A User's Guide to the International HapMap Project Web Site

Gudmundur A. Thorisson[1]*, Albert V. Smith*, Lalitha Krishnan, and Lincoln D. Stein
Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724

[1] To whom correspondence should be addressed. Email mummi@cshl.edu.
**These authors contributed equally to the work.*

The HapMap website at http://www.hapmap.org is the primary portal to genotype data produced as part of the International Haplotype Map Project (Gibbs et al. 2003). In phase I of the project, over 1.1 million SNPs were genotyped in 270 individuals from 4 worldwide populations (Consortium 2005). The HapMap website provides researchers with a number of tools that allow them to analyze the data as well as to 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.

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 is available for unrestricted public 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 elsewhere available. Since it was opened to the public in November 2003, the HapMap data set has been downloaded over 500,000 times by researchers in more than 100 countries. The site currently serves more than 30,000 static page requests per month, of which 14,000 are bulk download requests, and more than 100,000 accesses/month to the interactive HapMap browser.

This paper describes the website and the tools that have been developed for viewing, retrieving and analyzing the project data. We show readers how to perform several useful and popular tasks, and give an overview of new tools under development or planned for the future.

The HapMap Web Site

The HapMap web site at www.hapmap.org is organized into three main sections, accessible from the banner at the top of the page. The home page gives an overview of the project and lists project news. The "About the Project" section describes the HapMap project in more detail and provides pointers to background information about genetic association mapping, the ethical issues raised by the project, protocols used within the project, and project administration. The "Data" section provides bulk downloads of HapMap data and analysis sets as well as interactive access to the HapMap database. This paper will be concerned almost exclusively with the Data section.

Recipe #1: Browse genotypes using the genome browser

Research into the genetic contributions to a human disease is often centered around a small number of candidate genes. In such a case the researcher will wish to know whether there are any common single nucleotide polymorphisms (SNPs) in the neighborhood of the candidate gene(s), what those SNPs' alleles are, and what are the relative frequencies of the alleles in the population. The researcher will also be particularly interested in coding SNPs (cSNPs), SNPs whose alleles change the amino acid sequence of the gene product and therefore may represent functional variations.

The genome browser at the HapMap web site provides access to small to medium sized-regions of the genome for this type of interactive exploration. This basic recipe shows how to start using the genome browser:

1. Using any modern web browser, go to www.hapmap.org.
2. Click the "Browse Project Data" link under the "Project Data" section of the hapmap.org home page. This page can also be reached directly at http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/.
3. This will take you to a genome browser based on the

GBrowse package (Stein et al. 2002) (Figure 1). Depending on your computer language settings, this page may appear in one of several languages. In this guide, we assume English.

4. Locate the “Landmark or Region” search box, and enter a search term. Any of the following types of search term will work:

    a. A chromosome name (e.g. “Chr10”)

    b. A chromosomal position, in the format *Chromosome:start..stop* (e.g. “Chr9:25000..300000”)

    c. The name of a SNP using its dbSNP “rs” name (e.g. “rs4285800”)

    d. A gene using its NCBI Refseq accession number (e.g. “NM_214279”).

    e. A gene using its common name (e.g. “BRCA2”)

    f. A chromosomal band (e.g. “10q23.1”)

5. After entering one of these landmarks, press the “Search” button (or hit Enter).

6. This will return a page showing the region surrounding the requested feature (Figure 2). If multiple features match, then the page will show you a graphical summary of all possible features and prompt you to choose one. At the top of the returned page is an “Overview” section that shows the cytogenetic map of the selected chromosome. A red box indicates the section of the chromosome in view. Beneath this is a “Detail” section that has horizontal tracks showing various types of data. The two most useful tracks are the “Genotyped SNPs” track that provides information on the position, alleles, and allele frequencies of each SNP characterized by the HapMap project, and the RefSeq mRNAs track, which shows the positions and structures of human protein-coding genes.

7. Use the controls at the top of the page to scroll left, right or to change the magnification of the region. You may also click anywhere on the Overview or the scale at the top of the Detail section in order to center the view on this position.

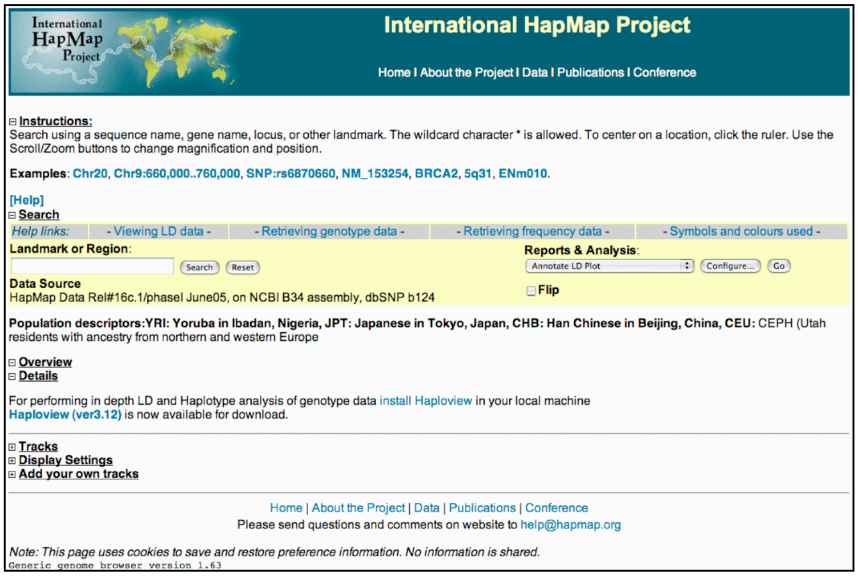
8. The genotyped SNP track changes its appearance in a manner appropriate to the scale of the image. At low magnifications, genotyped SNPs appear as equilateral triangles. The background color of the triangle indicates its coding status: SNPs that introduce non-synonymous amino acid changes are yellow, those that fall in an exon but create a synonymous nucleotide substitution are pink, and those that fall into a non-coding region are cyan. These colors can be customized by selecting the “Highlight SNP Properties” item in the “Reports and Analysis” menu.

