Genomics & Informatics Vol. 6(4) 173-180, December 2008

Identification and Characterization of Human Genes Targeted by Natural Selection

Ha-Jung Ryu[1], Young Joo Kim[2], Young-Kyu Park[2], Jae-Jung Kim[1], Mi-Young Park[1], Eul-Ju Seo[3], Han-Wook Yoo[3], In-Sook Park[3], Bermseok Oh[4] and Jong-Keuk Lee[1,3]*

[1]Asan Institute for Life Sciences, University of Ulsan College of Medicine, Seoul 138-736, Korea, [2]KRIBB, Daejeon 305-806, Korea, [3]Genome Research Center for Birth Defects and Genetic Disorders, Asan Medical Center, University of Ulsan College of Medicine, Seoul 138-736, Korea, [4]Department of Biomedical Engineering, Kyung Hee University School of Medicine, Seoul 130-702, Korea

Abstract

The human genome has evolved as a consequence of evolutionary forces, such as natural selection. In this study, we investigated natural selection on the human genes by comparing the numbers of nonsynonymous (NS) and synonymous (S) mutations in individual genes. We initially collected all coding SNP data of all human genes from the public dbSNP. Among the human genes, we selected 3 different selection groups of genes: positively selected genes (NS/S$\geq$3), negatively selected genes (NS/S$\leq$1/3) and neutral selection genes (0.9 $<$NS/S$<$1.1). We characterized human genes targeted by natural selection. Negatively selected human genes were markedly associated with disease occurrence, but not positively selected genes. Interestingly, positively selected genes displayed an increase in potentially deleterious nonsynonymous SNPs with an increased frequency of tryptophan and tyrosine residues, suggesting a correlation with protective effects against human disease. Furthermore, our nonsynonymous/synonymous ratio data imply that specific human genes, such as ALMS1 and SPTBN5 genes, are differentially selected among distinct populations. We confirmed that inferences of natural selection using the NS/S ratio can be used extensively to identify functional genes selected during the evolutionary adaptation process.

Keywords: single nucleotide polymorphism (SNP), natural selection, disease genes, ethnicity

*Corresponding author: E-mail cookie_jklee@hotmail.com
Tel +82-2-3010-4142, Fax +82-2-486-3312
Accepted 20 October 2008

Introduction

The human genome has evolved as a consequence of historical forces, such as natural selection. The present human genome contains unique signals that have been copied over time, either via positive or negative selection. Inferences on the patterns and distribution of natural selection on the human genome may thus provide important functional information (Nielsen, 2005). For example, highly skewed negative or positive natural selection indicates that a particular human gene has been selected by the actions of several evolutionary mechanisms, predominantly due to disease or adaptation to the surrounding environment. Accordingly, it may be possible to identify putative risk factors for genetic disease by determining regions of the human genome or genes currently under selection (Arbiza *et al*., 2006). Negative selection reduces the number of nonsynonymous (NS) mutations, whereas positive selection enhances this number, relative to the number of synonymous (S) mutations (Biswas & Akey, 2006). In previous studies, the NS/S ratio test led to the successful identification of specific positively selected genes, including human olfactory genes and human leukocyte antigen (HLA) loci (Salamon *et al*., 1999; Gilad *et al*., 2000). Therefore, the NS/S ratio test is a recognized tool for the effective detection of types of natural selection in protein-coding genes. Under conditions of no selection, we would expect a NS/S ratio of 1. In case of negative selection, NS/S is $<$1, and with positive selection, NS/S would be $>$1 (Biswas & Akey, 2006). Furthermore, the availability of large SNP datasets allowed us to determine where natural selection (either negative or positive) has effected variations in humans (Nielsen *et al*., 2007). In this study, we investigated natural selection on the human genes by comparing the simple ratios of nonsynonymous and synonymous coding SNPs (cSNPs) in individual protein-coding genes.

Methods

The dataset

We downloaded and analyzed all coding SNPs (cSNPs) with a validation code greater than 2 from the public dbSNP (build 125, http://www.ncbi.nlm.nih.gov/SNP/). Where necessary, we additionally used genotype data generated from the International HapMap Project with

90 Yoruba Nigerian (YRI), 45 Japanese (JPT), 45 Han Chinese (HCB) and 90 CEPH (CEU) individuals. The number of disease genes located on each chromosome was determined by searching the OMIM database (http://www.ncbi.nlm.nih.gov/omim/). All coding SNP (cSNP) data were used to perform natural selection mapping of human protein-coding genes at the individual gene levels.

