Genomics & Informatics Vol. 6(2) 57-63, June 2008

Development of KHapmap Browser using DAS for Korean HapMap Research

Hoon Jin[1], Seung-Ho Kim[2], Young Uk Kim[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

The Korean HapMap Project has been carried out for the last 5 years since it started in June, 2003. The project generated data for a sum of 1,764,000 Korean SNPs and formally registered the data to the dbSNP of NCBI (*The dbSNP website.* 2008). 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 (CEPH, Yoruba, Han Chinese, and Japanese populations). The KHapmap Browser was developed and integrated to provide haplotype retrieval and comparative study tools of human ethnicities for comprehensive disease association studies (http://www.khapmap.org). On that basis, GBrowse was adopted in the KHapmap Browser for inherent Korean genetic data, and a provision of extended services was pledged with the distributed sequence annotation system (DAS). The dynamic linking service of the KHapmap Browser to other tools in our intranetwork environment provides many enhanced functions over GBrowse without DAS. KHapmap Browser is expected to be an invaluable tool for the study of Korean and international Hapmap data.

Keywords: Korean, SNP, browser, DAS, Hapmap, association

Introduction

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 (The International HapMap Consortium. 2007). In the International HapMap Project, various types of genetic data are being gathered and analyzed from four populations (CEPH, Yoruba, Han Chinese, and Japanese populations). The International HapMap Project developed haplotype maps that describe what these variants are, where they occur in the genome, and how they are distributed among the persons within. 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 of complex diseases, such as diabetes, cancer, and stroke. As a partial result of this project, we generated and registered data for 1,764,000 Korean SNPs to the dbSNP of NCBI (*dbSNP website.* 2008).

The KHapmap Browser that is based on the generic genome browser (GBrowse) provides haplotype retrieval and comparative study tools of human ethnicities for comprehensive disease association studies (Stein *et al.*, 2002). GBrowse is a combination of databases and interactive web pages for manipulating and displaying annotations on genomes. In other words, GBrowse is a web-based application tool that is developed for navigating and visualizing the genomic features and annotations interactively for users. Through it, users can view a certain region of the desired genomes and search for genetic biomarkers. They may conduct a full-text search for most features of the genomes. They also can download SNP assay, genotype, and allele frequency information and generate customized sets of tag-SNPs for their association studies (Thorisson *et al.*, 2005). GBrowse utilizes a web-based display that can be used to show arbitrary features of a nucleotide or protein sequence and can accommodate genome-scale sequences that are megabases in length. The GBrowse system consists of various kinds of software modules and systems, such as web servers, database systems, and Perl libraries.

At present, many biological websites that provide genomic variants or portal services have been developed using GBrowse, including the following: the UCSC Genome Browser (Kuhn *et al.*, 2007), the International HapMap Project (Thorisson *et al.*, 2005), PlasmoDB (*The*

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

ApiDB/EuPathDB Project Team. 2008), WormBase (*The WormBase website.* 2008), the Perlegen Genotype Browser (*The Perlegen Sciences.* 2008), the Database of Genome Variants (*The Department of Genetics and Genomic Biology.* 2008), and the Database of Drosophila Genes & Genomes (*The FlyBase website.* 2008).

Users can describe their annotations by translating them to two types of file formats in GBrowse. The first file format is a GFF (GBrowse file format) format that is designed to be a light and easy way of describing most genomic annotations that range from simple one-element features to complex, multipart features, such as operons and their regulatory and structural elements (Stein *et al.*. 2008). The main limitation of the Bio::DB:: GFF schema is that it relies on a flat coordinate system to represent genomic features and can handle only one-level nesting of sequence features (Stein *et al.*, 2002). The other is the GadFly file format. It was designed to be capable of representing multiple levels of part-subpart relationships and takes advantage of controlled vocabularies to describe feature types and gene functions as a solution for the limitations of GFF (Mungall *et al.*, 2002). Because of the simplicity and fast speed that are used to treat GBrowse, GFF tends to be used more often. The GFF format is a flat tab-delimited file (Stein *et al.*, 2008), and a series of individual data is written by lines. One line contains nine tab-limited attributes; i.e., seqid, source, type, start position, end position, score, strand, phase, and group. If users want to show their genetic variants or annotations using GBrowse, one must provide newly generated data in GFF format to GBrowse.