9. At higher magnifications, the genotyped SNPs change to display the alleles associated with the SNP. The allele shown in blue is the allele present in the reference genomic sequence at that location, and the red allele is the other allele present in the SNP.

10. When zoomed in still further, the genotyped SNPs track changes to show pie charts representing the allele frequency for each genotyped population. The blue wedge of the piechart indicates the frequency of the allele that appears in the reference genome sequence. The red wedge is the frequency of the alternative allele.

11. Click on the glyph for an individual SNP to see a text-based page with detailed genotype and allele counts, and assay information

The piechart display that appears in step (9) provides the researcher with the ability to easily distinguish SNPs that are



International HapMap Project
International HapMap Project
Home I About the Project I Data I Publications I Conference

⊟ **Instructions:**
Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: Chr20, Chr9:660,000..760,000, SNP:rs6870660, NM_153254, BRCA2, 5q31, ENm010.

[Help]
⊟ Search
Help links:
- Viewing LD data -
- Retrieving genotype data -
- Retrieving frequency data -
- Symbols and colours used -

Landmark or Region:

Reports & Analysis: Annotate LD Plot

Data Source
HapMap Data Rel#16c.1/phaseI June05, on NCBI B34 assembly, dbSNP b124

□ **Flip**

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

For performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
Haploview (ver3.12) is now available for download.

Please send questions and comments on website to help@hapmap.org

Note: This page uses cookies to save and restore preference information. No information is shared.
Generic genome browser version 1.63

Figure 1. Start page for the HapMap Genome Browser
This image shows the HapMap Genome Browser before a region has been loaded. Use the “Data Source” menu to select whether to search the current release (the default), or previous releases. Type a chromosome name, genomic region, gene name, SNP name, or other identifier into the “Landmark or Region” textbox to access the region of interest.

highly polymorphic in all four of the HapMap populations, and therefore more likely to be polymoprhic in other populations as well. Alternatively, the researcher can identify SNPs that are more polymorphic in a single population and are therefore, suitable as markers in population-specific genetic screens.

The detailed view that is obtained when the researcher clicks on a genotyped SNP (step 10) provides the researcher with the information needed to generate an assay for the SNP, including the left and right flanking sequences needed to create PCR primers. A hypertext link to dbSNP (Wheeler *et al.* 2005) provides more information about how the SNP was first discovered and any other population genetic information that may exist for it outside the HapMap project, while a link to Ensembl (Birney *et al.* 2004) leads the researcher to a site where the structural impact of the SNP on coding sequence, splice sites, and other features of nearby genes can be examined.

Recipe #2: Generate a text listing of genotypes using the genome browser

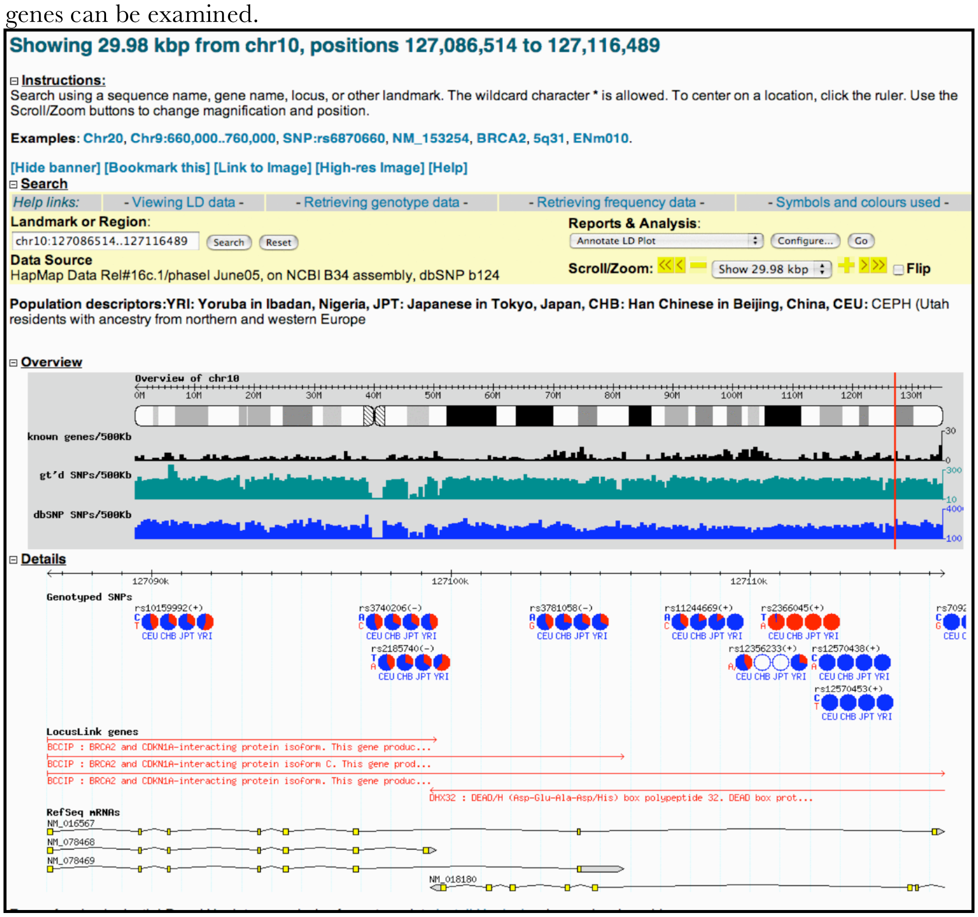
After a researcher has browsed a region graphically and has centered his view on a candidate gene and the region surrounding it, he may want to generate a space-delimited text dump of the genotyping results across this region.dump in the region. This data can then be imported into an Excel spreadsheet or other data analysis tools.

