J Hum Genet (2003) 48:213–216
DOI 10.1007/s10038-003-0011-9

ORIGINAL ARTICLE

**Jong-Keuk Lee · Hung-Tae Kim · Sung-Mi Cho
Kyung-Hee Kim · Hee-Jeong Jin · Gil-Mi Ryu
Bermseok Oh · Chan Park · Kuchan Kimm
Sangmee Ahn Jo · Sung-Chul Jung · Sook Kim
Sun Mi In · Jong-Eun Lee · Inho Jo**

Characterization of 458 single nucleotide polymorphisms of disease candidate genes in the Korean population

Received: 29 January 2003 / Accepted: 17 February 2003 / Published online: 20 March 2003

Abstract Single nucleotide polymorphisms (SNPs) are considered as very promising genetic markers for complex disease gene hunting. However, it has been demonstrated that there are significant ethnic differences in genetic variations. In order to investigate the genetic variations in the Korean population and their ethnic differences, a large number of SNPs of 161 disease candidate genes were collected from a publicly available SNP database and then tested for the distribution of allele frequency in the Korean population. Of all 458 SNPs tested, approximately 43.9% were polymorphic in the Korean population, whereas 44.5% were monomorphic. The remaining 11.6% were failed in the test. Significant differences have been observed when SNP allele frequency pattern of Koreans was compared with those of Caucasians and Africans, whereas this pattern was highly similar between Korean and Japanese populations. Our data indicate that although many of the SNPs available in publicly available database, especially coding-region SNPs (cSNPs), can be used as informative genetic markers for disease association studies, an extensive verification of public SNPs in a particular population studied should be undertaken prior to their association studies.

J.-K. Lee · H.-T. Kim · S.-M. Cho · K.-H. Kim · H.-J. Jin
G.-M. Ryu · B. Oh · C. Park · K. Kimm
National Genome Research Institute,
National Institute of Health, 5 Nokbun-dong,
Eunpyung-gu, Seoul 122-701, Korea

S. A. Jo · S.-C. Jung · I. Jo (✉)
Department of Biomedical Sciences,
National Institute of Health, 5 Nokbun-dong,
Eunpyung-gu, Seoul 122-701, Korea
E-mail: inhojo@nih.go.kr
Tel.: +82-2-380-1521
Fax: +82-2-388-0924

S. Kim · S. M. In · J.-E. Lee
DNA Link, Inc., 15-1 Yeonhee-dong,
Seodaemun-gu, Seoul 120-110, Korea

Keywords Single nucleotide polymorphism · SNP · Disease candidate genes · Korean · Ethnicity

Introduction

Genetic variations have been used for the identification of disease-related genes. Among those genetic variations, single nucleotide polymorphisms (SNPs) are the most common genetic variations between individuals, existing at a frequency of approximately 1 in every 300–1000 bases in the human genome (Brookes 1999; Cargill et al. 1999). Therefore, SNP can be used to facilitate genetic mapping studies that may lead to a better understanding of the genetic basis for complex diseases such as high blood pressure, diabetes, asthma and inflammatory diseases (Johnson et al. 2000; Horikawa et al. 2000; Hugot et al. 2001).

A large number of SNPs are deposited in the public SNP database, dbSNP (Sherry et al. 2001), at the U.S. National Center for Biotechnology Information (NCBI). Almost 5 million SNPs of the estimated 10 million SNPs have been identified to date and over 3 million SNPs have currently been assigned as "rsSNPs" at dbSNP (NCBI dbSNP build 110). However, there are many sequencing errors in SNP discovery steps and ethnic differences in SNP allele frequency patterns. In addition, most of the SNPs are located in intergenic regions. Although any SNPs can be used as useful genetic markers to identify disease loci, ultimately the functional SNP identified in the regulatory or coding region of the genes may be more important. Therefore, in order to test the accuracy of SNP information present in the public SNP database and also to compare ethnic differences in SNP allele frequency, approximately 458 SNP (mainly cSNP) sites of 161 disease candidate genes were collected from the public SNP database and the frequencies of the selected SNP sites were tested in the Korean population.

Materials and methods

Selection of candidate SNPs and DNA samples

Disease candidate genes mainly associated with immune responses were selected for the following diseases: atopic dermatitis, asthma, cardiovascular diseases, gastritis–hepatitis, and cancers. The candidate genes were chosen on the basis of their potential relevance to the selected common diseases. A large number of SNPs in coding regions (cSNPs) and some untranslated regions from the selected disease candidate genes were collected in the publicly available dbSNP database (http://www.ncbi.nlm.nih.gov/SNPs). A total of 43 healthy Korean women aged 34–62 years (53.2 $\pm$6.3, mean $\pm$ standard deviation), who did not have any pathological symptoms at the time of interview and blood test, were randomly selected. Informed written consent for participation was obtained from each individual. Blood was drawn into an ACD-A tube and the lymphocytes were isolated and transformed with EB virus. Genomic DNA was isolated from the EB-virus transformed lymphocytes with the standard method.

