Genomics & Informatics Vol. 6(1) 14-17, March 2008

Inbreeding Coefficients in Two Isolated Mongolian Populations - GENDISCAN Study

Joohon Sung[1,2], Mi Kyeong Lee[3] and Jeong-Sun Seo[3,4]*

[1]Department of Epidemiology, Seoul National University School of Public Health, Seoul 110-799, Korea, [2]Department of Cancer Prevention and Epidemiology, National Cancer Center, Goyang 410-769, Korea, [3]Ilchon Institute, Seoul National University College of Medicine, Seoul 110-799, Korea, [4]Department of Biochemistry, Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

GENDISCAN study (Gene Discovery for Complex traits in Asian population of Northeast area) was designed to incorporate methodologies which enhance the power to identify genetic variations underlying complex disorders. Use of population isolates as the target population is a unique feather of this study. However, population isolates may have hidden inbreeding structures which can affect the validity of the study. To understand how this issue may affect results of GENDISCAN, we estimated inbreeding coefficients in two study populations in Mongolia. We analyzed the status of Hardy-Weinberg Equilibrium (HWE), polymorphism information contents (PIC), heterozygosity, allelic diversity, and inbreeding coefficients, using 317 and 1,044 STR (short tandem repeat) markers in Orkhontuul and Dashbalbar populations. HWE assumptions were generally met in most markers (88.6% and 94.2% respectively), and single marker PIC ranged between 0.2 and 0.9. Inbreeding coefficients were estimated to be 0.0023 and 0.0021, which are small enough to assure that conventional genetic analysis would work without any specific modification. We concluded that the population isolates used in GENDISCAN study would not present significant inflation of type I errors from inbreeding effects in its gene discovery analysis.

Keywords: inbreeding coefficient, Mengolian population, STR, HWE, PIC

Introduction

The GENDISCAN (Gene Discovery for Complex traits in Asian population of Northeast area) study was launched in 2002 in order to elucidate genetic causes of complex diseases. This study attempted to incorporate designs that detect genetic signals with increased efficiency. These included using genetically homogeneous population, recruiting large families, and considering quantitative phenotypes as well as disease outcome (Peltonen *et al.*, 2001; Merikangas *et al.*, 2003). Large extended families still remaining in the Northeast Asia, enabled the project to adopt these designs.

Although there is no doubt that gene discovery of common complex diseases is one of the research priorities, the successful results have been very limited (Grant *et al.*, 2006).

The difficulty of replication across studies, mandates the use of internally valid study designs and proper methodologies. Using population isolates generally confers the advantage of increasing genetic homogeneity. However population isolates might have inbreeding structures, which deviates the basic assumptions of HWE. The presence of significant inbreeding necessitates modifications in genetic estimations using the population. Therefore, we attempted to estimate the status of HWE, and inbreeding coefficients in two ethnic groups of Mongolia using genome-wide short tandem repeat (STR) genetic markers. Compatibility with basic assumptions of population genetics can support the methodological validity of the overall GENDISCAN study,

Methods

Participants

The GENDISCAN study included non-selected families in Mongolia. The People's Republic of Mongolia (not including the Chinese territory) has 2.6 million people which comprise of more than 20 ethnic groups. The Orkhontuul are in Selenge Imag (Imag is an administrative district unit in Mongolia corresponding to a state in the United States) and the Dashbalbar area in Dornod Imag were selected. The Orkhontuul area has a population of 3,760 people, mainly consisting of Khalkha tribe, and maintains semi-urban life style. The Dashbalbar area is mainly habituated by about 4,000 people of Buryat ethnicity and has more traditional nomadic life style. Many large extended

*Corresponding author: E-mail jeongsun@snu.ac.kr
Tel +82-2-764-9444, Fax +82-2-740-4534
Accepted 16 March 2008

families, which fit the study purposes of the GENDISCAN study still remain in both areas.

