Allele frequency determination of publicly available cSNPs in the Korean population

Seong-Gene Lee, PhD[1], Yongsook Yoon[2], Sunghee Hong[2], Jieun Yoo[2], Insil Yang[2], and Kyuyoung Song, PhD[1,2]

Purpose: As a first step toward the construction of a single-nucleotide polymorphism (SNP) database of the Korean population, the authors determined the allele frequencies of 406 cSNPs selected from the public database. **Methods:** A pooled DNA sequencing approach was used to determine the allele frequencies of 406 cSNPs selected from 120 genes in 24 individuals. **Results:** Of 406 cSNPs, 53% were monomorphic in the Korean samples. Among tested SNPs, 292 SNPs (72%) were uncommon (minor allele <20%) and 114 SNPs (28%) were common (minor allele ≥20%) in our population. **Conclusion:** An extensive SNP characterization would be necessary, and the ethnic and population-based differences should be considered in the selection of SNPs for the study of complex diseases with association mapping methods. ***Genet Med* 2002:4(6, Supplement):49S–51S.**

Key Words: *publicly available cSNPs, allele frequency, ethnic variation, pooled DNA sequencing approach, Korean population, comparison of allele frequencies*

The central aim of genetics is to correlate specific molecular variation with particular phenotype changes. Since the human genome draft sequence was announced in June 2000, it has become possible to understand the spectrum of genetic variation in the human gene pool and its relation to diseases, individual responses to environmental factors, and biological processes such as development and aging. Geneticists have so far used familial linkage methods for studying simple single-gene diseases.[1] However, complex, high-incidence, multigene diseases are thought to require a different approach, that is, an association study, for which single-nucleotide polymorphisms (SNPs) can play a key role.[1–3]

SNPs consist of the most abundant form of genetic variation and have great potential for mapping complex genetic traits.[4–7] Because of their potential as genetic markers, scientists in the public and private sectors have begun to focus their attention on searching for SNPs throughout the human genome.[3,8–12] For whole-genome association studies, it is estimated that approximately 100,000 to 500,000 SNPs would be required.[5,6] Currently, a great deal of effort is being invested in the identification of SNPs. As of June 14, 2002, 4,275,093 SNPs had been deposited into public databases (http://www.ncbi.nlm.nih.gov/SNP) compared with 7,000 SNPs in April 1999. As the SNP identification process accelerates, it becomes necessary to characterize a large number of publicly available SNPs in a population. For an association mapping study, SNP allele frequencies in the population would be critical.

We decided to characterize publicly available cSNPs in the Korean population because (1) many SNPs are publicly available already, (2) those SNPs were discovered using rather limited samples, (3) some SNPs may not be common in a given population, and (4) SNP allele frequencies in the population would be critical for an association mapping study. Here, we present the allele frequencies of 406 publicly available cSNPs in 24 Koreans.

MATERIALS AND METHODS

Collection of candidate cSNPs

Candidate cSNPs were identified in the Whitehead Institute for Genome Research (http://www.genome.wi.mit.edu) databases. We picked SNPs for which allele frequencies were already available, and primers were ordered.

DNA samples

The study included 24 unrelated individuals: 12 males and 12 females. Genomic DNA was extracted from peripheral white blood cells by standard methods. Samples were diluted to approximate concentrations of 10 ng/μL on the basis of a spectrophotometer measurement at 260 nm and visual estimation by 0.7% agarose gel electrophoresis. The pooled DNA sample was prepared by combining equal amounts of genomic DNA from 21 individuals in one tube. The DNA of three individuals and the pooled DNA were sequenced for each marker.

DNA amplification

Polymerase chain reaction (PCR) was performed with 10 ng of genomic DNA in a final volume 15 μL of 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.25 mM $MgCl_2$, 0.033 mM concentration of each dNTP, 5 pmol oligonucleotide primer, and 0.25 units

From the [1]Asan Institute for Life Sciences, [2]Department of Biochemistry, University of Ulsan College of Medicine, Seoul, Korea.

Kyuyoung Song, PhD, Asan Institute for Life Sciences, Department of Biochemistry, University of Ulsan College of Medicine, 388-1 Pungnap-dong, Songpa-gu, Seoul 138-736, Korea.

Received: June 27, 2002.

Accepted: September 23, 2002.

DOI: 10.1097/01.GIM.0000040259.03889.E6

of AmpliTaq Gold *Taq* polymerase (Applied Biosystems, Foster City, CA). Thermocycling conditions were as follows: initial denaturation at 95°C for 12 minutes, with 35 cycles at 95°C for 30 seconds, 2 minutes of annealing at 56 to 60°C, and extension at 72°C for 40 seconds, followed by a final extension step at 72°C for 5 minutes on the Gene-Amp PCR System 9700 (Applied Biosystems, Foster City, CA). Three microliters of PCR products were checked on a 1.5% agarose gel.