1. Starting at step 6 from recipe #1, navigate to the region of interest.

2. Locate the "Reports and Analysis" menu (above the Detail panel) and select the menu item "Dump SNP genotype data." Next click the "Configure" button. This will open a configuration page that allows you to select the desired HapMap population, and whether to save the data to disk or view it in the web browser.

3. Choose the desired options, and click the "Go" button to retrieve the data and produce a report. The format is the same as the bulk-download files (see below). The dumper configuration settings are stored in a browser cookie, so next time you can click the 'Go' button on the main page and dump data directly, without having to configure the dumper first.

The text-dump format consists of a set of rows containing each SNPs dbSNP ID, the two reference and alternative allele, the position of the SNP on the genome, and the genotypes of the SNP on each of the individuals in the selected HapMap population. Because this format is identical to that used by the bulk downloadable files, it can be easily loaded into the HaploView program (Barrett *et al.* 2005) for detailed analysis on the researcher's local computer.



Showing 29.98 kbp from chr10, positions 127,086,514 to 127,116,489

⊟ Instructions:
Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: Chr20, Chr9:660,000..760,000, SNP:rs6870660, NM_153254, BRCA2, 5q31, ENm010.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help]

⊟ Search

Landmark or Region: chr10:127086514..127116489

Reports & Analysis: Annotate LD Plot

Data Source
HapMap Data Rel#16c.1/phaseI June05, on NCBI B34 assembly, dbSNP b124

Scroll/Zoom: Show 29.98 kbp

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

⊟ Overview

⊟ Details

Figure 2. HapMap Genome Browser displaying BRCA2 region
After typing "BRCA2" into the search box, the genome browser will display information about the region of the genome surrounding this gene. The "Overview" panel provides a birds-eye display of the whole chromosome with the region of interest indicated by a vertical red line. The "Details" panel shows the selected region. Multiple horizontal tracks provide information about genes, SNPs, and genotypes in the region.

Recipe #3: Generate a text listing of genotype frequencies using the genome browser

In a similar fashion, the researcher may wish to download a summary of the frequencies of alleles across a region of interest. The researcher can then select from this set those SNPs that meet certain criteria, for instance those that are most highly polymorphic in a particular population of interest. This recipe describes how to create a tab-delimited summary of HapMap allele frequency data across a particular genomic region.

1. Starting at step 6 from recipe #1, navigate to the region of interest.

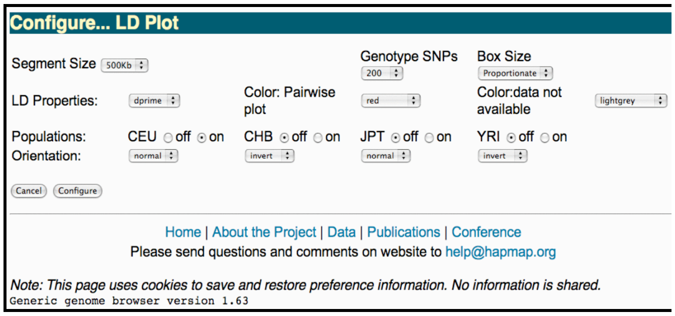
2. Locate the "Reports and Analysis" menu (above the Detail panel) and select the menu item "Dump SNP frequency data." Next click the "Configure…" button. This will open a configuration page that allows you to select the desired HapMap population, and whether to save the data to disk or view it in the web browser.

3. Click the "Go" button to retrieve the data and produce a report. The format is the same as the bulk-download files (see below). The dumper configuration settings are stored in a browser cookie, so next time you can click the 'Go' button on the main page and dump data directly, without having to configure the dumper first.

The report generated by this recipe contains one row for each SNP consisting of each SNP's dbSNP ID, its genomic position, the number of times each possible genotype was seen in the selected population, and the heterozygosity of the population for that SNP.

Recipe #4: View the extent of linkage disequilibrium in a region using the genome browser

When a researcher designs a study to detect the association between a common allelic variation of a gene and a disease of interest, knowledge of the extent of linkage disequilibrium (LD) in the region is essential for reducing the number of SNPs that need to be genotyped across the region. If there is high linkage disequilibrium in the region, then only a few SNPs need to be genotyped because their linkage to other SNPs in the region will serve as proxies for the genotypes of non-characterized SNPs. In contrast, a region of low linkage disequilibrium will need to be sampled more heavily because the allelic state of a genotyped SNP will be a poor predictor of the state of non-genotyped SNPs. The determination of patterns of LD in the four populations characterized by the HapMap project is, in fact, one of the major goals of this project.



Configure... LD Plot

Segment Size: 500Kb

Genotype SNPs: 200

Box Size: Proportionate

LD Properties: dprime

Color: Pairwise plot: red

Color:data not available: lightgrey

Populations: CEU ○ off ● on | CHB ● off ○ on | JPT ● off ○ on | YRI ● off ○ on

Orientation: normal | invert | normal | invert

Please send questions and comments on website to help@hapmap.org

Note: This page uses cookies to save and restore preference information. No information is shared.

```
Generic genome browser version 1.63
```

Figure 3. The configuration screen for the LD Plot viewer.
Use the LD Plot viewer configuration screen to select the size of the region to display LD values over, how many genotyped SNPs to display, what measure of LD to use (r^2, D' or LOD) and which populations to display. The "Box Size" configuration option allows you to select whether the plot will show the distances between SNPs proportionate to their genomic position, or to display the SNPs in the region as if there were a constant distance between them.

The International HapMap Project has precalculated patterns of LD among the genotyped SNPs. The data can be downloaded in bulk from the HapMap web site or browsed interactively using the HapMap genome browser. The latter method allows researchers to see patterns of LD in context with the distribution of genes of interest.