Currently, one of the major tasks of systems biology is integrating as much experimental and computational information as possible and thereby gaining biological insight into the properties and function of the macromolecules that are under observation (Olason *et al.*, 2005). DAS is one of the systems that support a distributed client-server network protocol. It allows sequence annotations to be decentralized among multiple third-party annotators and integrated on an as-needed basis by client-side software (Robin *et al.*, 2001). DAS clients make requests by fetching a defined URL from a DAS server and accordingly receive simple XML responses using HTTP (Prlić *et al.*, 2007). DAS-mediated data exchange and visualization is heavily used in popular genome browsers like Ensembl (Hubbard *et al.*, 2007), Wormbase (Stein *et al.*, 2000), and GBrowse (Stein *et al.*, 2002).

Another choice of support is the web service that is based on Service Oriented Architecture (SOA). SOA is composed of three components: service provider, consumer (service requestor), and service registry. BioMOBY is an open source research project that aims to generate an architecture for the discovery and distribution of biological data through web services (Mark *et al.*, 2002). BioMOBY is not supported by many groups yet.

The KHapmap Browser, based on GBrowse, provides haplotype retrieval and comparative study tools of human ethnicities for comprehensive disease association studies. In order to provide user-friendly Korean genetic data manipulation and a provision of extended services, we developed the KHapmap Browser and show several cases using the DAS system.

Methods

Architecture and Services

There are five components in constructing the KHapmap Browser system (Fig. 1). The system uses Apache web server, GBrowse, two MySQL database systems, and DAS as the main modules. As middlewares, the Perl and Bioperl modules have important roles in operating GBrowse. Then the KHapmap Browser was operated based on GBrowse and DAS. We developed the KHapmap Browser on a Linux machine (kernel release 2.6.9～67.0.7.ELsmp) that had Intel dual quad processors, 16 GB of memory, and software (Apache 2.0.48, MySQL 5.0.45, Bioperl 1.52, GBrowse 1.68, Perl 5.8.8 and Biodas 1.07).

In order to install the GBrowse system, a web server and a database system were needed as minimal system requirements. We used two independent database systems for faster service. In other words, two databases were located physically in separated nodes on the network, so that Korean hapmap data could be searched at higher speeds. Another important component of the KHapmap Browser was the DAS module. Through DAS, remote resources, such as the UCSC genome database, the Ensembl genome database, and the Perlegen genotype database, could be embedded into the KHapmap Browser system with ease.

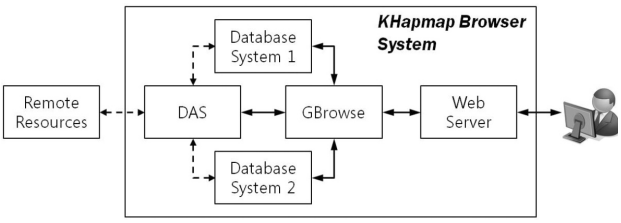


Fig. 1. Architectural flowchart of the KHapmap Browser system. The system comprises five components that use Apache web server, GBrowse, two MySQL database systems, and DAS as the main modules.

The KHapmap Browser was equipped with two services, the first one being a Korean hapmap data browsing service and the second being a mirror service of the International HapMap browser. In the Korean hapmap data browsing service, we showed the results of genotyped Korean SNPs that were produced by the Korean HapMap Project and also various types of analyzed data compared with the International HapMap Project (Lee *et al*., 2008).