Sequencing of PCR products

PCR products were treated with 1 μL each of 2 units of exonuclease I (USB, Cleveland, OH) and 0.4 units of shrimp alkaline phosphatase (USB) and incubated at 37°C for 15 minutes and then at 85°C for 15 minutes. For SNP sequencing, we performed a conventional dye-terminator sequencing reaction in a total volume of 10 μL containing 5 μL of enzyme-treated PCR product, 5 pmol of sequencing primer, 1 μL of dye ET terminator sequencing premix (Amersham Pharmacia Biotech, Piscataway, NJ), and 1 μL of half-TSII dilution buffer (GenPak, New Milton, UK—as of January 1, 2002, GenPak became Genetix). After ethanol precipitation, the reaction products were run on a MegaBACE 1000 sequencer according to the manufacturer's instructions. All PCR products were sequenced routinely in one direction; however, some PCR products with a low-quality sequence (below Phred score 20) and monomorphic peak pattern were sequenced in both directions.

Estimation of allele frequency in a pool

Sequences were base-called and assembled by the Phred and Phrap computer program (a kind gift from Dr. Phil Green, University of Washington, Seattle, WA), and the polymorphism in the tested population was investigated. Allele frequencies of the pool were estimated by comparing the peak heights of the corresponding bases between a heterozygous individual and the pooled DNA sample.[13] The normalized peak height from a heterozygous individual represents the signal contributed by 50% of the DNA template present that bears that of a particular allele. Therefore, we calculated allele frequencies of the pool from the ratio between the normalized peak height of a heterozygote and the pooled DNA. When a heterozygote was not identified in individual samples, but the peak pattern of the pool showed polymorphism, the reference heterozygote was identified by sequencing four individuals from the pool and used for its allele frequency estimation.

RESULTS AND DISCUSSION

Previously, we have published the results of the characterization of 300 publicly available SNPs in the Korean population using a pooled DNA sequencing approach.[14] In this report, we present allele frequencies of 406 publicly available cSNPs in 24 Koreans. Although a pooled DNA sequencing approach is not highly accurate for detecting SNPs with a minor allele frequency of less than 10%, it still is a useful approach for identifying informative SNP markers and estimating their allele frequencies in a cost-effective way.[13,14] We have shown that the accuracy of a pooled DNA sequencing approach is within 9% of that determined by sequencing each individual in the pool.[14] The markers were considered to be monomorphic only when the peak patterns in both directions were clear, without the background signal of a minor allele. When a weak signal of a minor allele was present in the pool, those markers were categorized to have minor allele frequencies of less than 10%, as shown on previous analyses.[14]

Table 1
Sources of Whitehead Institute cSNP markers

Sources	No. of SNPs
Noncoding SNPs	146
Coding SNPs	260
Nonsynonymous	116
Synonymous	144
Total	406

No. of genes = 120.

As shown in Table 1, 406 publicly available cSNPs (derived from 120 genes) were chosen from the Whitehead Institute for Genome Research database. These markers were originally identified in the genes relevant to cardiovascular disease, endocrinology, and neuropsychiatry[15] and were chosen because their allele frequency data were available. Of those, 146 cSNPs were from noncoding regions [untranslated regions (UTRs) and introns], 144 cSNPs were synonymous cSNPs, and 116 cSNPs were nonsynonymous cSNPs. Of 406 cSNPs, 53% were monomorphic in our samples (Table 2). This is more than twice the percentage of monomorphic markers we have shown in a previous report (23%) that included noncoding SNPs identified in 125 genes and 175 expressed sequence tags.[14] As was expected, this reflects that SNPs causing amino acid changes are rare compared with the ones in noncoding regions.[15,16] Among the screened SNPs, 92 SNPs (23%) showed minor allele frequency of 31 to 50%, 43 (11%) of 11 to 30%, and 271 (66%) lower than 10% (Table 2). From the distribution of minor allele frequencies, 74% of nonsynonymous cSNPs have a minor allele frequency lower than 10%, whereas 59% of noncoding SNPs fall in this category. Compared with our previous report in

