03	1.3E-02	3.3E-02	6.6E-02
	4%	*3.7E-04*	7.3E-04	3.3E-03	1.3E-02	3.4E-02	6.7E-02
	5%	6.0E-04	1.1E-03	4.5E-03	1.6E-02	4.2E-02	8.0E-02

Acknowledgments

S. Pratap is supported by the NIH/National Library of Medicine training grant T15-LM007450.

Address for correspondence

Siddharth Pratap / Vanderbilt Univ. / Dept. Biomedical Informatics
2209 Garland Ave. / Nashville, TN 37232 / USA