BMB
reports

KAREBrowser: SNP database of Korea Association REsource Project

Chang Bum Hong[1,#], Young Jin Kim[1,2,#], Sanghoon Moon[1,#], Young-Ah Shin[1,#], Yoon Shin Cho[1] & Jong-Young Lee[1,]*

[1]Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do 363-951,
[2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea

The International HapMap Project and the Human Genome Diversity Project (HGDP) provide plentiful resources on human genome information to the public. However, this kind of information is limited because of the small sample size in both databases. A Genome-Wide Association Study has been conducted with 8,842 Korean subjects as a part of the Korea Association Resource (KARE) project. In an effort to build a publicly available browsing system for genome data resulted from large scale KARE GWAS, we developed the KARE browser. This browser provides users with a large amount of single nucleotide polymorphisms (SNPs) information comprising 1.5 million SNPs from population-based cohorts of 8,842 samples. KAREBrowser was based on the generic genome browser (GBrowse), a web-based application tool developed for users to navigate and visualize the genomic features and annotations in an interactive manner. All SNP information and related functions are available at the web site http://ksnp.cdc. go.kr/karebrowser/. [BMB reports 2012; 45(1): 47-50]

INTRODUCTION

The frequency of SNPs has been made available through the efforts of the SNP Consortium (TSC) and HapMap (1, 2). The HapMap database includes four million SNP markers in 11 populations through phase I to III. Sets of tagging SNPs and common variants are available on commercial genotyping chips. Indeed, the advancement of microarray technology has enabled the researcher to identify numerous genomic loci accounting for the various aspects of phenotypes. Information, including genotype frequencies, linkage disequilibrium (LD), and recombination rates, across populations helps researchers conduct GWA analysis using millions of SNP markers. The differences in association results among populations for phenotypes of interest are partially explained by HapMap information such as population specific common variants and linkage disequilibrium blocks (3-5). Moreover, the phased haplotypes of HapMap samples are used as a reference for imputing untyped markers. One million SNPs can be increased to up to 2.5 million by imputing haplotypes from HapMap phased haplotypes based on the pattern of observed genotypes (6). Imputation methods enhance the resolution of association mapping for identifying susceptible loci for a particular phenotype. Although HapMap provides the characteristics of common variants in 11 populations of 1,184 individuals, there is still a lack of reference information for common variants for populations not available in HapMap. In imputation for such populations, investigators usually use the most closely located HapMap population as a reference (7). However, studying close populations often do not allow a full understanding of the genetic structure of the specific population. Therefore, population specific information for common variants is needed to avoid misleading conclusions. Previously, Koike *et al*. constructed the public repository database of 700 Japanese control samples (8). In this study, we built a database of common variants for a Korean population comprising 8,842 samples with 352,228 SNP markers and 1.5 million imputed genotypes.

*Corresponding author. Tel: +82-43-719-8870; Fax: +82-43-719-8908; E-mail: leejy63@nih.go.kr
#These authors contributed equally to this work and are listed in alphabetical order.
http://dx.doi.org/10.5483/BMBRep.2012.45.1.47

Received 5 August 2011, Revised 22 August 2011, Accepted 5 October 2011

Keywords: Database, GWAS, Human genome, SNP

RESULTS AND DISCUSSION

We constructed KAREBrowser by integrating GBrowse and genotype databases including KARE and HapMap data. In KAREBrowser, 1.8 million imputed and genotyped SNP data from KARE samples and four million SNP data from HapMap samples are available for investigators who are interested in GWAS and population genomics for Asian populations. KAREBrowser can be browsed using all kinds of internet web browsers and is available at the web site http://ksnp.cdc.go.kr/karebrowser/.