Statistical analyses of SNP data for determining natural selection

To investigate the natural selection process, we primarily employed the nonsynonymous vs synonymous cSNPs ratio. *In silico* prediction of the functional significance of nonsynonymous cSNPs resulting in amino acid alterations, was determined using PolyPhen (http://genetics.bwh.harvard.edu/pph/) (Ramensky *et al*., 2002) and/or SIFT (http://blocks.fhcrc.org/sift/SIFT.html) (Ng & Henikoff, 2003). As a measure of population differentiation, Wright's F_{ST} was estimated from genotypic data. Observed heterozygosity (Ho) and expected heterozygosity (He) were additionally obtained from HapMap database for the selected candidate genes. We subsequently investigated the differences (ΔH) between expected heterozygosity (He) and observed heterozygosity (Ho) to establish the direction of natural selection, such as heterozygote advantages or disadvantages. General statistical analysis was performed with SAS for windows V8 (SAS Institute Inc., Cary, NC, USA), and statistical significance was inferred at a two-tailed values of $p < 0.05$.

Results

Gene-Based Nonsynonymous/Synonymous Ratio Test

To detect human genes targeted by natural selection, we initially collected coding SNP (cSNP) data, including synonymous and nonsynonymous SNPs, of all human genes from the public dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). In total, 15,913 genes were selected due to the presence of at least one coding SNP. From the selected genes, 24,019 nonsynonymous and 24,641 synonymous SNPs were accumulated. To characterize the natural selection process, we classified and plotted all human genes on the basis of individual nonsynonymous (NS) and synonymous (S) SNP numbers. Subsequently, we selected three distinctive groups of genes representing each selection type (positive, negative and neutral), as shown in Fig. 1, classified based on the nonsynonymous/synonymous (NS/S) ratio. Among human genes with at least 10 synonymous or nonsynonymous SNPs each, we singled out 39 positively selected genes with $NS/S \geq 3$, 37 negatively selected genes with $NS/S \leq 1/3$, and 12 neutral selection genes with $0.9 < NS/S < 1.1$ for comparison.

Patterns of disease association by selection type

Following selection of genes from the three typical groups (Fig. 1), we analyzed the patterns of disease association by selection type. In the disease association analysis, we eliminated all hypothetical genes and pseu-

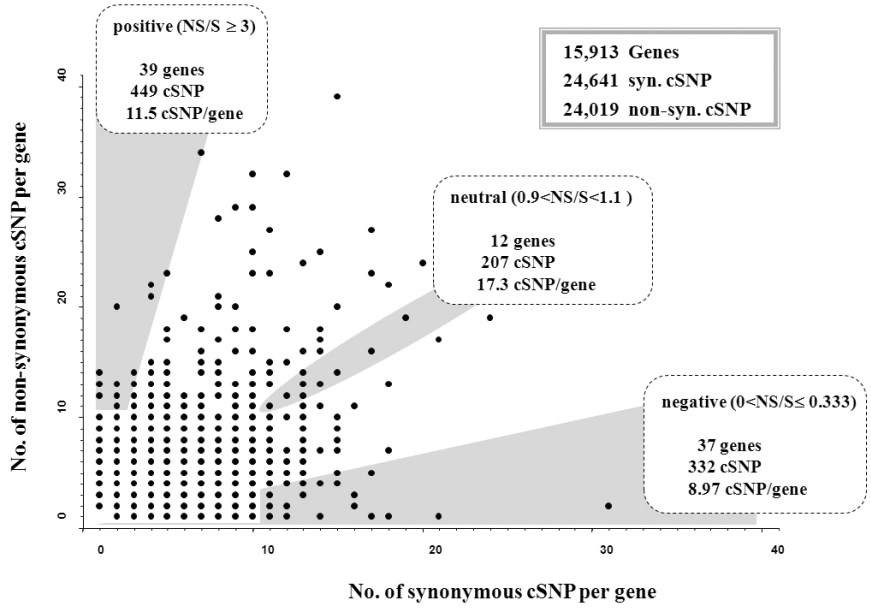

Fig. 1. Plot of human genes from the three selection types (positive, neutral and negative) based on individual nonsynonymous/synonymous (NS/S) ratios. All human genes (15,913) containing at least one coding SNP (cSNP) were plotted on the basis of the number of synonymous (S) and nonsynonymous (NS) SNPs. To resolve the characteristics of each selection type, extremely skewed selection groups were chosen on the basis of the NS/S ratio (positive selection, $NS/S \geq 3$; negative selection, $NS/S \leq 1/3$; neutral selection, $0.9 < NS/S < 1.1$), as indicated above in shadow.