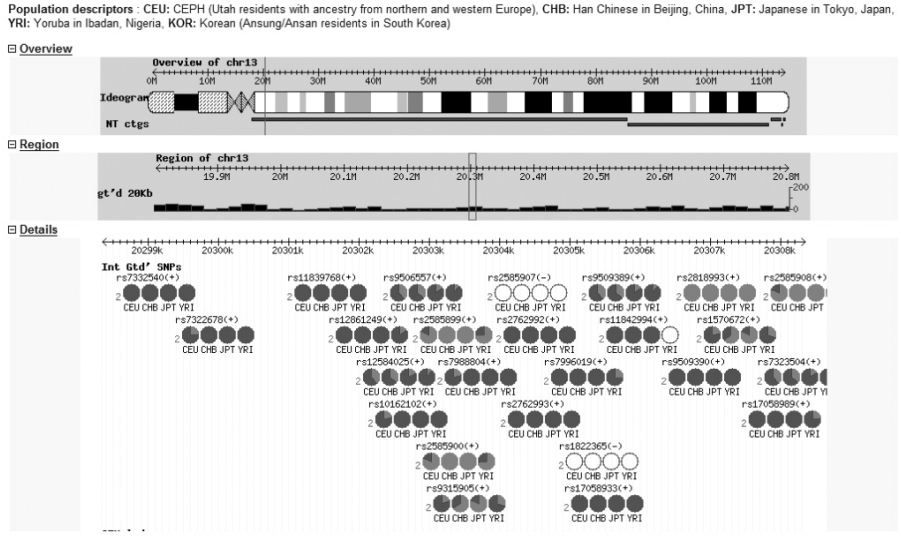
DAS to KHapmap Browser

While there are many kinds of web database systems that are equipped with GBrowse, only a few systems support the DAS system-for example, the International HapMap browser and the UCSC Genome Browser. We embedded DAS to the KHapmap Browser so that it could utilize the already developed tools of the International HapMap browser by linking. Our trials to extend the KHapmap Browser include: 1. Embedding the UCSC Genome databases to our mirror service; 2. Embedding our mirror service to the Korean hapmap data browsing service; and 3. Dynamic linking of the KHapmap Brower to other tools-in this case, FESD, the Functional Element SNPs Database, which categorizes functional elements in human genic regions.

Preparation of Korean HapMap Data

In order to utilize GBrowse, most of the hapmap results were converted into GFF format except for the genotype data. At present, the International HapMap browser provides 36 data attributes through nine tracks. We removed some auxiliary attributes to visualize hapmap re-

(A)



(B)

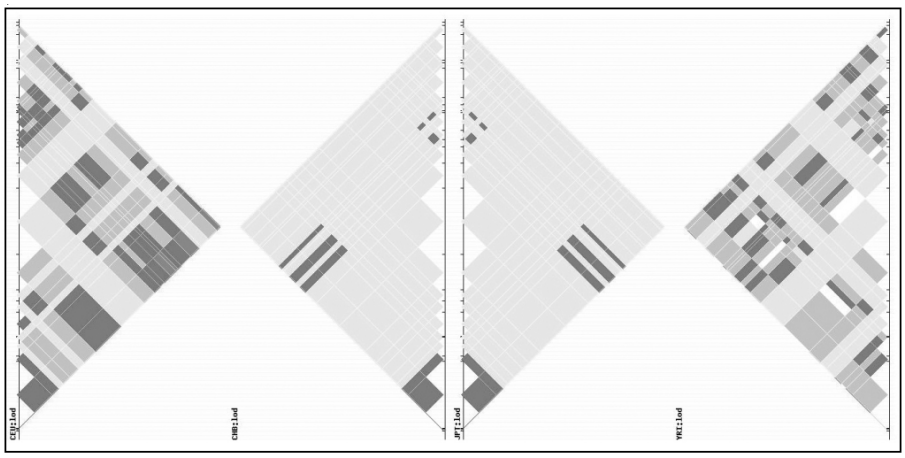


Fig. 2. A mirror service of the International HapMap Browser (http://www.khapmap.org/). (A) A graphical view of SNP allele frequency, (B) Linkage disequilibrium (LD) plots of four populations (European (CEU), Chinese (CHB), Japanese (JPT), and African (YRI) populations).