Database and browser

We integrated data combining our genome-wide genotyping data, imputed data, and HapMap genotyping data with the Generic Genome Browser (GBrowse). GBrowse is a combination of database and interactive web. This application tool has useful modules for manipulating and displaying annotations of the genome

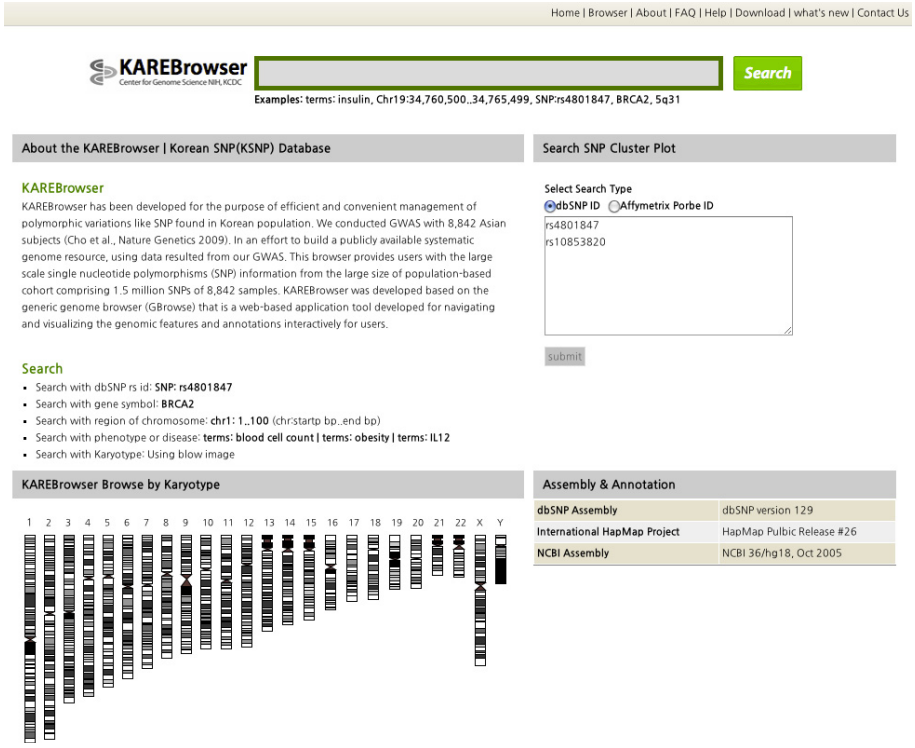

Fig. 1. KAREBrowser screenshot showing the main page.

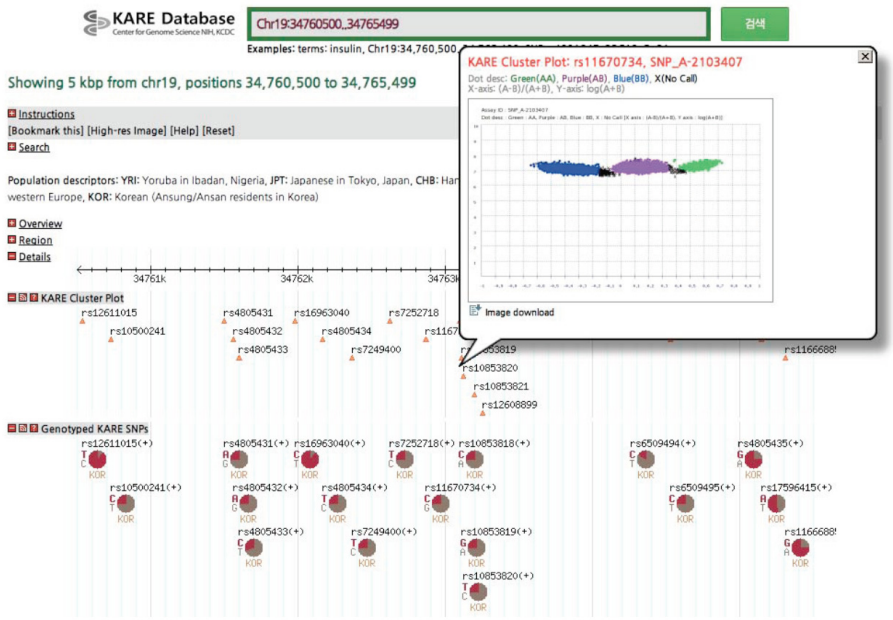

Fig. 2. KAREBrowser screenshot showing the search results.