Genotyping

Genomic DNA was extracted from peripheral leukocytes. The Orkhontuul samples (2004, n=1,080) were genotyped using the Applied Biosystems Inc. platform (ABI Prism^{TM} Linkage Mapping Set version 2.5 medium density, 400 markers) with average 10 cM resolution, and Dashbalbar samples (2006, n=1,020) were genotyped using the deCODE 1,000 STR marker platform with average of 3 cM resolution. For the Orkhontuul participants markers on the chromosome 1～4 were analyzed. For Orkhontuul data, markers with low call-rate (49 markers), and with more than 1% of genotype error rates (16 markers) and markers on X chromosome (18 markers) were excluded. For Dashbalbar genotype data, the 1,000 STR marker platform provided 1097 markers originally, however we excluded markers on X chromosome (49 markers) and markers with low call-rate and more than 1% of genotype error rates (4 markers). All participants provided informed consent.

Estimating Hardy-Weinberg Equilibrium (HWE), Information Contents and Inbreeding Coefficients

HWE and degree of inbreeding were assessed using the founders of each pedigree. Non-founders were excluded because their genotypes are dependent on those of the founders. HWE was estimated by comparing the expected and observed genotype frequencies. Expected genotype frequency was calculated from allele frequency. Chi-square goodness of fit test was used to determine whether HWE assumption was met. The Chi-square statistics (X^2) of multi-allelic loci is defined as equation as Equation 1, with k (k-1) degree of freedom, where k is the total number of alleles.

$$X_T^2 = \sum_u \frac{(n_{uu} - n\tilde{p}_u^2)^2}{n\tilde{p}_u^2} + \sum_u \sum_{v>u} \frac{(n_{uv} - 2n\tilde{p}_u\tilde{p}_v)^2}{2n\tilde{p}_u\tilde{p}_v}$$

(Equation 1)

where, nuu and nuv denote homozygotic and heterozygotic genotypes, while pu and pv denote allele frequency of each allele.

Information contents of the genetic markers were estimated as polymorphism information content (PIC), heterozygosity and allelic diversity. PIC is an index of the amount of information, which modifies the simple heterozygosity index by adjusting for the chance of mating between the same heterozygotic genotypes. PIC was calculated from Equation 2.

$$\mathrm{PIC} = 1 - \sum_{u=1}^{k} \tilde{p}_u^2 - \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} 2\tilde{p}_u^2\tilde{p}_v^2$$

(Equation 2)

where p_u and p_v denote allele frequency of each allele (Czika, 2005).

Inbreeding was estimated by the deviation from the assumption that each founder shares no Identity by descent (IBD). Generally genotype frequency of bi-allelic locus having *p* and *q* allele frequencies are predicted as p_2, $2pq$, q_2 respectively under HWE. However, if there are IBD sharing of FI between founders, above prediction can be re-written respectively as Equation 3.

$$p^2(1-F_I) + pF_I \quad 2pq(1-F_I) \quad q^2(1-F_I) + qF_I$$

(Equation 3)

where, F_I denotes inbreeding coefficient (Gillespie *et al.*, 2004).

In brief, inbreeding is characterized by the excess of homozygote over expected level. The inbreeding coefficient can be estimated as Equation 4 by solving Equation 3

$$F_I = 1 - \frac{H}{2pq}$$

(Equation 4)

where, H denotes observed heterozygotic, and 2pq denotes estimated heterozygotic proportions from allele frequency (Hart *et al.*, 2000).

HWE and estimations of expected and observed heterozygosity frequencies were obtained using SAS/Genetics program.

