Cell Research (2009) 19:913-915.

www.nature.com/cr

LETTER TO THE EDITOR

A design of multi-source samples as a shared control for association studies in genetically stratified populations

Cell Research (2009) **19**:913-915. doi: 10.1038/cr.2009.75; published online 23 June 2009

Dear Editors,

More and more genetic variants which contribute to human complex traits were identified recently in genome-wide association studies (GWAS), an approach that shows more efficiency than any other genetic approaches ever before [1]. It holds the promise to disclose genetic mechanism underlying the mystery of human diseases, since the most, if not all, of the genetic variants in the human genome could be investigated for their possible association with diseases by comparing the frequencies of alleles in the cases (patients) and controls (healthy subjects).

In GWAS, statistical power is the principle issue in the design of studies [2]. The statistical power of a study indicates the probability of identifying the risk locus if it is the one. Keeping a larger sample size of normal controls and cases is usually considered in order to achieve higher statistical power in association studies. Shared control for different studies, or common control, is a possible solution to achieve a large sample size, and the effort and cost of genotyping different sets of controls could also be reduced. Shared control has been successfully used in GWAS project [2]. However, this design may be impeded by the presence of population stratification in the cases as well as controls.

A slight genetic background difference between cases and controls is sufficient to inflate type I error rate [3], therefore, could drastically increase false positive results. Several different methods have been proposed to detect population stratification and to control the false positive rate for association studies with stratified samples [4-7]. All these approaches have shown some success in controlling the type I error rate.

Population stratification also affects statistical power in association studies. In a stratified population, hidden divergence among sub-populations leads to serious power changes when disease prevalence of sub-populations correlates with risk allele frequency and other population parameters [8]. Even for family-based association test (FBAT), power can be heavily affected by population differentiation since the estimated weights of families might provide poor approximation of the true theoretical optimal weights [9], even though the type I error rate in FBAT is little affected by the population stratification.

Population stratification is known to exist in Chinese populations across geographic area. Multiple studies suggested consistently the presence of a significant boundary between the populations of north and south in China [10, 11]. The genetic heterogeneity, therefore, poses a big challenge to association studies in Chinese populations [12], which can not be lightly ignored especially in light of the upcoming efforts of conducting large scale GWAS using a shared control design in China.

To investigate the magnitude of the divergence between representative Han populations in China, we conducted a genome-wide study including 1 987 Han individuals (representing both Southern and Northern Han) from 14 populations across China (Beijing, Shandong, Anhui, Shanghai, Zhejiang, Jiangsu, Guangdong, Taiwan etc.). Genetic divergence (measured by Fst) between paired populations are in a range from 0.0002 to 0.007 with an average of 0.003. Principle component analysis indicated that Han Chinese populations could be classified into three groups, i.e., Northern Han, Southern Han and Han from Zhejiang, Jiangsu & Shanghai, respectively. The average pairwise genetic divergence between any of the three groups is approximately 0.003.

The knowledge of genetic heterogeneity in Chinese populations laid a foundation for achieving a better, if not optimal, design of GWAS with shared controls. A better design is expected to yield higher power for a given sample size, which was calculated using a simulation algorithm throughout this study, given a predefined type I error (0.05) and other population parameters including sample size. The simulation algorithm incorporated a hierarchical model to address genetic divergence in multi-source scenarios. Given a risk allele frequency p_A and genetic divergence (Fst), allele frequency of each subpopulation p^i_A were generated randomly in a beta distribution

with mean p_A and variance $p_A(1\text{-}p_A)^{Fst}$. Statistical power on the given type I error (0.05) can be calculated for different study designs when the risk allele frequencies of subpopulations and other parameters were known (such as sampling strategy, prevalence of disease and relative risk of risk allele etc.) [8].