(9). Using the modules of GBrowse, we successfully integrated the genotype data from 8,842 Korean samples into four million SNPs of HapMap phase II samples. Genotype frequencies, allele frequencies, and linkage disequilibrium blocks are visualized by graphical modules of GBrowse.

Graphical user interface

A particular genome position can be assessed by keywords,

such as chromosome, genomic position, official gene symbol, dbSNP ID, gene accession numbers of reference gene, mRNA, and protein, and part of description of gene (Fig. 1). Users can search any keyword in a simple and convenient way using the specific search formats as shown at the middle left in Fig. 1. For example, SNP : rs12345 can be used for SNP search, BRCA2 for gene search, and chr1 : 1..1,000 for accessing a specific chromosomal region. As shown in the bottom left of Fig. 1, users can visit the specific location of each chromosome by clicking the designated region on the chromosomes. Moreover, users can view the genotype quality of SNPs for KARE samples by inspecting cluster plots of each SNP. On the first page of the browser, users can download cluster plots of SNPs by submitting a list of dbSNP IDs (Fig. 1). Users can visit the specific genomic location and view the genetic structure of Korean samples, along with four HapMap populations (Fig. 2). Users can view allele frequency, genotype frequency, and linkage disequilibrium (LD) information by selecting designated modules options. As shown in Fig. 2, the information related to SNPs is displayed as a track, along with chromosomal position. For example, a colored circle of the track at the bottom in Fig. 2 represents the allele frequency of a SNP and each allele is color coded. Detailed information is available by clicking the SNPs or by hovering over the SNPs. The genotype cluster plot can also be accessed by hovering the mouse cursor over the SNPs on the first track below the chromosomal position. The LD image is made by pairwise R-square values of SNPs.

MATERIALS AND METHODS

Genome-wide genotyping data

We have conducted GWAS of Korean individuals recruited from two population-based cohorts as part of the Korean Genome Epidemiology Study (KoGES). Standardized examinations were applied to 10,038 participants, aged 40 to 69 years. Genomic DNA genotyped on the Affymetrix Genome-Wide Human SNP array 5.0 (Affymetrix, Inc., Santa Clara, CA, USA) were isolated from peripheral blood drawn from participants. From 9,603 genotyped samples, we excluded samples with a high missing genotype rate (>4%, n = 401), high heterozygosity (>30%, n = 11), and gender inconsistencies (n = 41). Individuals who had any kind of cancer (n = 101) were also excluded. Additional exclusions for related or identical individuals whose computed average pairwise identity-by-state (IBS) value was higher than that estimated from first degree relatives of Korean sib-pair samples (>0.80, n = 601) resulted in 8,842 individuals for subsequent analyses. Methods to estimate heterozygosity and IBS have been described elsewhere (10). SNP markers with high missing genotype rate (>5%), low MAF (<0.01), and significant deviation from Hardy-Weinberg equilibrium (P < 1E-6) were excluded, leaving a total of 352,228 markers.

Genotype imputation

Imputation was carried out using IMPUTE program (6). On the basis of NCBI build 36 and dbSNP build 126, we initially used 90 individuals from JPT and CHB founders in HapMap as a reference panel comprising about 3.99 million SNPs (release 22). After removing SNPs with MAF < 0.01 and SNP missing rate > 0.05, we combined the remaining 1.8 million imputed SNPs with the directly typed KARE SNPs for constructing the database.

Hapmap genotyping data

Four million SNPs of four populations (HapMap phase II) were downloaded from the HapMap website. We converted the genotype data to PLINK binary genotype data using PLINK (11). Genotype frequencies, allele frequencies, and P values of Hardy-Weinberg equilibrium were calculated using PLINK.

Acknowledgements

This work was supported by grants from Korea Centers for Disease Control and Prevention (090-091-4800-4845-301) and an intramural grant from the Korea National Institute of Health (2010-N73002-00), the Republic of Korea.