SNP genotyping for allele frequency

SNP genotyping was performed by SNP-IT™ assays using SNPstream 25K™ System (Orchid Biosciences, New Jersey, USA). Briefly, the genomic DNA region spanning the polymorphic site was PCR-amplified using one phosphothiolated primer and one regular PCR primer. The amplified PCR products were then digested with exonuclease. The 5′ phosphothiolates were used in this study to protect one strand of the PCR-product from exonuclease digestion. The single-stranded PCR template generated from exnuclease digestion was overlaid onto a 384 well plate that precoated covalently with the primer extension primers, SNP-IT™ primers. These SNP-IT™ primers were designed to hybridize immediately adjacent to the polymorphic site. After hybridization of template strands, SNP-IT™ primers were then extended by a single base with DNA polymerase at the polymorphic site of interest. The extension mixtures contained two labeled terminating nucleotides (one FITC, one biotin) and two unlabeled terminating nucleotides. The final single base incorporated was identified with serial colorimetric reactions with anti-FITC-AP and streptavidin-HRP, respectively. The results of blue and/or yellow color developments were analyzed with an ELISA reader and the final genotyping (allele) calls were made with the QCReview™ program.

Korean-SNP database

The Korean-SNP database was constructed at the National Genome Research Institute (National Institute of Health, Korea). All SNP allele frequency data described in this study are currently available at the Korean-SNP database (http://152.99.72.69/~SNP/ksnp.html).

Results and discussion

Selection of SNPs in disease candidate genes from the public SNP database

Prior to association studies by using SNP for searching complex disease genes, we initially collected several disease candidate genes relevant to asthma, atopic dermatitis, hepatitis, and cancers. The lists of disease candidate genes (total 161 genes) were obtained from several disease genome centers in Korea and a total of 458 SNPs of the disease candidate genes were selected from the public dbSNP database. All SNPs and names of selected disease candidate genes used in this study are available at the Korean-SNP database web site as described in Materials and methods.

Distribution of SNP allele frequencies of disease candidate genes in the Korean population

To investigate the distribution of SNP allele frequency in the Korean population, a total of 458 SNP sites, selected from 161 disease candidate genes, were genotyped in 43 unrelated female Korean individuals using SNP-IT™ methodology. As shown in Fig. 1(A), among 458 SNP sites tested in this study, 201 SNP sites were polymorphic in the Korean population, indicating that 43.9% of SNP sites selected from the public SNP database were polymorphic in the Korean population. The allele frequency distribution of 201 polymorphic SNPs with respect to the frequency of the minor allele was shown in Fig. 1(B). The allele frequencies of 201 polymorphic SNPs in the Korean population were not uniformly distributed. The highest proportion (33.8%) of the rare SNPs having a minor allele frequency of less than 10% was observed (Fig. 1B), perhaps representing the true distribution of SNPs in nature. The average minor allele frequency of 21.3% in the Korean population was similar with a previous study in the Japanese population that reported an average minor allele frequency of 24% (Haga et al. 2002). In addition, among 201 polymorphic SNPs, approximately two thirds (66%) had greater than 10% minor allele frequency in the Korean population, indicating that those SNP sites can be used as useful genetic markers for searching complex disease genes in the association studies.

Ethnic differences in SNP allele frequency

At the present time, the public SNP database is very useful means for checking the presence of SNP allele frequency in a particular population, especially major ethnic groups such as Caucasians, Africans and Asians. In order to compare the allele frequency of those ethnic groups with that of Korean population, both The SNP Consortium (TSC; http://snp.cshl.org) database and JSNP database (http://snp.ims.u-tokyo.ac.jp) were used. Among 458 SNPs tested in this study, only 7 and 32 SNPs were matched with TSC data and JSNP data, respectively, which have allele frequency data of other ethnic groups. This result indicates that SNPs available at the public SNP database have very limited numbers of allele frequency data in a particular population. Furthermore, this result, together with previous available data showing only limited SNP characterization using the Korean population (Lee et al. 2001), suggests that, using the Korean population, an extensive characterization of publicly available SNPs should be undertaken