sults in four data attributes for four populations; i.e., genotyped SNPs, LD plot, Phased Haplotype Display, and CNV-related genomic variants and deletions. The UCSC Genome Browser provides over 100 data attributes through 12 tracks for humans, and we selected five hapmap-related attributes: HapMap SNPs, HapMap LD Phased, Tajima's D SNPs, copy number variations (CNVs), and ENCODE Regions. Considering these, we selected a total of seven attributes for the KHapmap Brower, which were CNVs, genotyped SNPs, average of 10 kb D' and R2 values (Reich *et al*., 2001), LD blocks by the Gabriel method (Gabriel *et al*., 2002), nine ENCODE regions (The ENCODE Project Consortium, 2007), and LD plots (Barrett *et al*., 2005).

To calculate Korean CNV data, we manipulated the segment files of 90 individual samples from 500-kb Affymetrix chips and computed CNVs by using the Redon *et al*. 2006 method (Redon *et al*., 2006). D' value means the difference between the observed frequencies and the expected frequencies of two loci. In this paper, we defined the D' value as the average value of 20-kb moving windows. The R2 value was calculated by dividing the D2 value by four types of allele frequencies at two loci. LD blocks were generated by running Haploview (Barrett *et al*., 2005). All of the data that were calculated were programmed by using the Perl language and were converted to the GFF format.

Results

The KHapmap Browser system comprises five components, which use Apache web server, GBrowse, two MySQL database systems, and DAS as the main modules. Two, physically separated, MySQL database systems were adopted to deal with those complex and huge data of linkage disequilibrium (LD) and other annotated results. The KHapmap Browser can be accessed freely at http://www.khapmap.org/.