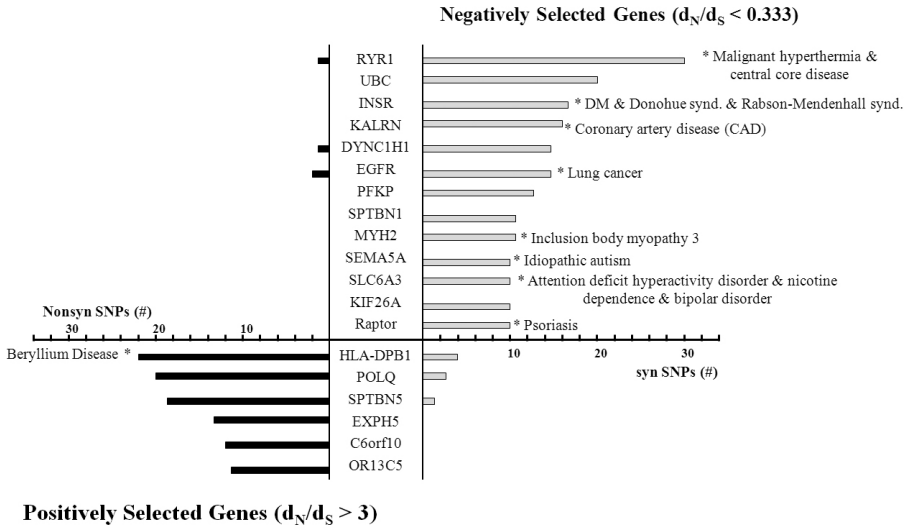

Fig. 2. Association of positively or negatively selected genes with human disease. Genes subjected to extreme positive or negative selection were identified as specified in Fig. 1, and evaluated for known disease association by searching the OMIM database. Diseases linked to selected genes are presented. Notably, negatively selected genes ($0 < NS/S < 0.333$) were strongly associated with human diseases, but not positively selected genes ($NS/S > 3$).

dogenes from each selection group. As shown in Fig. 2, negatively selected genes were strongly associated with disease. Among the 13 negatively selected genes, 8 showed disease association, specifically, RYR1 with malignant hyperthermia and central core disease, INSR with diabetes mellitus and Rabson-Mendenhall syndrome, KALRN with coronary artery disease, EGFR with lung cancer, MYH2 with inclusion body myopathy 3, SEMA5A with idiopathic autism, SLC6A3 with attention deficit hyperactivity disorder, nicotine dependence and bipolar disorder, and RAPTOR with psoriasis. In contrast, no positively selected genes were correlated with human disease, except HLA-DPB1 (linked to beryllium disease). The finding that negatively selected genes (reduced rate of nonsynonymous substitutions) are strongly correlated with involvement in human disease is consistent with previous data (Bustamante *et al.*, 2005).

Increase in potentially deleterious nonsynonymous cSNPs by positive selection

Recently, computational methods have been applied to predict the deleterious effects of nonsynonymous SNPs in humans (Ramensky *et al.*, 2002; Ng & Henikoff, 2003). In this study, we used the *in silico* computational procedure to predict the functionality of nonsynonymous amino acids by selection type: 1) positive, 2) negative, and 3) neutral, using PolyPhen and SIFT programs. We observed the lowest number of deleterious SNPs in negatively selected genes, a low number in neutrally selected genes, and the highest SNP incidence in positively selected genes. PolyPhen analysis (Fig. 3A) disclosed that the proportion of damaging amino acids gradually increased from negative (19%) to neutral (26%) and positive selection (31%). Similar patterns