1. To view available LD data precalculated from HapMap genotypes, first browse to a region of interest (see Recipe #1).

2. After browsing to a region of interest, select the "Annotate LD plot" plugin from the "Reports and Analysis" menu and click the "Configure" button to bring up a configuration page that will allow you to adjust the display properties to your liking. Key parameters on this page are the HapMap populations to display, which measure of LD to use (choice of D', r^2 or LOD), whether the triangle plot should be oriented with the vertex pointing upward or downward, color scheme and whether the box size in the plot should be proportional to genomic distance between markers or of uniform size (see Figure 3).

3. After configuring the parameters as desired, click on the 'Configure' button to return to the main display. The display will now show one triangle plot for each population selected (see Figure 4). The triangle plot is constructed by connecting every pair of SNPs along lines at 45 degrees to the horizontal track line. The color of the diamond at the position that two SNPs intersect indicates the amount of LD: more intense colors indicate higher LD. A grey diamond indicates that data is missing. Figure 4 shows a typical region of LD which demonstrates "patches" of high LD separated by relatively well-defined boundaries of low LD.

4. In regions with many genotyped SNPs, the LD plugin adds significantly to the time it takes for the web page to load. You may turn off the LD display at any time to deselecting the appropriate checkbox in the "Tracks" section of the browser. The LD plugin settings are stored in a browser cookie, so there is no need to visit the configuration page each time the plugin is turned on.

The traditional D' and r^2 metrics reflect the degree of pairwise LD between two SNPs, but differ in their sensitivity and specificity across different size scales. See (Mueller 2004) for a discussion of the practical application of these measurements. The LOD metric used by this HapMap web site display is described in (Daly *et al.* 2001).

Recipe #5: Generate a text listing of linkage disequilibrium values using the genome browser

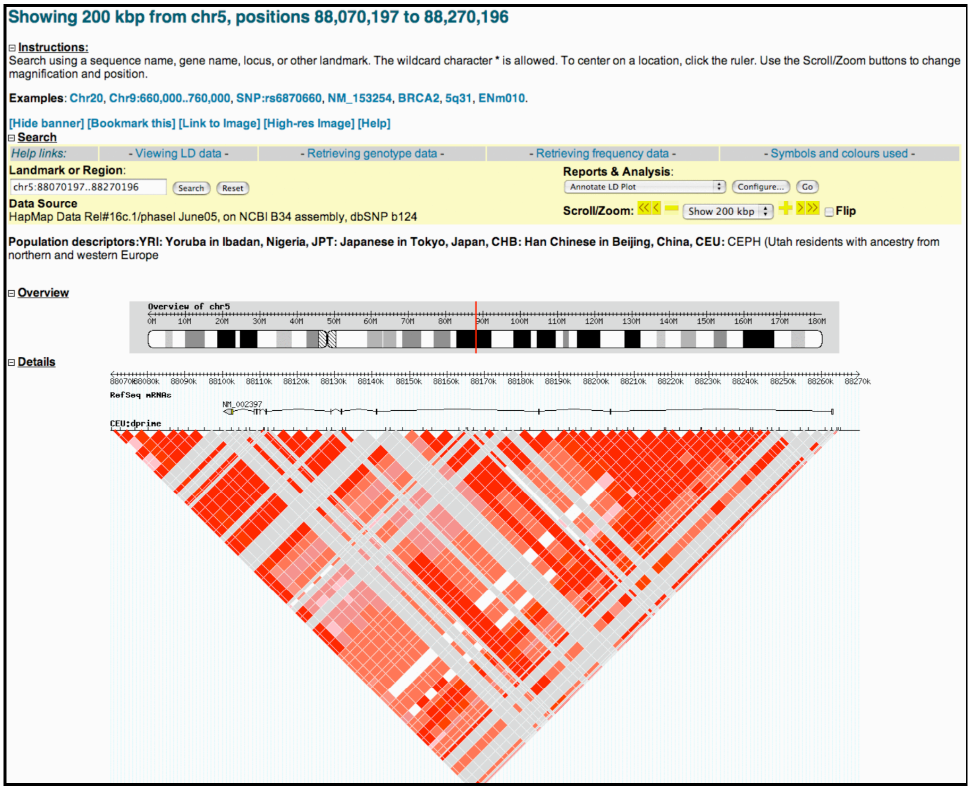
After selecting a region of interest and visually inspecting the extent of LD across the gene or genes of interest, the researcher may wish to download a tab-delimited numeric summary of the LD values in the selected region. This information can be used to select a set of "tag" SNPs that will act as proxies for other SNPs that are in high LD with them.

1. Starting at step 6 from recipe #1, navigate to the region of interest.

2. Locate the "Reports and Analysis" menu (above the Detail panel) and select the menu item "Dump HapMap LD Data". Next click the "Configure…" button. This will open a configuration page that allows you to select which HapMap population to dump data for and whether to save the data to disk or view it in the web browser.

3. Click the "Go" button to retrieve the data and produce a report. The format is the same as the bulk-download files (see below). The dumper configuration settings are stored in a browser cookie, so the next time you wish to perform this operation you can simply click the 'Go' button on the main page and dump the data directly, without having to go through the configuration step.

The report generated by this recipe will show pairwise LD between all SNPs within 500 kb of each other (this window may be reduced in size in future releases of the database). Each row of the report corresponds to one pair of SNPs. The first two columns indicate the positions of the SNPs on the chromosome, the third column is the population for which the LD values were calculated, and the fourth and fifth columns indicate the dbSNP IDs for the



Showing 200 kbp from chr5, positions 88,070,197 to 88,270,196

⊟ **Instructions:**
Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: Chr20, Chr9:660,000..760,000, SNP:rs6870660, NM_153254, BRCA2, 5q31, ENm010.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help]

⊟ **Search**

Landmark or Region: chr5:88070197..88270196

Reports & Analysis: Annotate LD Plot

Data Source
HapMap Data Rel#16c.1/phaseI June05, on NCBI B34 assembly, dbSNP b124

Scroll/Zoom: Show 200 kbp

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

Figure 4. LD viewer displayed on the Genome Browser.
This image shows 200 kbp on chr5 for the CEU samples, positions 88,070,197 to 88,270,196. The strength of the the linkage disequilibrium is shown in increasing shades of red for higher LD values. This particular option is the "LOD" plot option.