Results

The demographic characteristics of the subjects geno-

Table 1. Demographic characteristics of two Mongolian populations used in the study

Area	Orkhontuul	Dashbalbar
Ethnicity	Khalkha	Buryat
Number of Genotyped Founders	280	142
Men	99	90
Women	181	52
Total number of participants	1,032	1,020
Mean age of the founders	49.60	50.69

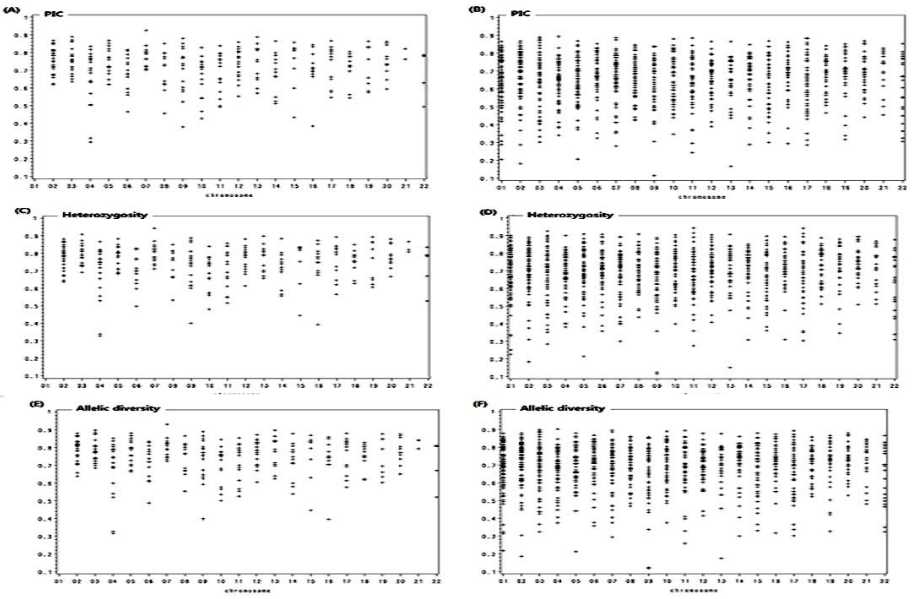

Fig. 1. Distribution of polymorphism information content (PIC), heterozygosity index, and allelic diversity across the 22 chromosomes among Orkhontuul (A, C, E) and Dashbalbar (B, D, F) populations

Table 2. Hardy-Weinberg Equilibrium status in two Mongolian populations used in the study

	Orkhontuul	Dashbalbar
Total effective markers	317	1,044
Number of markers with HWE (%)	281 (86.1%)	984 (94.2%)
Number of markers satisfying HWE ($p < 0.05$)	245	924
Number of markers satisfying HWE ($p < 0.01$)	251	944

typed are shown in Table 1. There were 280 (99 men and 181 women) and 142 (90 men and 52 women) founders in Orkhontuul and Dashbalbar populations. Non-founders' genotype. were excluded, since theirs do not independently contribute to a gene pool. The information contents in terms of PIC for single marker, range between 0.2 and 0.9, as shown in Fig. 1. Average PIC was 0.72 and 0.71 for Orkhontuul and Dashbalbar populations, respectively which are relatively high for single marker information contents. There was no significant difference in PIC across the chromosomes or populations. The high PIC level enabled accurate estimation of other population genetic parameters. HWE was satisfied among 88.6 % and 94.2%, respectively, of all markers in Orkhontuul and Dashbalbar populations (p-value <0.05). If we apply the criteria of p-value 0.01, 90.5% and 95.3% of all markers were in HWE status All the markers including those which were not in HWE, were used for estimating the inbreeding coefficients,. Inbreeding coefficient was estimated to be 0.0023 and 0.0021 in Orkhontuul and Dashbalbar populations.

Table 3. Inbreeding coefficients in two Mongolian populations used in the study

	Orkhontuul	Dashbalbar
Total effective markers	317	1,044
Average number of alleles for the loci	11.7	8.5
Inbreeding coefficient	0.0023	0.0021

Discussion

Population isolates are generally considered to be one of the most ideal populations for genetic study (Pajukanta *et al.*, 2003; rcos-Burgos *et al.*, 2002; Escamilla *et al.*, 2001). However, possible inbreeding can cause deviation from general assumptions on which most analyses depend. Presence of inbreeding can be problematic, because, if exits, I the genetic relationships between unrelated as well as related persons could be underestimated. This underestimation of IBD can result in inflation of type I errors for linkage analysis (Hossjer *et al.*, 2006 Nomura *et al.*, 2005), linkage disequilibrium estimations and haplotype reconstructions (Zhang *et al.*, 2004). The inbreeding coefficient found in this study (about 0.2% in each population), does not necessitate any adjustment for genetic analyses such as IBD calculation, classic or non-parametric linkage analysis, and variance component-based linkage analysis. By estimating the last common ancestor, 0.2% of inbreeding coefficient corresponds to 10 or 11 generations (Jensen-Seaman *et al.*, 2001; Santos-Lopes *et al.*, 2007).