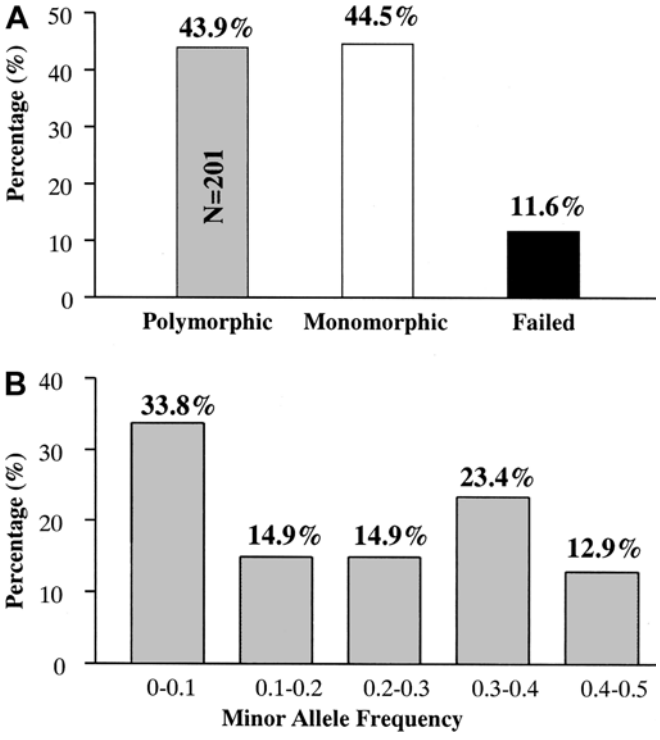

Fig. 1A, B Summary of SNP genotyping in the Korean population. A total of 458 SNPs were selected from the public SNP database and the presence of SNPs was determined in the Korean population using the SNP-IT™ system. Among 458 SNPs tested in this study, 201 SNPs were polymorphic in the Korean population (**A**). The distribution of minor allele frequency of verified 201 SNPs was shown (**B**)

prior to use of those candidate SNPs for disease association studies.

It is interesting that the 7 SNPs overlapping with TSC data showed striking differences in allele frequency among ethnic groups (Table 1). For example, no SNP polymorphism of the EGF-R gene (rs#884225) was observed in Caucasians and Africans, whereas the same SNP site was polymorphic with frequencies of 0.2 and 0.37 in Asians and Koreans, respectively. In contrast, the SNP of the NAT2 gene (rs#1208) was nearly monomorphic in Asians and Koreans, whereas this SNP was polymorphic with frequencies of 0.29 and 0.33 in Caucasians and Africans. In the majority of 7 SNPs, similar allele frequency patterns were found between Africans and Caucasians except one SNP (IL-12Rβ2, rs#1495963). The latter IL-12Rβ2 gene was virtually monomorphic in Caucasians, whereas this gene is polymorphic in the other three ethnic groups, having similar allele frequencies. Furthermore, a major "C" allele of CD70 (rs#1862511) in Caucasians and Africans was inversed to a minor allele in Asians and Koreans. As shown in the EGF-R gene (rs#884225) and IL-12Rβ2 gene (rs#1495963), considerable differences in allele frequency between Koreans and Asians were found. These differences may be due to sampling variability since only Japanese and Chinese, but not Koreans, were included among the Asian samples of TSC data.

Although significant differences in SNP allele frequency were detected among different ethnic groups (Table 1), the comparison of allele frequencies between Koreans and Japanese showed a high similarity of SNP allele frequency patterns having less than 10% allele frequency differences in all SNPs tested except EGF (rs#2302135) (Table 2). Recently, it has been reported that the Asian population had the smallest number of distinct SNP haplotypes. Furthermore, allele frequency patterns between Korean and Japanese in our data were shown to be comparable with previous data within the Japanese population (Okuda et al. 2002), suggesting that both Korean and Japanese populations may share a common origin of ancestry, as expected from the close geographical location of the two countries. Our study has, however, one important limitation, which is a finding that the ethnic differences in allele frequencies between Koreans and Japanese are based on only a small sample size. Since genotype distribution with small sample size is sometimes different from true distribution with large populations, further studies, using a larger sample size, need to be conducted to reach an accurate evaluation.

In summary, a total of 458 SNPs selected from the 161 disease candidate genes were characterized in the Korean population. Our results will be further utilized to determine the experimental strategies for studying complex diseases using SNPs selected from publicly available SNP databases.

