Evaluating Cost Efficiency of SNP Chips in Genome-wide Association Studies

Chun Li,[1,2]* Mingyao Li,[3] Ji-Rong Long,[4] Qiuyin Cai,[4] and Wei Zheng[4]

[1]*Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee*
[2]*Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, Tennessee*
[3]*Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania*
[4]*Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee*

Genome-wide association (GWA) studies have recently emerged as a major approach to gene discovery for many complex diseases. Since GWA scans are expensive, cost efficiency is an important factor to consider in study design. However, it often requires extensive and time-consuming computer simulations to compare cost efficiency across different single nucleotide polymorphism (SNP) chips. Here, we propose two simulation-free approaches to cost efficiency comparisons across SNP chips. In the first method, the overall power under a given disease model is calculated for each SNP chip and various sample sizes. Then SNP chips can be compared with respect to the sample sizes required to achieve the same level of power. In the second method, for a desired level of genomic coverage, the effective r^2 threshold values are calculated for each SNP chip. Since r^2 is inversely proportional to the sample size to achieve the same power, the required sample sizes can then be compared among SNP chips. These two methods are complementary to each other. The first approach provides direct power comparisons, but it requires information on disease model and may not be reliable for SNP chips that contain many non-HapMap SNPs. The second approach allows sample size comparisons based on the coverage of SNP chips, and it can be modified for SNP chips that contain non-HapMap SNPs. These methods are particularly relevant for large epidemiological studies in which enough subjects are available for GWA screening and follow-up stages. We illustrate these approaches using five currently available whole genome SNP chips. *Genet. Epidemiol.* 32:387–395, 2008. © 2008 Wiley-Liss, Inc.

Key words: genome-wide association; cost efficiency; HapMap

Contract grant sponsor: NIH; Contract grant number: RO1C64277, RO1CA90899; Contract grant sponsor: University of Pennsylvania.

*Correspondence to: Chun Li, Center for Human Genetics Research, Vanderbilt University School of Medicine, 519 Light Hall, Nashville, Tennessee 37232. E-mail: chun.li@vanderbilt.edu

Received 13 September 2007; Revised 7 November 2007; Accepted 2 January 2008
Published online 12 February 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20312

INTRODUCTION

Genome-wide association (GWA) studies have now become feasible due to recent developments in genotyping technologies, a rapid decline in genotyping costs [Hirschhorn and Daly, 2005; Wang et al., 2005; Hinds et al., 2005], and the completion of the International HapMap Project [The International HapMap Consortium, 2005, 2007]. In GWA studies, investigators typically rely on commercial genotyping products that attempt to provide adequate coverage across the genome. Whole genome single nucleotide polymorphism (SNP) chips from both Affymetrix and Illumina have been successfully used to identify disease genes. Since multiple SNP chips are available, when designing a GWA study, it is necessary to evaluate and compare all available SNP chips and select the one that can provide high-quality data in a cost-efficient manner. To date, the comparisons across SNP chips have been mostly on the chips' coverage of the genome [Barrett and Cardon, 2006; Pe'er et al., 2006], irrespective of genotyping cost. Although SNP chips with a higher genomic coverage are more desirable, such chips generally also cost more. Because a GWA scan is expensive, cost efficiency is an important factor to consider in study design [Skol et al., 2006]. However, it has been difficult to compare cost efficiency across different SNP chips without extensive and time-consuming computer simulations.

In this paper, we propose two simulation-free approaches to cost efficiency comparisons among SNP chips. Both approaches rely on the information of linkage disequilibrium (LD). For each SNP chip and each SNP in the genome, we calculate LD as measured by r^2 between the SNP and the SNPs on the chip and obtain the maximum r^2. The two methods differ in their summarization of the LD information. The first method calculates the power for each SNP and then the overall power for all SNPs across the genome assuming they are equally likely to be associated with the disease. To calculate power for each SNP, information on disease model is needed. This method allows us to compare the cost efficiency of different SNP chips either by comparing power with a