In this study, both ABI and deCODE STR markers were genotyped with standardized procedure and any markers with more than 1% of genotype errors were discarded. The genotype errors were confirmed within the pedigree structure. Any Mendelian inconsistency was deleted and markers with possible double-recombination were also deleted. Generally, genotyping in family-based study is more accurate than in studies using individuals only. Thus, It is not likely that any genotype error could have been biased our findings.

In conclusion, we have estimated inbreeding coefficients in two population isolates in Mongolia.,. We found that they fall in negligible range, allowing related genetic studies to be performed without any modification or adjustment for possible inbreeding effects. This finding validates the ability of The GENDISCAN study to add to the growing body of evidence which associates specific genetic variations with complex disorders.

Acknowledgements

Korean government approved and has supported the pilot study, and phase I study (Ministry of Science and Technology, Grant # M10306030002-03B4803-00210).

References

Peltonen, L., McKusick, V.A. (2001). Genomics and medicine. Dissecting human disease in the postgenomic era. Science 291,1224-1229.

Merikangas, K.R., Risch, N. (2003) Genomic priorities and public health. *Science* 302, 599-601.

Grant, S.F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K.P., Walters, G.B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R.L., Reilly, M.P., Rader, D.J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J.R., Kong, A., and Stefansson, K. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet.* 38, 320-323.

Czika, W. (2005). SAS/Genetics 9.1.3 User's Guide. *SAS Institute Inc.*, Cary, NC, USA.

Gillespie, J.H. (2004). Population Genetics, A concise guide. 2nd Edition, The Johns Hopkins University Press, Baltimore MA, USA.

Hartl, D.L. (2000). Primer of population genetics 3 sub-Edition. *Sinauer Associates*, USA.

Pajukanta, P., Allayee, H., Krass, K.L., Kuraishy, A., Soro, A., Lilja, H.E., Mar, R., Taskinen, M.R., Nuotio, I., Laakso, M., Rotter, J.I., de Bruin, T.W., Cantor, R.M., Lusis, A.J., and Peltonen, L. (2003). Combined analysis of genome scans of dutch and finnish families reveals a susceptibility locus for high-density lipoprotein cholesterol on chromosome 16q. *Am J Hum Genet.* 72, 903-917.

rcos-Burgos, M., and Muenke, M. (2002). Genetics of population isolates. *Clin Genet.* 61, 233-247.

Escamilla, M.A. (2001). Population isolates: their special value for locating genes for bipolar disorder. *Bipolar Disord.* 3, 299-317.

Hossjer, O. (2006). Modeling the effect of inbreeding among founders in linkage analysis. *Theor Popul Biol.* 70, 146-163.

Nomura, T. (2005). Joint effect of selection, linkage and partial inbreeding on additive genetic variance in an infinite population. *Biom J.* 47, 527-540.

Zhang, W., Collins, A., Gibson, J., Tapper, W.J., Hunt, S., Deloukas, P., Bentley, D.R., and Morton, N.E. (2004). Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc Natl Acad Sci USA.* 101, 18075-18080.

Jensen-Seaman, M.I., Deinard, A.S., and Kidd, K.K. (2001). Modern African ape populations as genetic and demographic models of the last common ancestor of humans, chimpanzees, and gorillas. *J Hered.* 92, 475- 480.

Santos-Lopes, S.S., Pereira, R.W., Wilson, I.J., and Pena ,S.D. (2007). A worldwide phylogeography for the human X chromosome. *PLoS ONE.* 2, e557.