The International HapMap Project Web site

Gudmundur A. Thorisson,[1,2,3] Albert V. Smith,[1,2] Lalitha Krishnan,[1] and Lincoln D. Stein[1]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA*

The HapMap Web site at http://www.hapmap.org is the primary portal to genotype data produced as part of the International Haplotype Map Project. In phase I of the project, >1.1 million SNPs were genotyped in 270 individuals from four worldwide populations. The HapMap Web site provides researchers with a number of tools that allow them to analyze the data as well as download data for local analyses. This paper presents step-by-step guides to using those tools, including guides for retrieving genotype and frequency data, picking tag-SNPs for use in association studies, viewing haplotypes graphically, and examining marker-to-marker LD patterns.

[Supplemental material is available online at www.genome.org.]

The goal of the International HapMap Project (International HapMap Consortium 2005) is to map and understand the patterns of common genetic diversity in the human genome in order to accelerate the search for the genetic causes of human disease. The first major milestone of the project was the genotyping of 1.1 million SNPs across four populations, a goal reached in the spring of 2005. Another 4.6 million SNPs are being genotyped in the second phase of the project, and are scheduled for completion in fall 2005.

The project data are available for unrestricted public use at the HapMap Web site, located at http://www.hapmap.org. This site offers bulk downloads of the data set, as well as interactive data browsing and analysis tools that are not available elsewhere. Since it was opened to the public in November 2003, the HapMap data set has been downloaded >500,000 times by researchers in >100 countries. The site currently serves >30,000 static page requests per month, of which 14,000 are bulk download requests, and >100,000 accesses per month to the interactive HapMap browser.

This paper describes the Web site and the tools that have been developed for viewing, retrieving, and analyzing the project data.

The HapMap Web site

The HapMap Web site at http://www.hapmap.org (Fig. 1) is organized into three main sections, accessible from the banner at the top of the page. Reflecting the international nature of the project, the home page and much of the internal site is available in the languages of the countries that participated in the project: English, French, Chinese, Japanese, and Yoruba. The Web site automatically selects the language to display based on the user's browser settings.

The site's "Home Page" gives an overview of the project and lists project news. Users will also find links here to recent publications, events of interest, related projects, and affiliated Web sites.

The "About the Project" section describes the HapMap project in more detail. It provides an introduction to genetic association mapping, describes the ethical issues raised by the project and how they were addressed, and provides guidelines for using HapMap data. This is also the place to find background information on the populations sampled for the project and to obtain project protocols.

The "Data" section is the largest part of the Web site. It provides bulk downloads of HapMap data and analysis sets as well as interactive access to the HapMap database. The Supplemental material for this paper provides detailed "recipes" for using the facilities available in the Data section to study patterns of common variation in the human genome and to generate sets of SNP-based markers suitable for genetic association studies.

Interactive access to the data

The Data section provides interactive access to the HapMap database via a graphical genome browser (Stein et al. 2002). The browser allows users to search the genome for a gene or region of interest and then to visualize the 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. These facilities are described in Supplemental information Recipes #1–#8.

Another feature available through the genome browser allows users to download genotyping data across a region in a format suitable for analysis using the desktop application Haploview (Barrett et al. 2005). This is described in Recipe #9.

Data mining

The HapMart facility allows users to generate genome-wide extracts of the HapMap data set based on combinations of criteria, such as whether a SNP causes a non-synonymous amino acid change, or what its degree of polymorphism is in a selected population. Based on the BioMart data warehousing system (Gilbert 2003), this facility is an attractive alternative to downloading and filtering the entire data set manually. It is described in Recipe #10.

Bulk download of the data

Lastly, the Data section contains a link to a bulk download section where the entire HapMap data set can be download as a series of text files. Available data include assay design information, allele frequency information, raw genotypes, and analytic results including pairwise linkage disequilibrium between SNPs

[2]These authors contributed equally to the work.
[3]Corresponding author.
E mail mummi@cshl.edu; fax (516) 367-8389.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.4413105. Freely available online through the *Genome Research* Immediate Open Access option.