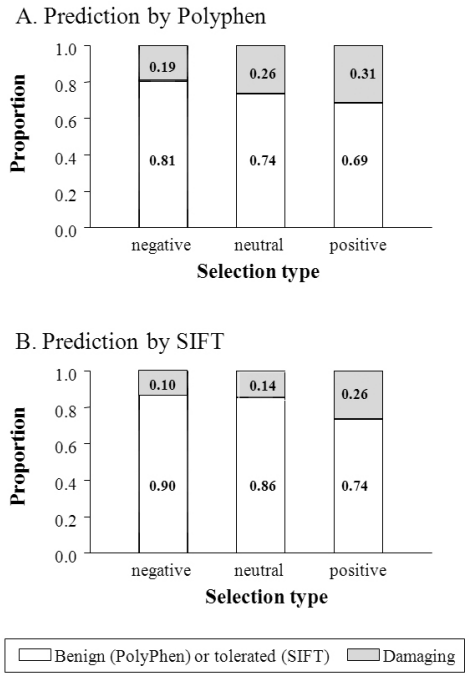

Fig. 3. Distribution of potentially deleterious nonsynonymous cSNPs by selection type (positive, negative and neutral). *In silico* prediction of nonsynonymous SNPs from genes classified in different selection groups (Fig. 1) was performed using PolyPhen (A) and SIFT (B) programs. The damaging group includes nonsense mutations, as well as all types of damaging results, such as possibly and probably damaging (PolyPhen).

were observed with SIFT analysis (negative: 10%, neutral: 14%, positive: 26%) (Fig. 3B). The gradual increase in deleterious SNPs (from negative or neutral to positive

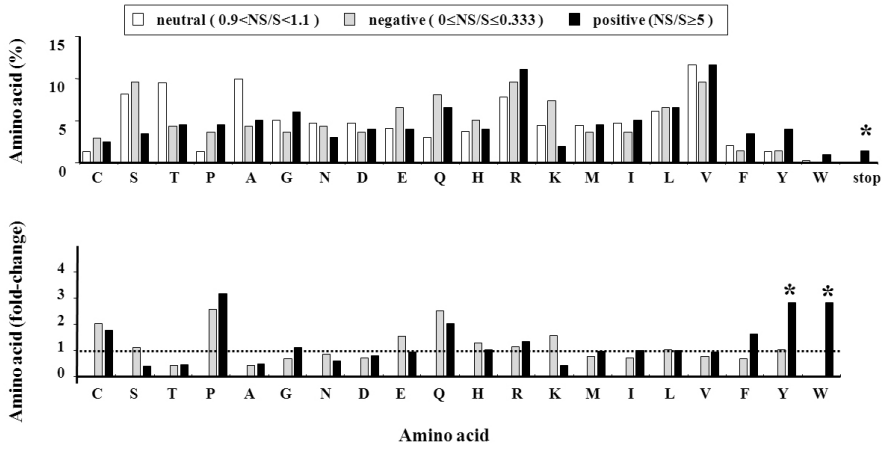

Fig. 4. Distribution of residues from protein-coding genes of all three selection types. The upper figure represents the percentage of each amino acid by selection type. The termination codon is presented as 'stop'. The lower figure depicts the proportion of amino acids subjected to positive and negative selection relative to neutral selection. Amino acids that are significantly changed are indicated with asterisks (*).

selection) suggests significant functional changes in positively selected genes. Deleterious SNPs in positively selected genes may have protective effects against human diseases.

Identification of significant amino acid residues by selection type

Depending on selection type, we observed significant changes in the functionality of amino acids altered due to nonsynonymous SNPs. Consequently, we attempted to identify the types of amino acids that are frequently observed in specific selection types, particularly positive selection. We searched all amino acid variations due to nonsynonymous SNPs from 6 positively selected genes ($NS/S \geq 5$), 29 negatively selected genes ($NS/S \leq 1/3$) and 12 neutrally selected genes ($0.9 < NS/S < 1.1$). For the selection on amino acid type, in this analysis, we increased the NS/S threshold to ≥ 5 (not ≥ 3) in positive selection in order to use similar numbers of amino acids from each selection group. Positively and negatively selected genes were compared to those that underwent neutral selection. Since critical residues in negatively selected genes have been eliminated in the population by natural selection, we mainly focused on positively selected genes. As shown in Fig. 4, we observed an increased frequency of tryptophan and tyrosine residues in positively selected protein-coding genes. Moreover, significantly increased frequency of termination and a marginal increase in phenylalanine were evident (Fig. 4). Our results support the theory that these residues generated as a consequence of positive selection play important biological roles.