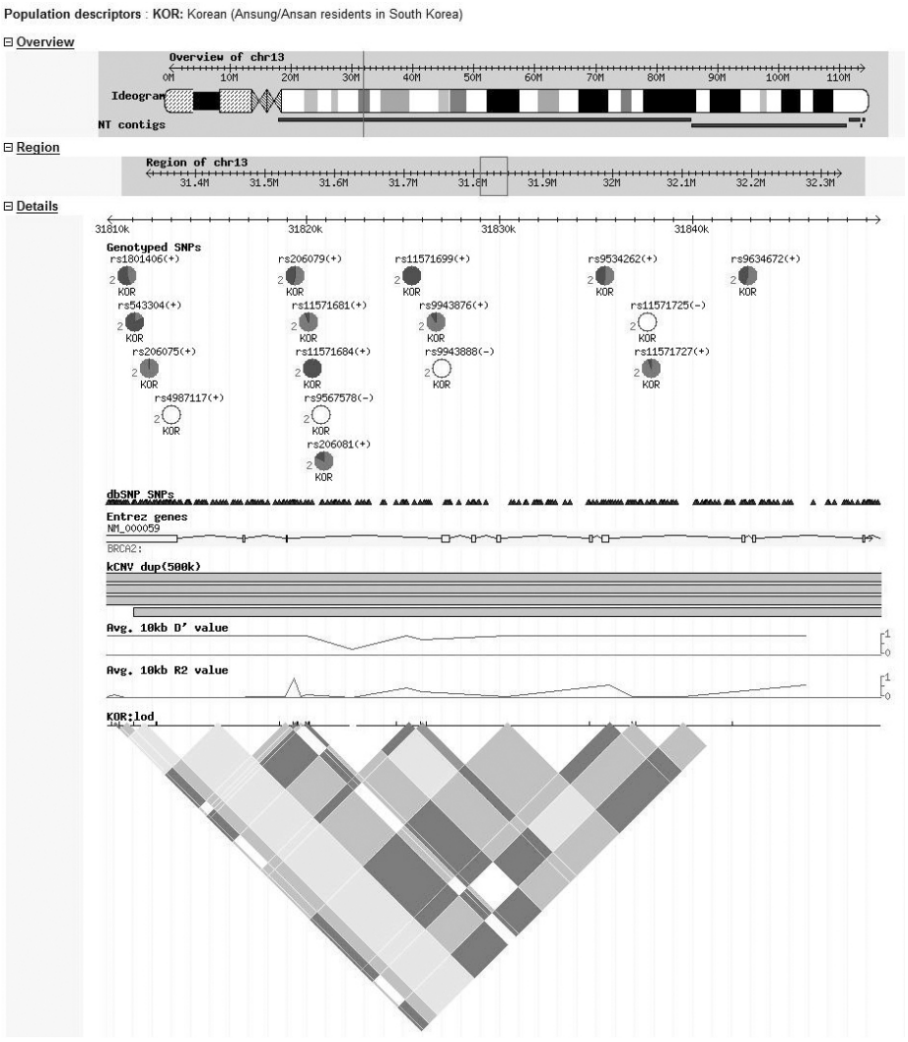


Fig. 3. Search results of the KHapmap Browser. A graphical view of Korean (KOR) SNP allele frequency and features such as copy number variation (CNV), D' value, R2 value, and LD patterns are shown for Koreans.

To service the International HapMap Browser as a mirror service, International HapMap data (NCBI build 36) were downloaded from the designated site (http://www.hapmap.org/downloads/gbrowse/latest/). By using a script that was provided for data input from the URL, all of the GFF files were generated and stored into the generic KHapmap Browse database. A graphical view of SNP allele frequency for the four populations (European (CEU), Chinese (CHB), Japanese (JPT), and African (YRI) populations) is shown in Fig. 2A. LD is known to occur in a block-like structure across the genome, whereby conserved haplotype blocks of tens to hundreds of kilobases are punctuated by "hot spots" of recombination (Daly *et al*,. 2001). LD plots that were drawn for the same locus also are shown (Fig. 2B).

To service those users who are interested in only Korean haplotype data, we constructed a Korean HapMap data browsing service. A graphical view of Korean SNP allele frequency and features, such as copy number variation (CNV), D' value, R^2 value, and LD patterns, clearly were shown for Koreans (Fig. 3).

To utilize other Genome browser services, the local Korean Browser was extended to remote resources, in this case, the UCSC Genome browser by DAS (Fig. 4). Through this process, we give researchers an integrated service for SNP-related studies. In the KHapmap Browser service, users also can see the compared results between Korean and other populations by using DAS (Fig. 5).

We integrated the KHapmap Browser with the FESD system, the Functional Element SNPs Database, which categorizes functional elements in human genic regions (Kang *et al*., 2005; Kim *et al*., 2007) (Fig. 6). The figure shows that the results of FESD may be more powerful and graphically more dynamic by linkage to the KHapmap Browser.

Discussion

The KHapmap Browser has been developed as one of the results of the Korean HapMap Project. The KHapmap Browser utilized GBrowse and DAS in order to bring about three differences compared with the International HapMap Browser. First, we made Korean hapmap data from the genotyped SNPs and processed them so that users could analyze the data better and powerfully. For these purposes, we decided what information should be selected and made for the study of Korean genetic variants as compared with the international hapmap samples. Second, the KHapmap Browser provides two independent services: one is a mirror service of the International HapMap browser (Fig. 2) and the

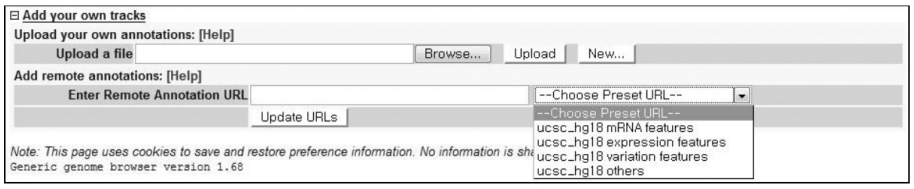


Fig. 4. Track information of the extended mirror service by DAS using the UCSC Genome browser (HG18).