REFERENCES

1. Thorisson, G. A. and Stein, L. D. (2003) The SNP Consortium website: past, present and future. *Nucleic Acids Res.* **31**, 124-127.
2. International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426**, 789-796.
3. Kato, N., Takeuchi, F., Tabara, Y., Kelly, T. N., Go, M. J., Sim, X., Tay, W. T., Chen, C. H., Zhang, Y., Yamamoto, K., Katsuya, T., Yokota, M., Kim, Y. J., Ong, R. T., Nabika, T., Gu, D., Chang, L. C., Kokubo, Y., Huang, W., Ohnaka, K., Yamori, Y., Nakashima, E., Jaquish, C. E., Lee, J. Y., Seielstad, M., Isono, M., Hixson, J. E., Chen, Y. T., Miki, T., Zhou, X., Sugiyama, T., Jeon, J. P., Liu, J. J., Takayanagi, R., Kim, S. S., Aung, T., Sung, Y. J., Zhang, X., Wong, T. Y., Han, B. G., Kobayashi, S., Ogihara, T., Zhu, D., Iwai, N., Wu, J. Y., Teo, Y. Y., Tai, E. S., Cho, Y. S. and He, J. (2011) Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat. Genet.* **43**, 531-538.
4. Soranzo, N., Spector, T. D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., Salo, P., Voight, B. F., Burns, P., Laskowski, R. A., Xue, Y., Menzel, S., Altshuler, D., Bradley, J. R., Bumpstead, S., Burnett, M. S., Devaney, J., Doring, A., Elosua, R., Epstein, S. E., Erber, W., Falchi, M., Garner, S. F., Ghori, M. J., Goodall, A. H., Gwilliam, R., Hakonarson, H. H., Hall, A. S., Hammond, N., Hengstenberg, C., Illig, T., Konig, I. R., Knouff, C. W., McPherson, R., Melander, O., Mooser, V., Nauck, M., Nieminen, M. S., O'Donnell, C. J., Peltonen, L., Potter, S. C., Prokisch, H., Rader, D. J., Rice, C. M., Roberts, R., Salomaa, V., Sambrook, J., Schreiber, S., Schunkert, H., Schwartz, S. M., Serbanovic-Canic, J., Sinisalo, J., Siscovick, D. S., Stark, K., Surakka, I., Stephens, J., Thompson, J. R., Volker, U., Volzke, H., Watkins, N. A., Wells, G. A., Wichmann, H. E.,

Van Heel, D. A., Tyler-Smith, C., Thein, S. L., Kathiresan, S., Perola, M., Reilly, M. P., Stewart, A. F., Erdmann, J., Samani, N. J., Meisinger, C., Greinacher, A., Deloukas, P., Ouwehand, W. H. and Gieger, C. (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41**, 1182-1190.

5. Kim, Y. J., Go, M. J., Hu, C., Hong, C. B., Kim, Y. K., Lee, J. Y., Hwang, J. Y., Oh, J. H., Kim, D. J., Kim, N. H., Kim, S., Hong, E. J., Kim, J. H., Min, H., Kim, Y., Zhang, R., Jia, W., Okada, Y., Takahashi, A., Kubo, M., Tanaka, T., Kamatani, N., Matsuda, K., Park, T., Oh, B., Kimm, K., Kang, D., Shin, C., Cho, N. H., Kim, H. L., Han, B. G. and Cho, Y. S. (2011) Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits. *Nat. Genet.* **43**, 990-995.
6. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906-913.
7. [Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. *Annu Rev Genomics Hum. Genet.* **10**, 387-406.
8. Koike, A., Nishida, N., Inoue, I., Tsuji, S. and Tokunaga, K. (2009) Genome-wide association database developed in the Japanese Integrated Database Project. *J. Hum. Genet.* **54**, 543-546.
9. Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.* **12**, 1599-1610.
10. Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H. J., Yoon, D., Lee, M. H., Kim, D. J., Park, M., Cha, S. H., Kim, J. W., Han, B. G., Min, H., Ahn, Y., Park, M. S., Han, H. R., Jang, H. Y., Cho, E. Y., Lee, J. E., Cho, N. H., Shin, C., Park, T., Park, J. W., Lee, J. K., Cardon, L., Clarke, G., McCarthy, M. I., Lee, J. Y., Oh, B. and Kim, H. L. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* **41**, 527-534.
11. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575.