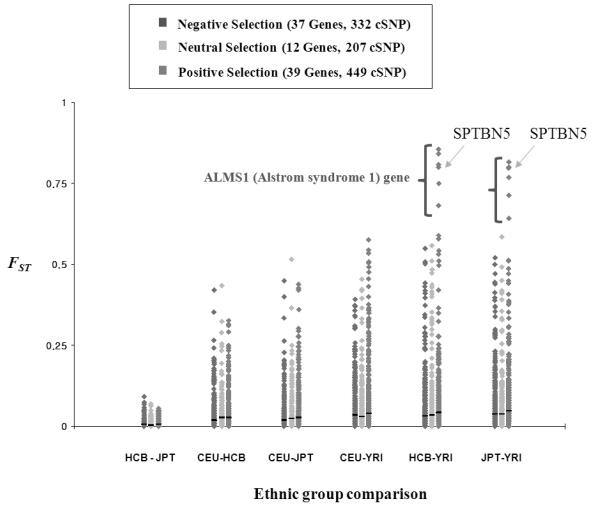

Fig. 5. Distribution of population differentiation (F_{ST}) across ethnic groups by selection type. The standardized variance (F_{ST}) across populations was calculated for each SNP selected from three different selection type groups using HapMap genotype data (CEU, Caucasian; HCB, Han Chinese; JPT, Japanese; YRI, Yoruba African). Significant differences (F_{ST} values) were observed for 9 SNPs and 1 SNP of positively selected genes, ALMS1 and SPTBN5, respectively, between Asian and African populations. Strong F_{ST} values of the ALMS1 and SPTBN5 genes are marked. Black bars represent the median values.

Population differentiation (F_{ST}) by selection type

To evaluate the population divergence of the three natural selection types, we estimated the F_{ST} (fixation index) among different ethnic groups. As shown in Fig. 5, no significant differences in F_{ST} were evident between negative and neutral selection groups among several human

populations. However, in the positive selection group, a significant distribution difference was observed, particularly between Asian and African populations. In particular, high F_{ST} values (range of F_{ST}=0.64～0.86) were obtained for 9 SNPs (rs6724782, rs3813227, rs6546837, rs6546839, rs2056486, rs6546838, rs10193972, rs1052162 and rs6546836) in the ALMS1 gene in Asian and African populations. Ethnic differences in the ALMS1 gene were additionally observed between Caucasian and Africans, with lower F_{ST} values (range of F_{ST}=0.40～0.58). Extreme differences in allele frequencies of ALMS1 among populations have been reported previously (The International HapMap Consortium, 2005). Since ALMS1 is a causative gene for Alstrom syndrome 1 resulting in obesity, type 2 diabetes and neurosensory degeneration (Collin *et al*., 2002), the variable incidence of this syndrome among ethnic groups is attributed to the natural selection process. The ALMS1 gene displayed high F_{ST} values between Africans and Asians as well as between Africans and Caucasians. In addition, a SNP (rs1456235) in SPTBN5 showed a very high F_{ST} value (F_{ST}=0.80) between Africans and Asians, but low F_{ST} between other ethnic groups (range of F_{ST}=0.21～0.29). These data indicate that SPTBN5 is an important Asian gene in the positive selection group. Significant population variations of positively selected genes (ALMS1 and SPTBN5) imply differential positive selection pressure in distinct ethnic populations.