Since most of the patient samples, i.e., cases, were pre-collected from large hospitals in cosmopolitan areas, it is sensible to assume that they often come from multiple but unknown sources or subpopulations. To simplify the demonstration of the results, we first consider 4 000 patients from four different areas (1 000 cases per population). The overall size of the control group is fixed (8 000) for different designs (Figure 1). The divergence among populations is set to Fst = 0.003, the relative risk of the risk allele is 1.2 and disease prevalence is 0.1 for all the populations. Statistical power was evaluated with two different risk allele frequencies (0.3, 0.1) and type I error rate was controlled in 0.05 with a method similar to genomic control.

For both risk allele frequencies, we showed that power increases with increasing sources of controls (Figure 1A), and the single-source control yields the lowest power. Even when the case and control are completely miss-matched, statistical power of a multi-source design is still higher than the single-source one, although partially matched controls could deliver an even better performance. It should be noted that a perfect match can hardly be achieved given the complexity of the sources of patients in China.

Another question of utmost importance is whether the shared control could benefit future association studies in

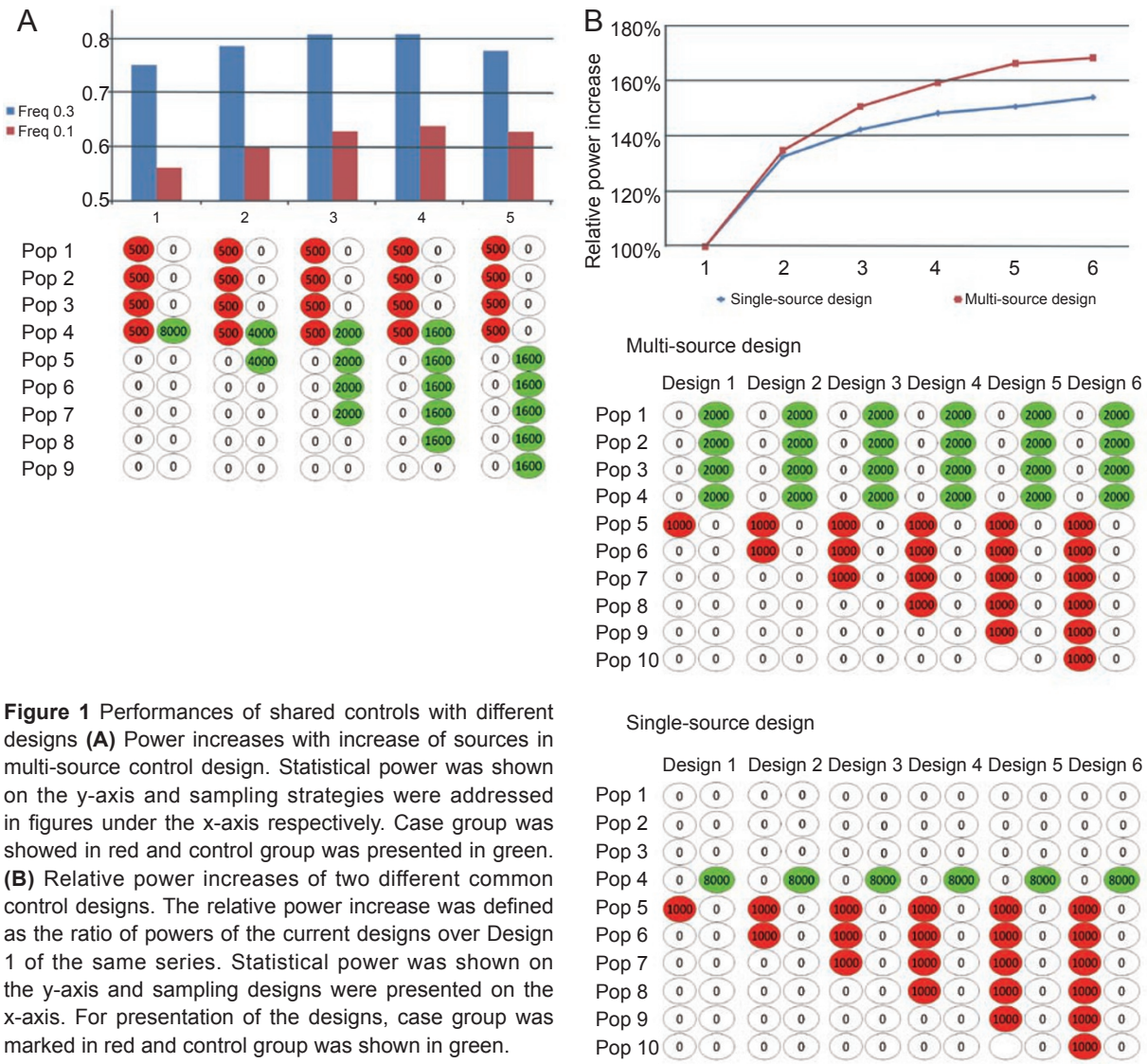

Figure 1 Performances of shared controls with different designs **(A)** Power increases with increase of sources in multi-source control design. Statistical power was shown on the y-axis and sampling strategies were addressed in figures under the x-axis respectively. Case group was showed in red and control group was presented in green. **(B)** Relative power increases of two different common control designs. The relative power increase was defined as the ratio of powers of the current designs over Design 1 of the same series. Statistical power was shown on the y-axis and sampling designs were presented on the x-axis. For presentation of the designs, case group was marked in red and control group was shown in green.

which the diversity of the sources of samples is expected to increase. To address this question, we investigated the power change assuming that 1 000 new cases will be added for each new study (Figure 1B). At a worse scenario, the new cases will come from the source not overlapping with those of the controls. The frequency of the risk allele was set to 0.1, while other parameters were taken as before. Two series of study designs were compared: one with single-source control of 8 000 subjects and the other with a control from four sources (2 000 subjects per source). The results of relative power are presented in Figure 1B which was defined as the ratio of powers of