SNP pair. These are followed by the D', r^2 and LOD scores for LD between the two SNPs.

Recipe #6: Pick tag-SNPs using the Genome Browser

Tag SNPs are a reduced set of SNPs that capture much of the LD in regions; they can be used in association studies to reduce the number of SNPs needed to detect LD-based association between a trait of interest and a region of the genome. For small regions it is possible to select tag SNPs by hand using the graphical and numeric displays of LD generated by the previous two recipes, but for best results it is recommended that the researcher use an algorithm that chooses tag SNPs by formally maximizing the number of linked SNPs captured by the tag set.

There is no single set of tag-SNPs that will satisfy the diverse requirements of every association study design. Researchers may wish to select SNPs that work well with a particular genotyping system (for example, those that have been included on a particular "SNP chip") and may be willing to accept different tradeoffs between the cost of genotyping a study population and the strength of the association they can detect. For this reason, the HapMap web site does not offer a static set of preselected tag-SNPs, but instead offers researchers a tool for interactively selecting tag-SNPs based on user-provided criteria.

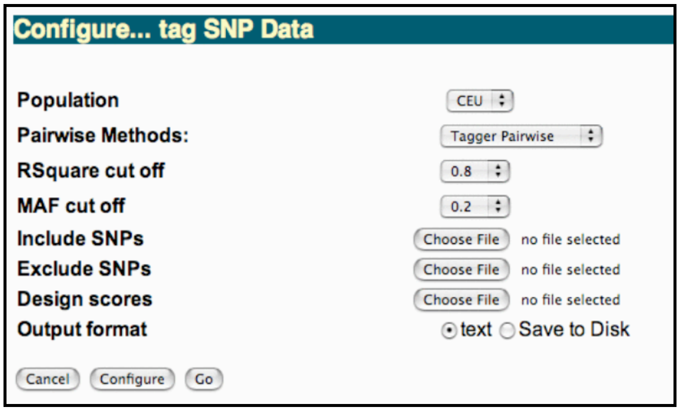
There is no single set of tag-SNPs that will satisfy the diverse requirements of every association study design. Instead of offering a static set of tag-SNPs, the HapMap web site offers a tool for To find tag-SNPs in a region first browse to the region (see Recipe #1).

1. Starting at step 6 from recipe #1, navigate to the region of interest.

2. Select the "Annotate tag SNP Picker" option under the "Reports and Analysis " menu.

3. Press "Configure" to select the desired options for tag SNP selection (see figure 5). Options include selecting a population and an algorithm, uploading a list of SNP IDs to be included in the set of tag-SNPs, uploading a list of SNP IDs to be excluded from the set of tag-SNPs, uploading a list of design scores (priorities) for each SNP, and selecting cutoffs for minimum acceptable LD value and allele frequency for SNPs to be included in the set.

4. After setting the desired options, click the 'Configure' button to run the analysis and return to the main display. Results are shown on a new feature track (see Figure 6, under the track labeled "tSNPs_Tagger_CEU"). As with the LD display above, settings are stored in a browser cookie and the plugin track can be turned off when it is not needed.



Configure... tag SNP Data

Population: CEU

Pairwise Methods: Tagger Pairwise

RSquare cut off: 0.8

MAF cut off: 0.2

Include SNPs: Choose File no file selected

Exclude SNPs: Choose File no file selected

Design scores: Choose File no file selected

Output format: ● text ○ Save to Disk

Figure 5. Options screen for the tag SNP Picker
The tag SNP Picker configuration page allows you to select which population to pick SNPs on, what algorithm to use for the selection, and which SNPs to include or exclude from the resulting tag SNP set. Additional options allow you to fine-tune the tag selection algorithms. The fine-tuning options that are displayed change according to the algorithm selected.

The tag-SNP lists are generated from algorithms in the Tagger program (de Bakker 2005). In the near future we will enhance the tag-picking service in the future by adding additional tag-SNP selection algorithms, and we welcome inquiries from the authors of such algorithm.

Recipe #7: Generate a text listing of tag-SNPs using the genome browser

Using the interactive tag-SNP selection track described in the previous recipe, the researcher can adjust selection criterion until he is satisified with the characteristics of the tag set. This recipe describes how to generate a text dump of the SNPs in this set so that they can be used to create a screening set when combined with information from other HapMap-generated reports.

1. First browse to a region of interest (see Recipe #1).

2. Locate the "Results and Analysis" menu (above the Detail panel) and select the menu item "Dump tag SNP Data". Press "Configure" to set up the interactive options for tag SNP selection as described in Recipe #6. Options include selecting a population and an algorithm, uploading a list of SNPs to necessarily be included as a tag-SNP, uploading a list of SNPs to be excluded from the tag-SNP list, and setting cut-offs for LD and allele frequency values.

3. Click the "Go" button to retrieve the data and produce a report.

The generated report contains a tab-delimited list of tag SNP names, chromosome, position, and allele frequency in the region. This is followed by a section that lists the tag-SNPs, the non tag-SNPs that they capture, and the strength of LD between each tag-SNP and the non tag-SNPs it captures.