Distribution of heterozygosity by selection type among ethnic groups

Observed heterozygosity (Ho) and expected heterozygosity (He) of genes were plotted from all three selection groups (data not shown). In Africans, median heterozygosity (both observed and expected) was similar (ranging from 0.188 to 0.203) for different types of selection. However, the distribution of median heterozygosity was different in Caucasian and Asian populations. Median heterozygosity of neutral selection genes in these populations was as high as that in Africans, ranging from 0.20 to 0.244. However, median heterozygosity in the positive selection group was very low in both populations (0.045～0.133). Our data suggest that the positive selection process occurs over a short evolutionary period in Caucasian and Asian populations. In addition, we calculated the differences between observed and expected heterozygosity (Ho - He) to identify the direction of selection and genes undergoing ethnic group-specific selection. A high positive value signifies a heterozygote advantage, whereas a high negative value is suggestive of heterozygote disadvantage or genomic deletions. In neutral selection genes, SNPs were randomly distributed between negative and positive values, as expected (Fig. 6). However, olfactory receptor genes (OR6K6 & OR10C1 genes) in the positively selected group were differentially selected in distinct populations. For example, the OR6K6 gene has a heterozygote advantage in Africans, whereas OR10C1 has a heterozygote disadvantage in the Asian population. On the

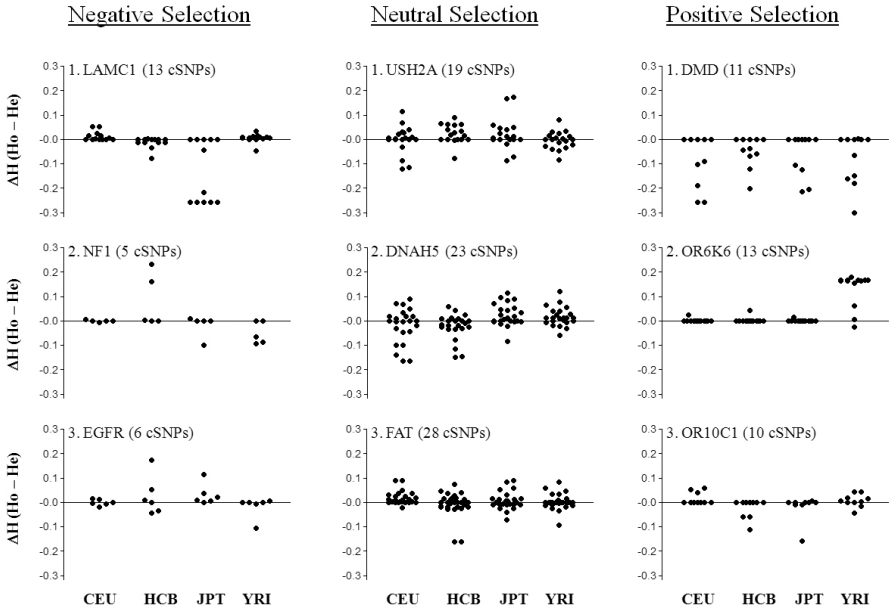

Fig. 6. Identification of genes undergoing ethnic group-specific selection. Genes subjected to ethnic-specific selection were identified by subtracting the expected from observed heterozygosity (Ho - He). Heterozygosity was calculated using HapMap SNP genotype data (CEU, CEPH individuals; HCB, Han Chinese; JPT, Japanese; YRI, Yoruba African). The DMD gene was used as a control showing strong negative values in all ethnic groups due to its location on the X chromosome.

other hand, among negatively selected genes, LAMC1 displayed a significant decrease in observed heterozygosity in Asians (especially Japanese), suggesting a heterozygosity disadvantage of this gene in the Japanese population. In contrast, observed heterozygosity was markedly increased for NF1 and EGFR genes in Asians (Fig. 6). The difference between observed and expected heterozygosity reveals that positively or negatively selected genes undergo different directions of selection pressure, such as heterozygote advantage or disadvantage, in distinct populations.