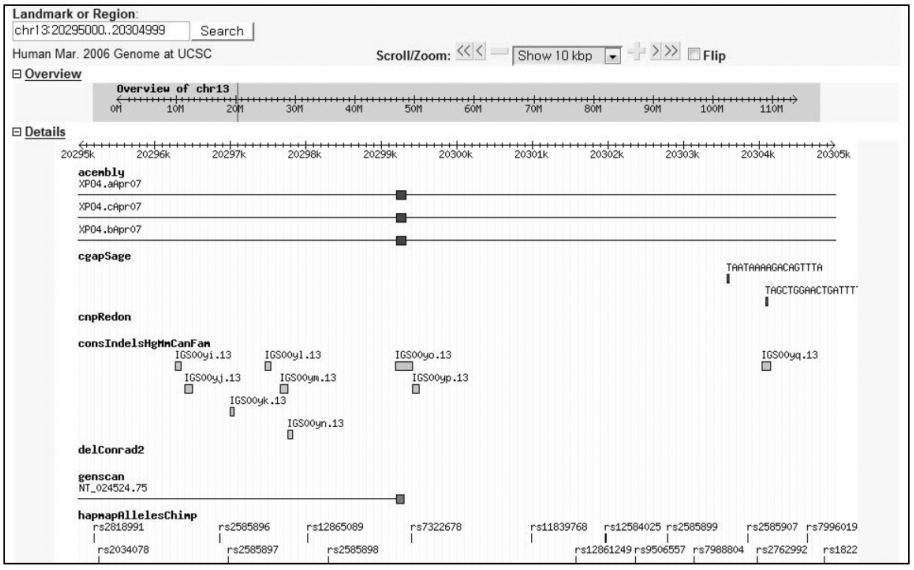


Fig. 5. Imported DAS tracks of the Korean HapMap data browsing service from the UCSC Genome browser.

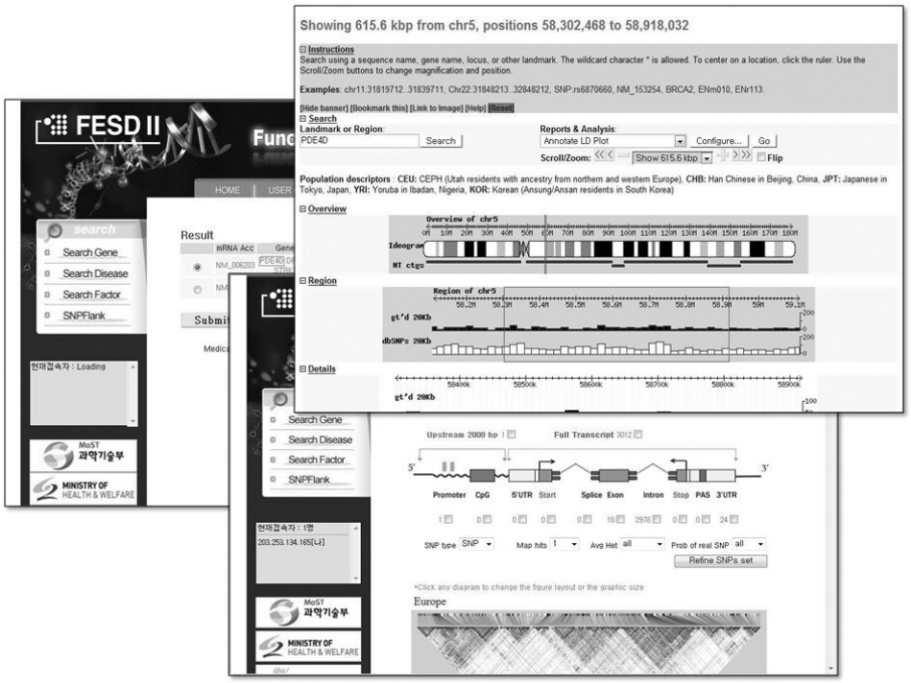


Fig. 6. An application of the KHapmap Browser to FESD, the Functional Element SNPs Database, which categorizes functional elements in human genic regions.

other is the Korean HapMap data browsing service (Fig. 3). Also, the services were embedded together for complementary uses by DAS. Third, through the dynamic link to other tools, we may provide more powerful and valuable results, as shown in Fig. 6. Any remote biological resources that support DAS also can be embedded for extended services. The dynamic linking service of the KHapmap Browser to other tools in our intranetwork environment provides many enhanced functions over GBrowse without DAS. The KHapmap Browser is expected to be an invaluable tool for the study of Korean and International Hapmap data (Lee *et al*., 2008).