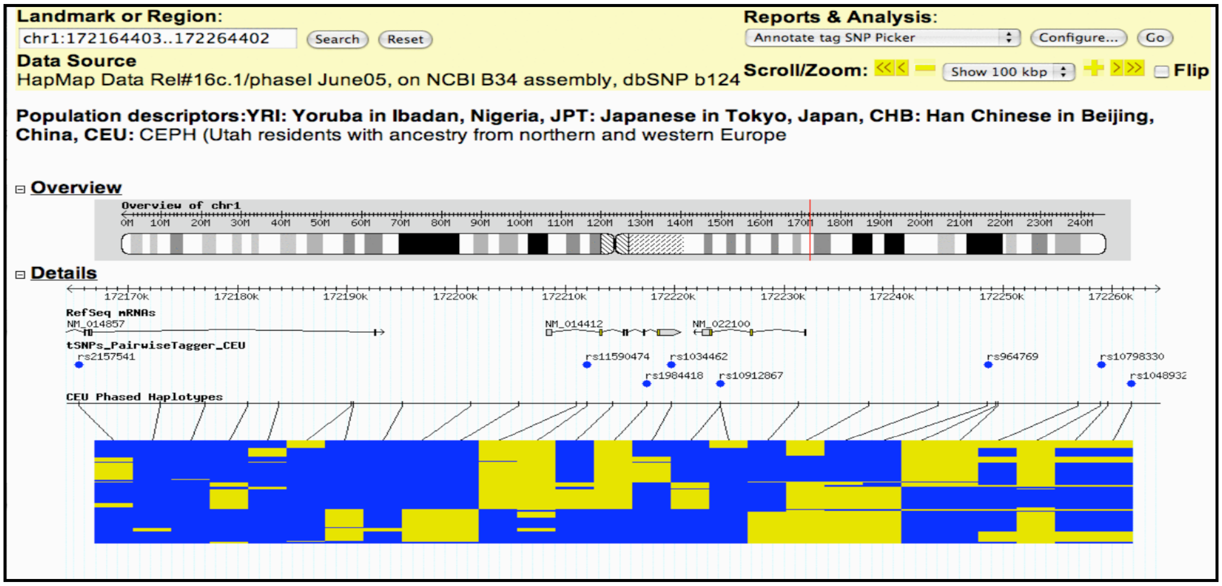
Recipe #8: View phased haplotypes on the genome browser

A researcher may wish to correlate the tag-SNP set selected by the tag-SNP picker algorithm with the underlying haplotype structure of the region. One way to do this is to turn both the tag-SNP and pairwise LD tracks on simultaneously, as described in recipes 4 and 6. An alternative, however, is to activate a track that displays the phased haplotypes themselves.

The phased haplotype data described in this recipe was generated by the International HapMap Project Consortium using the program PHASE version 2.0 (Stephens and Donnelly 2003). During phasing, each allele in a genotype is assigned to one or the other parental chromosome, using a maximum likelihood algorithm that uses trio (lineage) information in the HapMap population groups, or, if trio information is not available, by fitting the data to a model that minimizes the number of implied historical crossovers in the population.

The phased haplotypes are displayed as a graphic in which each chromosome of the individuals sampled by the project is represented as a line one pixel high and each SNP allele is arbitrarily colored blue or yellow. A region of high LD will appear as a region in which there are long runs of SNPs whose alleles are the same color, indicating that there is little recombination among them. A region of low LD will appear as an area where the runs are shorter and more fragmentary.

1. Browse to a region of interest (see Recipe #1).
2. Locate the "Results and Analysis" menu (above the Detail panel) and select the menu item "Annotate Phased Haplotype Display". Press "Configure" to set options for Haplotype display. The options give you the ability to select the population for which to display haplotype information.
3. After selecting the desired population(s), click the 'Configure' button to return to the main display. A new feature track will appear for each population selected. Each track shows the haplotypes for that population using the two color scheme described earlier (Figure 6, track "haplo_CEU"). The order of chromosomes is determined by a fast hierarchical clustering methodology, which places chromosomes that share similar haplotypes together.
4. To retrieve the detailed phased genotypes, click on the track of the desired population. This will take you to a page that provides the haplotype information in tabular



Landmark or Region: chr1:172164403..172264402

Data Source
HapMap Data Rel#16c.1/phaseI June05, on NCBI B34 assembly, dbSNP b124

Reports & Analysis: Annotate tag SNP Picker

Scroll/Zoom: Show 100 kbp

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

Figure 6 Graphical display of tag SNPs and phased haplotypes.
The phased haplotype track represents each phased chromosome in the selected population as a one-pixel high horizontal line. The alleles are represented using a simple two-color scheme, and the chromosomes are sorted in order to maximize the lengths of shared alleles

form. Each row of the table is an individual chromosome, and each column is an individual SNP. The background of each table entry is set to a color corresponding to what is seen in the graphical track.

The advantage of this display over the pairwise LD “triangle display” described earlier is that it is more compact and therefore better suited for the display of large regions. This makes it easy to correlate the position of long common haplotypes with SNPs chosen by the tag-SNP picker. The disadvantage of this display is that it conceals much of the fine-structure of LD in the region, in particular strong linkage among SNPs that are not adjacent to one another.

Recipe #9: Manipulate HapMap data using Haploview

Advanced users who wish to exercise finer control over the display of regions of high LD, or who wish to experiment with new algorithms for tag-SNP picking, may wish to analyze HapMap data using the Haploview program (Barrett *et al.* 2005). This program works well in combination with the HapMap genome browser.

1. To install Haploview, go to http://www.broad.mit.edu/mpg/haploview/ and follow the download link.

2. Haploview requires that the Java Runtime Environment (JRE) is installed on the local computer. If the JRE is not installed, the newest version can be found at http://www.java.com.

3. Download the Haploview program appropriate for your operating system. For Windows computers, download the windows installer file. Double clicking the installer file will create a Haploview folder on the Start Menu. For MacOS X and Unix, download the HaploView.jar file.

4. Download genotypes from a region of interest using recipe #2.

5. To start Haploview, open the Haploview.jar file. On Windows computers, open Haploview.jar from the Haploview folder on the Start Menu. On other operating systems, double click on the Haploview.jar file.

6. To load genotypes click on the Load HapMap Data button, which appears in the Haploview welcome window. Browse to the downloaded file containing genotypes, and open the file.