Acknowledgements This work was supported in part by an IMT-2000 research grant (01-PJ11-PG9-01BT05-0003) from The

Table 1 Comparison of allele frequency of overlapping Korean SNP data with TSC data

Gene	SNP ID (rs#)	Allele	Caucasian	African	Asian	Korean	Ethnic difference[a]
CD70	1862511	C:T	0.70:0.30	0.88:0.12	0.33:0.67	0.29:0.71	0.59
L-Selectin	909628	C:T	0.04:0.97	0.05:0.95	0.05:0.95	0.04:0.97	0.01
EGF-R	884225	G:A	0:1	0:1	0.20:0.80	0.37:0.63	0.37
F9	6048	G:A	0.20:0.80	0.11:0.89	0.01:0.99	0:1	0.20
NAT2	1208	G:A	0.29:0.71	0.33:0.67	0.02:0.98	0.01:0.99	0.31
IL-12Rβ2	1495963	G:A	1:0	0.67:0.33	0.83:0.17	0.70:0.30	0.33
LIGHT	344560	G:A	0.98:0.02	0.94:0.06	0.88:0.12	0.92:0.08	0.10

[a]Ethnic difference in allele frequency was calculated by subtracting the lowest allele frequency from the highest allele frequency of minor allele among ethnic groups at each SNP site

Table 2 Comparison of allele frequency of verified Korean SNPs and JSNP database data

Gene	SNP (rs#)	Allele	Japanese	Korean	Ethnic difference[a]
BRCA1	16940	C:T	0.35:0.65	0.35:0.65	0
BRCA1	799917	C:T	0.65:0.36	0.65:0.35	0.01
CYP19	2236722	C:T	0.03:0.97	0.07:0.93	0.04
DLK1	2273606	G:A	0.93:0.07	0.95:0.05	0.02
EDNRA	5333	C:T	0.26:0.74	0.23:0.77	0.03
EDNRA	5342	G:A	0.46:0.54	0.44:0.56	0.02
EDNRA	2292764	C:T	0.01:0.99	0.01:0.99	0
EDNRA	5335	C:G	0.51:0.49	0.5:0.5	0.01
EGF	2302135	G:A	0.22:0.78	0.37:0.63	0.15
EGF	2237051	G:A	0.31:0.69	0.26:0.74	0.05
EGF-R	2293347	G:A	0.64:0.36	0.64:0.36	0
ESELE	5368	C:T	0.81:0.19	0.80:0.20	0.01
F7	6042	G:A	0.93:0.07	0.94:0.06	0.01
HDC	2073440	C:A	0.03:0.97	0.04:0.96	0.01
HLA	2071555	G:T	0.80:0.20	0.80:0.20	0
HLA	1042337	G:A	0.29:0.71	0.38:0.62	0.09
ICAM-1	2071440	C:T	0.97:0.03	0.97:0.04	0.01
IL12-RB2	1495963	G:A	0.79:0.21	0.70:0.30	0.09
IL1-RN	315952	C:T	0.65:0.35	0.69:0.31	0.04
LIGHT	2291667	C:T	0.99:0.01	1:0	0.01
LIGHT	2291668	C:T	0.63:0.37	0.70:0.30	0.07
LRP2	2228171	G:A	0.43:0.57	0.43:0.57	0
LT-A	2071589	C:T	0.98:0.03	0.99:0.01	0.02
LT-A	2239704	C:T	0.53:0.47	0.56:0.44	0.03
MUC5B	2292636	C:T	0.99:0.01	1:0	0.01
NGF-R	2072446	C:T	0.90:0.10	0.88:0.12	0.02
NOS2A	1137933	C:T	0.92:0.08	0.87:0.13	0.05
NOS2A	1060822	C:T	0.78:0.22	0.78:0.22	0
NPR3	2270915	G:A	0.22:0.78	0.26:0.74	0.04
PRCP	2298668	C:A	0.10:0.90	0.04:0.96	0.06

[a]Ethnic difference in allele frequency was calculated by subtracting the lowest allele frequency from the highest allele frequency of minor allele among ethnic groups at each SNP site. The comparison was performed by searching JSNP database data with verified SNPs in the Korean population

Ministry of Health and Welfare and The Ministry of Science and Technology, Korea.

References

Brookes AJ (1999) The essence of SNPs. Gene 234:177–186

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millenium Genome Project: Identification of 190562 genetic variations in the human genome. J Hum Genet 47:605–610

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz, PEH, Bosque-Plata LD, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nature 26:163–175

Hugot J-P, Chamaillard M, Zouali H, Lesage S, Cezard J-P, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel J-F, Sahbatou M, Thomas G. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Johnson GCL and Todd JA (2000) Strategies in complex disease mapping. Curr Opin Genet Dev 10:330–334

Lee SG, Hong S, Yoon Y, Yang I, Song K (2001) Characterization of publicly available SNPs in the Korean population. Hum Mutat 17:281–284

Okuda T, Fujioka Y, Kamide K, Kawano Y, Goto Y, Yoshimasa Y, Tomoike H, Iwai N, Hanai S, Miyata T (2002) Verification of 525 coding SNPs in 179 hypertension candidate genes in the Japanese population: identification of 159 SNPs in 93 genes. J Hum Genet 47:387–394

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311