The Korean HapMap Project Website

Young Uk Kim[1], Seung-Ho Kim[2], Hoon Jin[1], Young-Kyu Park[1], Mihyun Ji[1] and Young Joo Kim[1]*

[1]Medical Genomics Research Center, [2]Instrumental Analysis Laboratory, KRIBB, Daejeon 305-806, Korea

Abstract

Single nucleotide polymorphisms (SNPs) are the most abundant form of human genetic variation and are a resource for mapping complex genetic traits. A genome is covered by millions of these markers, and researchers are able to compare which SNPs predominate in people who have a certain disease. The International HapMap Project, launched in October, 2002, motivated us to start the Korean HapMap Project in order to support Korean HapMap infrastructure development and to accelerate the finding of genes that affect health, disease, and individual responses to medications and environmental factors. A Korean SNP and haplotype database system was developed through the Korean HapMap Project to provide Korean researchers with useful data-mining information about disease-associated biomarkers for studies on complex diseases, such as diabetes, cancer, and stroke. Also, we have developed a series of software programs for association studies as well as the comparison and analysis of Korean HapMap data with other populations, such as European, Chinese, Japanese, and African populations. The developed software includes HapMapSNPAnalyzer, SNPflank, HWE Test, FESD, D2GSNP, SNP@Domain, KMSD, KFOD, KFRG, and SNP@WEB. We developed a disease-related SNP retrieval system, in which OMIM, GeneCards, and MeSH information were integrated and analyzed for medical research scientists. The kHapMap Browser system that we developed and integrated provides haplotype retrieval and comparative study tools of human ethnicities for comprehensive disease association studies (http://www.khapmap.org). It is expected that researchers may be able to retrieve useful information from the kHapMap Browser to find useful biomarkers and genes in complex disease association studies and use these biomarkers and genes to study and develop new drugs for personalized medicine.

*Corresponding author: E-mail yjkim8@kribb.re.kr
Tel +82-42-879-8127, Fax +82-42-879-8119
Accepted 10 June 2008

Summary

The International HapMap Project started in October, 2002 to develop a haplotype map of the human genome, the HapMap, which describes common patterns of human DNA sequence variation (*The International HapMap Consortium*, 2007). HapMap is expected to be a key resource for researchers to find genes that affect health, disease, and responses to drugs and environmental factors. The project was a collaborative effort between scientists in Japan, the U.K., Canada, China, Nigeria, and the U.S. This motivated us to start the Korean HapMap Project in June, 2003 in order to support Korean HapMap infrastructure development as well as to accelerate finding genes that affect health, disease, and individual responses to medications and environmental factors (Thorisson *et al.*, 2005). The project was undertaken as a collaboration between Korean funding agencies (Ministry of Education, Science and Technology and Korean National Institute of Health), experimental academia (Ulsan Medical Institute, SungKyunKwan Medical Institute, and Korea Advanced Institute of Science and Technology), and corporations (DNA Link, SNP-Genetics, and Samsung Advanced Institute of Technology) (Yoo *et al.*, 2006; Lee *et al.*, 2008).

Resulting from the project, a Korean SNP and haplotype database system was developed to help those researchers who study high-frequency, complex Korean diseases and changes in ethnic global migratory variants. In the project, we tried to accomplish a number of goals. First, the system should be able to provide essential information that is needed for gene discovery of complex Korean diseases. Second, the system should contain basic and advanced tools that may apply to applications such as diagnostics, treatment, and prevention of diseases. Third, the database system should provide Korean-specific SNPs and haplotype information that are common in the Korean population.

We have developed a series of software programs for association studies as well as for the comparison and analysis of Korean HapMap data with four other populations (Yorubans in Ibadan, Nigeria; Centre d'Etude du

Polymorphisme Humain reference families; Japanese in Tokyo; and Han Chinese in Beijing) (*The International HapMap Consortium*, 2005). The software includes HapMapSNPAnalyzer, SNPflank, HWE Test, FESD (Kang and Choi *et al.*, 2005; Kim and Kim *et al.*, 2007), D2GSNP (Kang and Hong *et al.*, 2006), SNP@Domain (Han and Kang *et al.*, 2006), KMSD, KFOD, KFRG, SNP@WEB, and KHapMap Browser. As a partial result