7. Once the data is loaded, Haploview provides you with options to view a high-resolution “triangle plot” of LD across the region, view shared haplotypes and their recombination frequencies, and select tag-SNP sets by a number of methods. You may select among these analyses and visualizations by choosing the appropriately-labeled tab along the top of the Haploview window (see Figure 7).

8. The “Display” and “Analysis” menus allows you to change the size and coloring scheme of the LD triangle plot, as well as to select among a variety of algorithms for defining “haplotype blocks,” regions of SNPs that are in high mutual LD.

A big advantage of Haploview over the HapMap web site genome browser is that it displays simultaneous high and low-power views of regions of LD, and gives immediate feedback during scrolling and zooming operations. Its current disadvantage is that it does not display gene structures or other genomic features, although this feature is planned as a future enhancement.

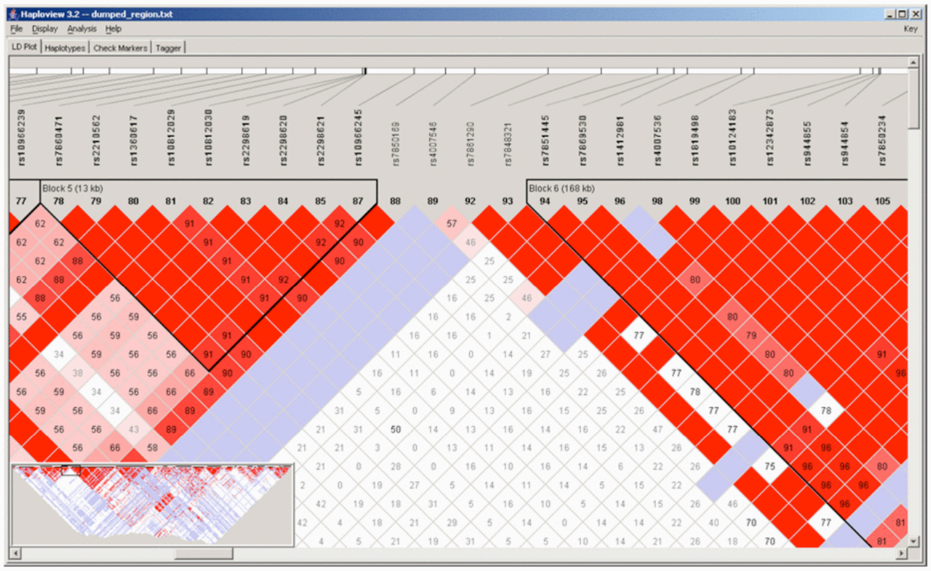


Figure 7 Haploview LD ‘triangle plot’.
This image shows a high-resolution LD ‘triangle plot’ of two regions with high levels of linkage disequilibrium. The color scheme is similar to that used in the LD viewer in the genome browser (see Figures 3 and 4).

Recipe #10: Retrieve HapMap data using HapMart

Because of performance considerations, interactive access to HapMap data via the genome browser is limited to regions no more than 5 Mb wide. Researchers who wish to obtain data for chromosome- or genome-wide data have two choices: bulk download or HapMart access. The former (described in recipe 10) provides text dumps of the entire HapMap data set, and while complete, does not provide any filtering or selection services. HapMart, described in this section, allows researchers to select SNPs using diverse criteria and to display just those aspects of the data set that they are interested in.

1. Open the MartView interface at http://www.hapmap.org/BioMart/martview and click the 'next' button to start a new query using the default database and dataset.

2. On the filter page (see Figure 8), select SNPs to be retrieved on the basis of numerous criteria (either individually or in combination). You can apply filters in any order and revise existing filters by using the "next" and "back" buttons. As you apply filters, the number of SNPs selected are shown in the summary panel on the right. Available filters include:

   a. Lists of SNPs to be included or excluded.

   b. The mininum minor allele frequency of the selected SNPs.

   c. Limit SNPs to those in Intronic Regions, mRNA/UTR regions, coding non-synonymous SNPs, or coding synonymous SNPs.

   d. Limit SNPs to a specific genomic region.

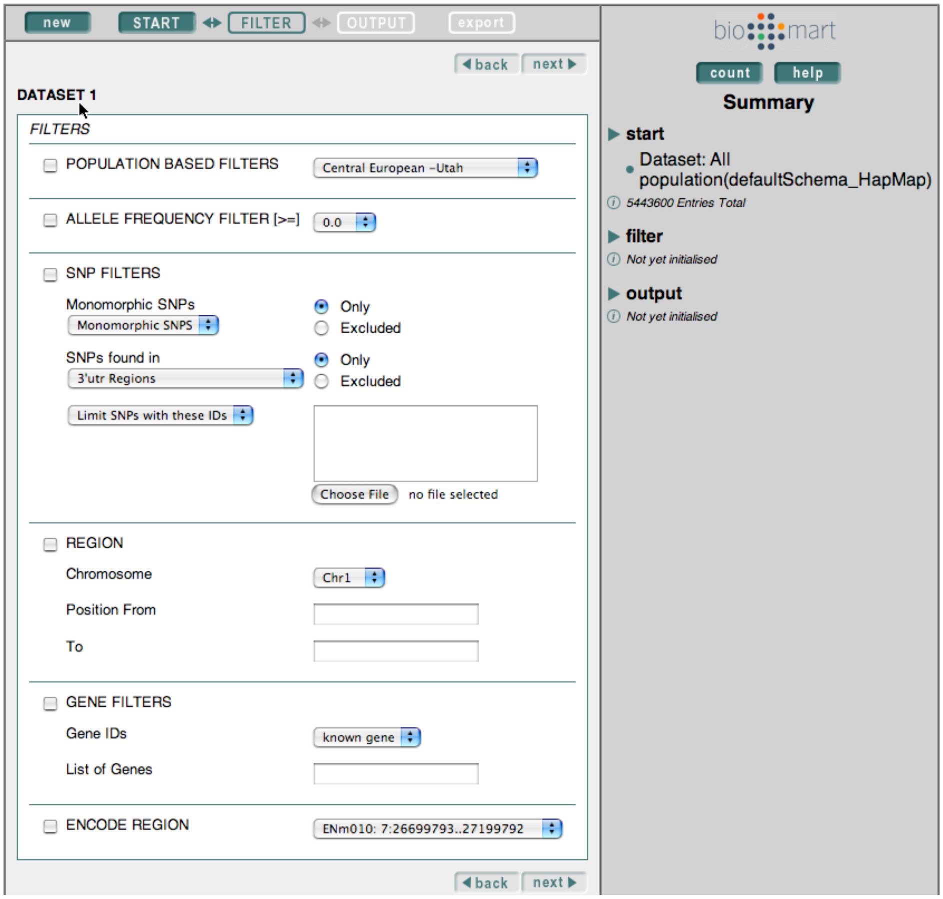
   e. Limit SNPs to those which overlap specific gene ID(s).