Discussion

The occurrence of genetic variations in non-functional regions is inconsequential, whereas mutations in the functionally critical regions of genes affect the fitness of organisms, leading to natural selection (Bamshad & Wooding, 2003; Nielsen, 2005). In this study, we identified several human genes that display signatures of natural selection detected using the NS/S ratio. The abundant SNP data obtained from the international HapMap project (The International HapMap Consortium, 2003; 2005; 2007), particularly in coding regions of human genes, renders the NS/S ratio test very powerful for investigating the natural selection of the human genome and genes. In addition, there is increased awareness that regions of the human genome targeted by natural selection are generally of functional importance. Disease- causing mutations should affect organism fitness. Therefore, there is an intimate relationship between disease and selection that can potentially be exploited to identify candidate disease loci and SNPs. In this study, we observed a strong correlation of negatively selected genes (high NS/S ratio) with human disease, such as RYR1 with malignant hyperthermia, INSR with diabetes mellitus, and EGFR with lung cancer. Strong correlation of negative selection with disease genes has been demonstrated in earlier studies (Bustamante *et al*., 2005; Biswas & Akey, 2006), indicating that negative selection constitutes a major evolutionary pressure against human diseases. These findings are consistent with a previous report showing that strong selective pressure with NS/S $<$ 0.1 is significantly associated with human disease (Arbiza *et al*., 2006). However, positively selected genes were not linked to human disease, except HLA-DPB1, which is associated with chronic beryllium disease (Richeldi *et al*., 1993). It is established that HLA regions have undergone positive selection (Salamon *et al*., 1999). As evident with HLA genes, low correlation of positively selected genes with disease is probably a result of their protective effects against human diseases, which may be involved predominantly in the longevity of human life. Positive selection may lead to the generation of a diverse spectrum of amino acids in protein-coding genes, which, in turn, ensure more flexibility in responses to detrimental environments. However, the detailed biological evolutionary functions of positively selected genes remain to be clarified.

To identify the important amino acids that have undergone adaptation by natural selection in the human population, we mainly focused on positive selection genes, since the critical SNPs and residues of negative selection genes have been already eliminated from the human population. Only non-functional SNPs or residues should be abundant in negatively selected genes. This concept was confirmed by *in silico* prediction of nonsynonymous SNPs using PolyPhen in that the frequency of deleterious or damaging nonsynonymous SNPs was low in negatively selected genes (19%), intermediate in neutral selection genes (26%) and high in positively selected genes (31%). The high frequency of deleterious nonsynonymous SNPs in positively selected genes is indicative of significant changes in the functions of corresponding proteins. Furthermore, the high incidence of tryptophan, tyrosine, phenylalanine and termination codons in positively selected genes signifies critical roles of these residues. Additionally, tryptophan, cysteine, arginine and glycine are possibly associated with genetic diseases (Vitkup *et al*., 2003). Therefore, tryptophan appears important for both positive and negative selection, whereas tyrosine, phenylalanine and termination codons are preferentially associated with positively selected genes.

Natural selection can occur in specific populations (Sabeti *et al*., 2006). Several genes have been selected in different ethnic populations, specifically the Duffy antigen (FY) gene for protection against malaria infection in Africa and lactase gene (LCT) for lactose intolerance in Europe (Hamblin & Di Rienzo, 2000; Bersaglieri *et al*., 2004). Population differences were mainly observed for positively selected human genes, indicating variations in positive selection, depending on the population. In particular, it has been reported that the ALMS1 gene has been positively selected in human populations (The International HapMap Consortium, 2005). Furthermore, French Acadians and other ethnically or geographically isolated populations display a higher frequency of Alstrom syndrome (Marshall *et al*., 1997; Deeble *et al*., 2000). The SPTBN5 gene additionally displayed a signal of population-specific selection. However, the biological functions of this gene remain to be characterized. In addition, there are a number of limitations in distinguishing the effects of population-specific selection and migration of the human population, leading to high population dif-

ferentiation (F_{ST}) of SNPs among ethnic groups.

On the other hand, the heterozygosity test is a useful tool to differentiate gene selection in diverse populations (Halliburton, 2004). We detected low median heterozygosity of positively selected genes in Asian and Caucasian populations, suggesting that these populations have a short evolutionary time for positive gene selection, compared to Africans. Another contributory factor to this difference is migration of populations. In addition, differences in observed and expected heterozygosity indicate that the heterozygosity test is a very useful method to identify genes with heterozygote advantages and/or balancing selection in different populations. The OR6K6 gene displayed a strong heterozygote advantage in the African population only, whereas NF1 and EGFR genes displayed heterozygote advantages in the Asian population. Higher somatic mutations in the EGFR gene were evident in Asians (10% European versus 30% Japanese lung cancer patients) (Johnson & Janne, 2005), and mutations were directly correlated to the clinical response to the anti-cancer drug (Gefitinib; EGFR inhibitor). The high incidence of EGFR mutations in cancer may explain the heterozygote advantage of this gene in Asians. Application of the NS/S ratio test previously showed that human olfactory receptor genes as well as human leukocyte antigen (HLA) loci (Salamon *et al*., 1999) have undergone positive selection (Gilad *et al*., 2000). Interestingly, olfactory receptor (OR6K6 & OR10C1) genes were differentially selected in distinct populations. Specifically, the OR6K6 gene had a heterozygote advantage in Africans, whereas OR10C1 exhibited heterozygote disadvantages or genomic alterations, such as copy number variations, in the Asian population, although we do not fully understand the reasons. Ethnic-specific heterozygote advantages of positively or negatively selected genes signify different types of natural selection in distinct populations.