the current designs over Design 1 of the same series. The multi-source designs out-performed the single-source designs in terms of gain of power. We also observed that the power continually improves with increasing sources of new subjects for both types of designs. However, the power increase was nearly stopped for the single-source control design when number of studies is more than 3 (Design 4-6), while the multi-source control design continues to gain more statistical power. This result demonstrates that not much power can be scratched in a 'bad' design by recruiting more patients for the study.

To summarize, this study showed that a good design of shared control for Han Chinese should include samples from multiple sources. A good multi-source design will not only benefit GWAS study using the existing cases but also the future studies when more patient samples are added.

Yungang He[1, 2], Shuhua Xu[1, 2], Chuan Jia[3], Li Jin[1, 2, 3]

[1]CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai 200031, China; [2]Key laboratory of Computational Biology at CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai 200031, China; [3]MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China

Correspondence: Li Jin
Tel: +86 021 65643714; Fax: +86 021 55664885
E-mail: ljin007@gmail.com

References

1 McCarthy MI, Abecasis GR, Cardon LR, *et al*. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**:356-369.
2 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**:661-678.
3 Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000; **66**:1933-1944.
4 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997-1004.
5 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000; **67**:170-181.
6 Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping using a mixture model for complex traits. *Genet Epidemiol* 2002; **23**:181-196.
7 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**:904-909.
8 He Y, Jiang R, Fu W, Bergen AW, Swan GE, Jin L. Correlation of population parameters leading to power differences in association studies with population stratification. *Ann Hum Genet* 2008; **72**:801-811.
9 Ding X, Weiss S, Raby B, Lange C, Laird NM. Impact of population stratification on family-based association tests with longitudinal measurements. *Stat Appl Genet Mol Biol* 2009; **8**: Article 17. DOI: 10.2202/1544-6115.1398
10 Xue F, Wang Y, Xu S, *et al*. A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. *Eur J Hum Genet* 2008; **16**:705-717.
11 Du R, Xiao C, Cavalli-Sforza LL. Genetic distances between Chinese populations calculated on gene frequencies of 38 loci. *Sci China C Life Sci* 1997; **40**:613-621.
12 Shi Y, Zhao X, Yu L, *et al*. Genetic structure adds power to detect schizophrenia susceptibility at SLIT3 in the Chinese Han population. *Genome Res* 2004; **14**:1345-1349.