3. After selecting and refining the appropriate filters, click the 'next' button to go to the output selection page. This page allows you to select the fields you wish to be output in the report. There are many output options, organized as in a series of tabs along the top of the screen. To retrieve genotypes, for example select the "Genotypes" tab at the top of the page. Then select the "genotypes" checkbox. Other output options include SNP chromosome position, alleles, genotype frequencies and allele frequencies.

5. (optional) If the number of SNPs to be retrieved is large, you may wish to select "gzip file compression." This will compress the report prior to sending it to your browser, and may reduce the time it takes to download the report.

6. To retrieve the results, select the "Export" button.

The reports generated by HapMart are in a tab-delimited text format suitable for importation into Excel or for loading into a relational database. The engine underlying HapMart is a generic data-mining framework named BioMart (Gilbert *et al.* 2003).



DATASET 1

FILTERS

□ POPULATION BASED FILTERS: Central European -Utah

□ ALLELE FREQUENCY FILTER [>=]: 0.0

□ SNP FILTERS

Monomorphic SNPs: Monomorphic SNPS — ● Only ○ Excluded

SNPs found in: 3'utr Regions — ● Only ○ Excluded

Limit SNPs with these IDs

Choose File no file selected

□ REGION

Chromosome: Chr1

Position From

To

□ GENE FILTERS

Gene IDs: known gene

List of Genes

□ ENCODE REGION: ENm010: 7:26699793..27199792

Summary

▶ start

Dataset: All population(defaultSchema_HapMap)

5443600 Entries Total

▶ filter

Not yet initialised

▶ output

Not yet initialised

Figure 8 Filter selection page for the HapMart report generator.
The HapMart report generator allows you to filter the HapMap data set by genomic region or various characteristics of the SNPs and derived genotypes. After adjusting the filters you will be asked to select which fields to include in a tab-delimited report.

Recipe #11: Retrieve data via bulk download

Finally, users can obtain the entire unfiltered HapMap data set by batch download. This recipe describes what downloads are available.

1. To access bulk data, browse to the download page (http://www.hapmap.org/downloads/). Links and descriptions can be found for each category of available data.
2. To download genotypes, click the 'Genotypes' link to go to the genotypes download directory. The 'latest/' subdirectory always points to the current data freeze.
3. The download repository can also be accessed via anonymous FTP at ftp://www.hapmap.org. The two main categories of data are on one hand the full, whole-chromosome dumps found in the 'full/' subdirectory, and on the other hand specialized data dumps such as that found in 'ENCODE/' (data from the 10 ENCODE regions in which all available SNPs are genotyped). Datafiles in both categories are split by chromosome/region and population, and come in three varieties, each within its own subdirectory:
   a. *non-redundant/ Cleaned* These datasets contain only one set of genotypes per SNP/population. All genotype sets have passed quality control checks, and multiple submissions of the same SNP (which occur because of QA exercises, corrected submissions, and planned redundancy in the project) have been removed. This is the dataset most users will want.
   b. *redundant-filtered/ All* These datasets have passed quality control checks, but redundant data has not been removed.
   c. *redundant-unfiltered/ All* This dataset contains all SNPs genotyped by the project irrespective of quality control checks. This constutes the 'raw' data for users who want to look at potentially biologically interesting data that are normally filtered out by project quality control checks.
3. To download LD values, click the “LD Data” link to go to the LD data download directory. The ‘latest/’ directory again points to the most recent datafreeze available. LD values are represented as D’, LOD and r^2 values.
4. To download phased genotypes, click the ‘Phased Data’ link . This will take you to a directory containing data files representing the output of the PHASE program.

Discussion

A number of public online resources have been developed as portals to high-volume genome-wide datasets. 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.* 2004) is a repository for information on single nucleotide polymorphisms, but does not yet contain extensive information on the relationships among those SNPs.

The HapMap web site, located at http://www.hapmap.org, 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 provided 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 both Ensembl information (e.g. “find all genes that contain a zinc-finger domain and a strong homologue in mouse”) with HapMap queries ("find all tag-SNPs for this list of zinc-finger genes").

Acknowledgements

We wish to 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 JC, Fry B, Maller J, and Daly MJ. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21**:263-265.

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

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. (2001). High-resolution haplotype structure in the human genome. *Nat Genet*. **29**:229-232.

de Bakker, P. (2005) Tagger.
http://www.broad.mit.edu/mpg/tagger/.

Gibbs, R.A. J.W. Belmont P. Hardenbol T.D. Willis F. Yu H. Yang L.Y. Ch'ang W. Huang B. Liu Y. Shen et al. 2003. The International HapMap Project. *Nature* **426**: 789-796.

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

International HapMap Consortium 2005. International HapMap Consortium Paper. *Nature.* (in press).

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

Mueller, J.C. (2001). Linkage disequilibrium for different scales and applications. *Brief Bioinform.* **5**:355-364.

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

Stephens, M. and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**: 1162-1169.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33:D39-45.