15:1592–1593 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

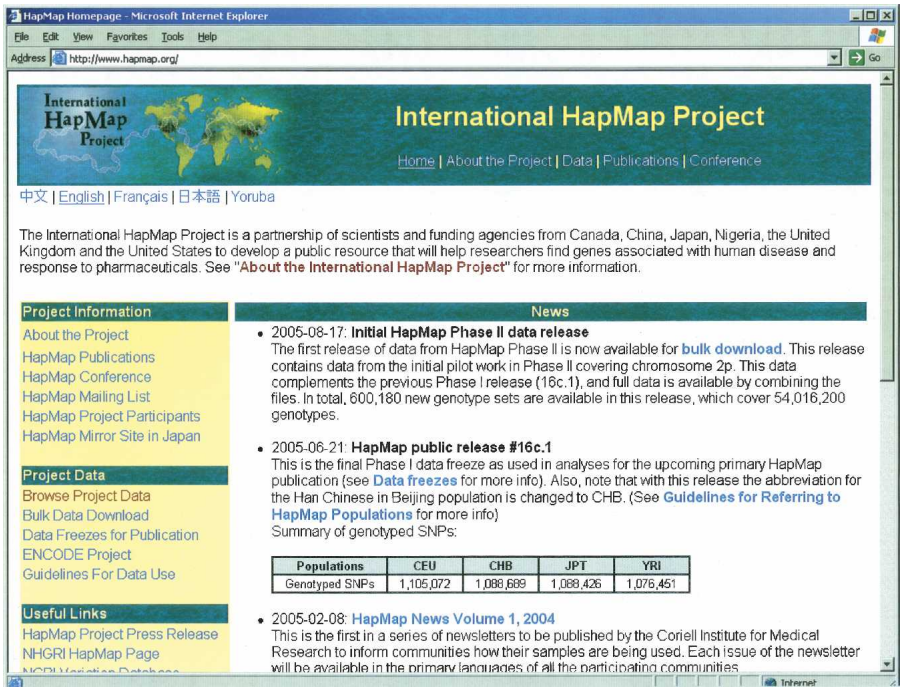

Figure 1. The HapMap Web site. This image shows the home page of the HapMap Web site. Links to the major subdivisions of the site are located in the banner at the *top* of the page as well as the navigation panel on the *left* border.

and phased haplotypes. Access to these data sets is described in Recipe #11.

Discussion

A number of public online resources have been developed as portals to high-volume genome-wide data sets. The UCSC Genome Browser (Kent et al. 2002) and the Ensembl project (Birney et al. 2004) have developed multispecies genome browsers that display genomic annotations graphically and offer retrieval of the underlying data. dbSNP (Wheeler et al. 2005) is a repository for information on SNPs, but does not yet contain extensive information on the relationships among them.

The HapMap Web site has a distinct focus. It aims to be a resource in the display, retrieval, and analysis of high-throughput, high-quality, genome-wide human genetic data, with an emphasis on the support of tools for facilitating disease association studies. Although the resource is still in development, it currently provides the basic tools for visualizing patterns of common polymorphism among the populations surveyed by the HapMap project, selecting tag-SNP sets based on a variety of criteria, and generating customized extracts of the data set.

In the future, the HapMap Web site will evolve to provide more services to those designing and interpreting genetic association studies. In the near future, we will integrate the HapMap genome browser more tightly with other genome browsers, for example by sharing tracks with the UCSC Genome Browser and Ensembl projects. This will provide researchers with the ability to see HapMap data in the context of many other genomic features, particularly those relating to evolutionary conservation. Over a somewhat longer term, we will provide tools that will allow researchers to upload genetic association data (in a secure and anonymous manner) and view association data on top of the LD map, genes, and other genomic features. This feature will be integrated with databases providing information on biological pathways, protein–protein interactions, and known disease genes, allowing researchers to correlate their association data with what is known about the biological processes involving the genes in the region.

We will add to the tag-SNP picker a suite of tools to help researchers create SNP sets tuned for genome-wide association studies, for association studies directed at a particular region or regions, and for different types of study design. We also hope to provide increasingly sophisticated visualization services that assist in interpreting the results of association studies and comparing the results of one association study to another.

Finally, because the BioMart system allows queries to span multiple databases, we will make it possible to perform simultaneous queries across HapMart and the EnsMart genome annotation database at Ensembl. This will allow researchers to make queries that combine Ensembl information (e.g., "find all genes that contain a zinc-finger domain and a strong homolog in mouse") with HapMap queries ("find all tag-SNPs for this list of zinc-finger genes").

Acknowledgments

We thank the four anonymous reviewers for their constructive comments during the preparation of this work. This work was supported by grants from The SNP Consortium and the Genome Institute of the National Institutes of Health.

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.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts T., et al. 2004. An overview of Ensembl. *Genome Res.* **14:** 925–928.

Gilbert, D. 2003. Shopping in the genome market with EnsMart. *Brief. Bioinformatics* **4:** 292–296.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* (in press).

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

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.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., et al. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **33:** D39–D45.

Web site references

http://www.hapmap.org; HapMap Web site.

Received July 11, 2005; accepted in revised form September 6, 2005.