Conclusion

During the Korean HapMap Project, we generated data for a total number of 1,764,000 Korean SNPs and formally registered them to the dbSNP. The KHapmap Browser was developed and integrated to provide haplotype retrieval and comparative study tools of human ethnicities for comprehensive disease association studies. Based on GBrowse, the KHapmap Browser utilized DAS in order to provide both powerful views of inherent Korean genetic data and a provision for extended services. A dynamic linking service of the KHapmap Browser to other tools in our intranetwork environment may be profitable and flexible because it can extend the originally limited functions of the tools. The KHapmap Browser is expected to be an invaluable tool for the study of Korean and International Hapmap data.

Acknowledgments

The authors thank Aravinda Chakravarti, Peter Chen, and Carl Kashuk at McKusick-Nathans Institute of Genetic Medicine at Johns Hopkins University for valuable advice and comments on the manuscript. The authors gratefully acknowledge the partial financial support of the Korean HapMap Project of the Ministry of Education, Science and Technology (MEST), and the Sasang Constitution SNP Database Project of Korea Research Council of Fundamental Science & Technology.

References

Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., *et al*. (2001). Extent and distribution of linkage disequilibrium in three genome regions. *Am. J. Hum. Genet.* 68, 191-197.

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.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. Scientific American Magazine.

Mungall, C.J., Misra, S., Berman, B.P., Carlson, J., Frise, E., Harris, N., Marshall, B., Shu, S., Kaminker, J.S., Prochnik, S.E., Smith, C.D., Smith, E., Tupy, J.L., Wiel, C., Rubin, G.M., and Lewis, S.E. (2002). An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biology* 3, research0081.1-0081.11

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and

Lander, E.S. (2001). High resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229-232.

Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R., and Stein, L. (2001). The distributed annotation system. *BMC Bioinformatics* 2, 7.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., De-Felice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225-2229.

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.

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.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., Thakkapallayil, A., Sugnet, C.W., Stanke, M., Smith, K.E., Siepel, A., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pedersen, J.S., Hsu, F., Hinrichs, A.S., Harte, R.A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G.P., Baertsch, R., Haussler, D., and Kent, W.J. (2007). The UCSC genome browser database: update 2007. *Nucleic Acids Res.* 35, D668-673.

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.

Olason, P.I. (2005). Integrating protein annotation resources through the Distributed Annotation System. *Nucleic Acids Res.* 33, W468-470.

Prlić, A., Down, T.A., Kulesha, E., Finn, R.D., Kahari, A., and Hubbard, T.J. (2007). Integrating sequence and structural biology with DAS. *BMC Bioinformatics* 8, 333.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., *et al.* (2006). Global variation in copy number in the human genome. *Nature* 444, 444-454

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., and Lander, E.S. (2001). Linkage disequilibrium in the human genome. *Nature* 411, 199-204.

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.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2000). WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.* 28, 82-86.

Stein, L.D., and the GMOD development team. (2008). GBrowse Configuration HOWTO.

Subramaniam, S. (1998). The Biology Workbench--a seamless database and analysis environment for the biologist. *Proteins* 32, 1-2.

Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. (2005). The International HapMap Project Web site. *Genome Research* 15, 1591-1593.

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

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

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

The ApiDB/EuPathDB Project Team. (2008). The PlasmoDB web site. *http://www.plasmodb.org/plasmo.*

The Department of Genetics and Genomic Biology. (2008). *http://projects.tcag.ca/variation/cgi-bin/gbrowse/hg18.*

The ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799-816.

The FlyBase web site. (2008). *http://flybase.net.*

The Perlegen Sciences. (2008). *http://genome.perlegen.com/browser/index_v2.html.*

The WormBase web site. (2008). *http://www.wormbase.org.*