Table 2
Distribution of minor allele frequencies

Minor allele frequency (%)	SNP type: *n* (%)			
	Nonsynonymous	Synonymous	Noncoding	Total
0	73 (63)	76 (53)	68 (47)	217 (53)
0–10	13 (11)	24 (17)	17 (12)	54 (13)
11–30	11 (10)	17 (12)	15 (10)	43 (11)
31–50	19 (16)	27 (19)	46 (32)	92 (23)
<20	91 (78)	106 (74)	95 (65)	292 (72)
≥20	25 (22)	38 (26)	51 (35)	114 (28)
Total	116	144	146	406

Table 3
Comparison of minor allele frequencies between Korean and publicly available data

	Nonsynonymous		Synonymous		Noncoding	
Allele frequency (%)	Korean	WI	Korean	WI	Korean	WI
>15	25 (21)	22 (19)	38 (27)	51 (36)	54 (37)	60 (41)
5–15	15 (13)	21 (18)	21 (15)	27 (19)	20 (14)	39 (27)
<5	76 (65)	73 (63)	85 (59)	66 (46)	72 (49)	47 (32)
Total	116	116	144	144	146	146

Values represent *n* (%). WI, Whitehead Institute.

which we found that 24% were uncommon (minor allele <20%) and 76% common (minor allele ≥20%), in this study, 72% were uncommon and 28% were common. This could be due to the fact that the markers chosen in this experiment were well annotated and had more nonsynonymous and synonymous cSNPs compared with the previously characterized 300 cSNPs, of which 99% were from noncoding regions. It is unlikely that the rather high percentage of monomorphic and low allele frequency markers in our samples are due solely to the small sample size, detection limit of the pooled sequencing approach, or the presence of false positives in the public SNP databases.

We compared minor allele frequencies of 406 cSNPs in the Korean population with the publicly available data obtained from a mix of Europeans, Asians, African Americans, and African Pigmies (Table 3). Of those, 57% of cSNPs (233 markers) have minor allele frequency less than 5% in the Korean sample, whereas 45.8% have minor allele frequency less than 5% (186 markers) in the mix sample. The trends in the distribution of minor allele frequencies of the markers are quite similar between the Korean and Whitehead Institute data. When the allele frequencies of individual markers were compared, 66% of nonsynonymous cSNPs, 57% synonymous cSNPs, and 41% noncoding cSNPs were common both in the Korean and the mix samples. The fact that the nonsynonymous cSNPs were enriched in low-frequency alleles compared with the noncoding cSNPs in both samples suggests that cSNPs causing amino acid changes are deleterious, and selection was working against them. Our data suggest that an extensive SNP characterization would be necessary, and the ethnic and population-based differences should be considered in the selection of SNPs for the study of complex diseases with association mapping methods.

The number of available SNPs in public databases is expected to grow exponentially in the next few years. Thus, it is important to estimate the allele frequencies in a variety of populations. Because so many SNPs are already available and will continue to increase, rare SNPs are going to be avoided in the mass screening of a population. Allele frequency information would be a prerequisite in the selection of useful markers for future association studies. The comparative sequencing approach provides reasonable accuracy and the capacity to handle a large number of markers in a cost-effective way. We are currently in the process of increasing the SNP database in the Korean population on the basis of previously described variants.

Acknowledgments

This work was supported by grants to Kyuyoung Song from the Functional Analysis of the Human Genome Project in the 21C Frontier R&D Program of MOST (Ministry of Science and Technology of Korea) and the Asan Foundation through the Asan Institute for Life Sciences. The authors thank the individuals who agreed to participate in the characterization of SNPs in the Korean population.

References

1. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037–2048.
2. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516–1517.
3. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resources for research on human genetic variation. *Genome Res* 1998;8:1229–1231.
4. Kruglyak L. The use of genetic map of biallelic markers in linkage studies. *Nat Genet* 1997;17:21–24.
5. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999;22:139–144.
6. Brookes AJ. The essence of SNPs. *Gene* 1999;234:177–186.
7. Chakravarti A. Population genetics: making sense out of the sequence. *Nat Genet* 1999;21:56–60.
8. Marshall E. "Playing chicken" over gene markers. *Science* 1997;278:2046–2048.
9. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997;278:1580–1581.
10. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077–1082.
11. Buetow KH, Edmonson MN, Cassidy AB. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet* 1999;21:323–325.
12. Marshall E. Drug firms to create public database of genetic mutations. *Science* 1999; 284:406–407.
13. Kwok PW, Carlson C, Yager TD, Ankener W, Nickerson D. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* 1994;23:138–144.
14. Lee S, Hong S, Yoon Y, Yang I, Song K. Characterization of publicly available SNPs in the Korean population. *Hum Mutat* 2001;17:281–284.
15. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalvanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of a single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999;22:231–238.
16. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 1999;22:239–247.