In summary, in spite of some limitations such as biases in distribution and density of SNPs as well as the ascertainment bias, we demonstrated that natural selection analysis of human genes using the NS/S ratio test is an efficient method to elucidate the evolutionary history of individual human genes. We additionally confirmed that inferences of natural selection can be used extensively to identify functional genes selected during the evolutionary adaptation process. Enhanced knowledge of natural selection in human genes will provide novel opportunities for comprehensive analysis of the human genes and genome.

Acknowledgements

This work was supported by a grant from the Ministry of Health & Welfare, Republic of Korea (01-PJ10-PG6-01GN15-001) and a grant from the Research Project on the Production of Bio-organs (No. 200606021001 & No. 20070401034029), Ministry of Agriculture and Forestry, Republic of Korea. YJ Kim was partially supported by a grant from the Cerebrovascular Disease Oriental Medicine Project of MOST (Ministry of Science & Technology).

References

Arbiza, L., Duchi, S., Montaner, D., Burguest, J., Pantoja-Uceda, D., Pineda-Lucena, A., Dopazo, J., and Dopazo, H. (2006). Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J. Mol. Biol.* 358, 1390-1404.

Bamshad, M., and Wooding, S.P. (2003). Signatures of natural selection in the human genome. *Nat. Rev. Genet.* 4, 99-111.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111-1120.

Biswas, S., and Akey, J.M. (2006). Genomic insights into positive selection. *Trends Genet.* 22, 437-446.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., Civello, D., Adams, M.D., Cargill, M., and Clark, A.G. (2005). Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153-1157.

Collin, G.B., Marshall, J.D., Ikeda, A., So, W.V., Russell-Eggitt, I., Maffei, P., Beck, S., Boerkoel, C.F., Sicolo, N., Martin, M., Nishina, P.M., and Naggert, J.K. (2002). Mutations in ALMS1 cause obesity, type 2 diabetes and neurosensory degeneration in Alstrom syndrome. *Nat. Genet.* 31, 74-78.

Deeble, V.J., Roberts, E., Jackson, A., Lench, N., Karbani, G., and Woods, C.G. (2000). The continuing failure to recognise Alstrom syndrome and further evidence of genetic homogeneity. *J. Med. Genet.* 37, 219.

Gilad, Y., Segre, D., Skorecki, K., Nachman, M.W., Lancet, D., and Sharon, D. (2000). Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nat. Genet.* 26, 221-224.

Halliburton, R. (2004). *Introduction to Population* (NJ, USA: Pearson Education Inc.).

Hamblin, M.T., and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* 66, 1669-1679.

Johnson, B.E., and Janne, P.A. (2005). Epidermal growth factor receptor mutations in patients with non-small cell lung cancer. *Cancer Res.* 65, 7525-7529.

Marshall, J.D., Ludman, M.D., Shea, S.E., Salisbury, S.R., Willi, S.M., LaRoche, R.G., and Nishina, P.M. (1997). Genealogy, natural history, and phenotype of Alstrom

syndrome in a large Acadian kindred and three additional families. *Am. J. Med. Genet.* 73, 150-161.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812-3814.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.* 39, 197-218.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. (2007). Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* 8, 857-868.

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900.

Richeldi, L., Sorrentino, R., and Saltini, C. (1993). HLA-DPB1 glutamate 69: a genetic marker of beryllium disease. *Science* 262, 242-244.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. *Science* 312, 1614-1620.

Salamon, H., Klitz, W., Easteal, S., Gao, X., Erlich, H.A., Fernandez-Viña, M., Trachtenberg, E.A., McWeeney, S.K., Nelson, M.P., and Thomson, G. (1999). Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics* 152, 393-400.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature* 426, 789-796.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437, 1299-1320.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-861.

Vitkup, D., Sander, C., and Church, G.M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biology* 4, R72.