of this project, we registered data that encompasses 1,764,000 Korean SNPs to the dbSNP of NCBI (*The dbSNP Website*, 2008). The KHapMap Browser, based on the generic genome browser, provides haplotype retrieval and comparative study tools of human ethnicities for comprehensive disease association studies (Stein *et al.*, 2002). This study provides a starting point for research to identify genetic variants that are based on in-
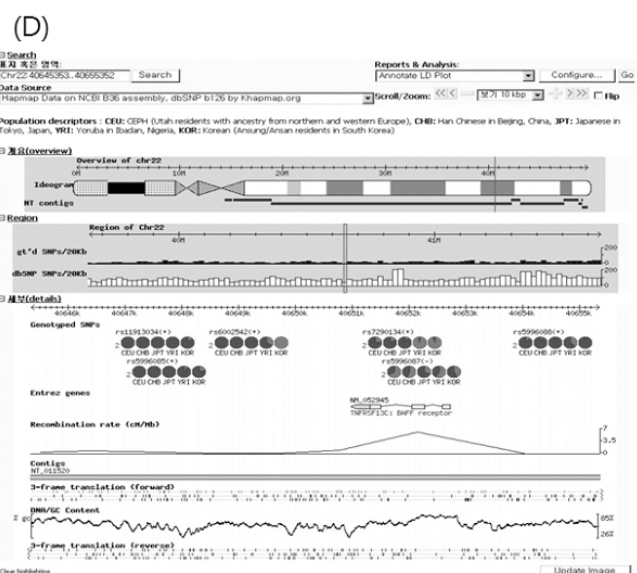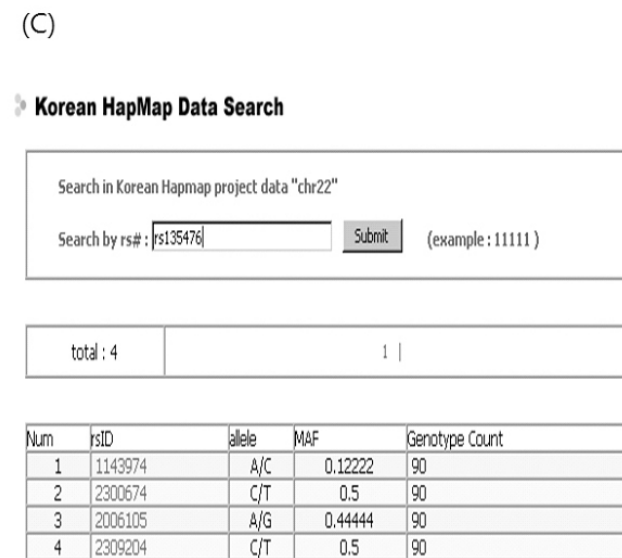
(A)

(B)

(C)

(D)

Fig. 1. The Korea HapMap Website (http://www.khapmap.org). (A) Links to the major menu of the site are located at the top of the page and the quick navigation panel on the left border. (B) Data search page is shown. Users can input rsID or chromosomal position with both a user-set minor allele frequency (MAF) and a user-set minimum genotyped sample number in the input area. (C) Search result shows rsID links to the dbSNP database, allele, MAF, and genotype count. (D) The KHapMap Browser provides interactive access to the HapMap database via a graphical genome browser. Users can compare Korean SNP data with other population descriptors (European (CEU), Chinese (CHB), Japanese (JPT), African (YRI), and Korean (KOR) populations). (Kim *et al.*)

herent Korean characteristics and catalog genetic similarities and differences compared with various ethnic samples that are generated by cohort-based studies. Also, it supports additional results in promoting the completeness of the International HapMap Project.

This paper presents a guide to using those tools, including the following: retrieving genotype and frequency data, picking tag-SNPs for use in association studies, viewing haplotypes graphically, and examining marker-to-marker LD patterns.

Introduction to the KHapMap Project Website

The results of the Korean HapMap Project are displayed at http://www.khapmap.org, named the KHapMap Project Website. The website is composed of six parts on display (Fig. 1A). The main menus are located in the upper part. "Home" is the starting page of the website. "Introduction" and "Projects" contain short descriptions and tutorials that concern the HapMap Project. "Services" is the starting point to the software that was developed during the project. "Data" allows access to the project's raw data for searching, downloading, and uploading. Users can search for SNP rsIDs in two ways: (i) rs number, and (ii) chromosome number and position with both a user-set minor allele frequency (MAF) and a user-set minimum number of genotyped samples (Fig. 1B). A search result shows rsID links to the dbSNP database, allele, MAF, and genotype counts (Fig. 1C). "Links" connects the user to the working groups of the project and other related sites.

As an auxiliary menu, "Quick Link" is located on the left part of the web page. "Quick Link" holds shortcuts to the programs that users may use more often, such as KHapMap Browser, HapMapSnp Analyzer, SNPflank, D2GSNP, FESD, DOMAINSNP, and KMSD. Most tools in Quick Link can be operated in the same window, except for FESD and KMSD, which are too large to operate in the same frame. In programming, in addition to Java/JSP, JSF (Java Server Faces) was adapted to the KHapMap Website in order to simplify building user interfaces for JavaServer applications. As a result, the website can be quickly and flexibly built by assembling reusable UI components in a page and connecting these components to application data sources (*The Sun Microsoft website*, 2008).

Sources for Korean HapMap

Genomic DNA from 90 unrelated Korean individuals for the Korean HapMap Project was generously provided from the Genomic Research Center in the Korean National Institute of Health (KNIH). The KNIH samples were collected as part of an epidemiological project; the samples represent urban and rural regions that are south of Seoul, Korea. The sex ratio was 0.5, and mean age was 50 years. Informed consent from all participating subjects was obtained through KNIH, and research approval came from the relevant ethics committees. DNA was isolated from peripheral blood leukocytes according to standard procedures with proteinase K-RNase digestion, followed by phenol-chloroform extraction.

Effective tools developed for data analysis

HapMapSnp Analyzer allows users to analyze HapMap data visually and extract information that is related to the HapMap data with ease. SNPflank is a web-based tool that provides flanking sequences of given SNPs to help PCR primer design. D2GSNP is a web-based server system that is designed to find SNPs within those genes that are known to be related to human diseases. With the web interface of D2GSNP, users can find SNPs within genes that are responsible for certain diseases and obtain their flanking sequences for further genotyping experiments (Kang and Hong *et al*. 2006). In D2GSNP, OMIM, GeneCards, and MeSH information were integrated and analyzed for medical research scientists.

FESD is a web-based integrated database for selecting sets of SNPs in putative functional elements in human genes (Kang and Choi *et al*. 2005; Kim and Kim *et al*. 2007). It provides a set of SNPs that are located in 10 different functional elements. SNP@Domain is a web resource of SNPs within protein domain structures and sequences (Han and Kang *et al*. 2006). Users can investigate SNPs within protein domains with 2D and 3D maps, if available. KMSD is a database that provides information on monomorphic SNPs that are found solely in the Korean population. It is composed of a SNP-searching tool for monomorphic SNP information by rsID and a drawing tool for viewing allele frequency pie graphs that compare Korean, Japanese, Chinese, European, and African populations.

The KHapMap Browser provides interactive access to the HapMap database through a graphical genome browser (Fig. 1D). The browser allows users to search the genome for a gene or a region of interest and then visualize the corresponding distribution of SNPs and patterns of common variation in the region. It also provides facilities for downloading SNP assay information, genotypes, and allele frequency information and for generating customized sets of tag-SNPs for association studies. Users can compare Korean SNP data with oth-

er population descriptors (European (CEU), Chinese (CHB), Japanese (JPT), African (YRI), and Korean (KOR) populations). The browser allows users to download genotyping data across a region in a format that is suitable for Haploview analysis (Barrett *et al*. 2005). It is expected that researchers will be able to use the browser to retrieve useful information and to find valuable biomarkers and genes in complex disease association studies. They may use the information to study and develop new drugs in the era of personalized medicine.

Acknowledgements

References

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-265.

Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V.A. (2002). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 30, 52-55.

Han, A., Kang, H.J., Cho, Y., Lee, S., Kim, Y.J., and Gong, S. (2006). SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. *Nucleic Acids Res*. 34, W642-W644.

Kang, H., Choi, K.O., Kim, B.D., Kim, S., and Kim, Y.J. (2005). FESD: a functional element SNPs database in human. *Nucleic Acids Res*. 33, D518-D522.

Kang, H., Hong, T.H., Chung, W.H., Kim, Y.U., Jung, J.H., Hwang, S.H., Han, A.R., and Kim, Y.J. (2006). D2GSNP: a web server for the selection of Single Nucleotide Polymorphisms within human disease genes. *Genomics & Informatics* 4, 45-47.

Kim, H.J., Kim, I.H., Shin, K.H., Park, Y.K., Kang, H., and Kim, Y.J. (2007). FESD II: a revised functional element SNP database of human ethnicities. *Genomics & Informatics* 5, 188-193.

Lee, J.E., Jang, H.Y., Kim, S., Yoo, Y.K., Hwang, J.J., Jun, H.J., Lee, K., Son, O., Yang, J.M., Ahn, K.S., Kim, E., Lee, H.W., Song, K., Kim, H.L., Lee, S.G., Yoon, Y., Kimm, K., Han, B.G., Oh, B., Kim, C.B., Jin, H., Choi, K.O., Kang, H., and Kim, Y.J. (2008). Chromosome 22 LD map comparison between Korean and other populations. *Genomics & Informatics* 6, 18-28.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., *et al*. (2002). The generic genome browser: A building block for a model organism system database. *Genome Res*. 12, 1599-1610.

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.

Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. (2005). The International HapMap Project Web site. *Genome Research* 15, 1592-1593.

Yoo, Y.K., Ke, X., Hong, S., Jang, H.Y., Park, K., Kim, S., Ahn, T., Lee, Y.D., Song, O., Rho, N.Y., Lee, M.S., Lee, Y.S., Kim, J., Kim, Y.J., Yang, J.M., Song, K., Kimm, K., Weir, B., Cardon, L.R., Lee, J.E., and Hwang, J.J. (2006). Fine-scale map of encyclopedia of DNA elements regions in the Korean population. *Genetics* 174, 491-497.

Websites

McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). (2000). Online mendelian inheritance in man, OMIM (TM). *http://www.ncbi.nlm.nih.gov/omim/*.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *http://www.ncbi.nlm.nih.gov/SNP/*.

The dbSNP website. (2008). *http://www.ncbi.nlm.nih.gov/projects/SNP*.

The International HapMap website. (2008). *http://www.hapmap.org*.

The Sun Microsoft website. (2008). *http://java.sun.com/javaee/javaserverfaces*.