6 Using the HapMap Web Site

Albert Vernon Smith

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724; Genthor ehf., 101 Reykjavik, Iceland, and Icelandic Heart Association, 201 Kopavogur, Iceland

INTRODUCTION

The primary goal of the International Haplotype Map Project (International HapMap Consortium 2005) has been to develop a haplotype map of the human genome that describes the common patterns of genetic variation, in order to accelerate the search for the genetic causes of human disease. Within the project, approximately 3.9 million distinct single-nucleotide polymorphisms (SNPs) have been genotyped in 270 individuals from four worldwide populations. The project data are available for unrestricted public use at the HapMap Web site, http://www.hapmap.org (Thorisson et al. 2005). This site, which is the primary portal to genotype data produced by the project, offers bulk downloads of the data set, as well as interactive data browsing and analysis tools that are not available elsewhere.

This chapter describes the Web site and tools that have been developed for viewing, retrieving, and analyzing the project data. Details are provided on how to perform several useful and popular tasks. Protocols include instructions for retrieving genotype and frequency data, picking tag-SNPs for use in genetic association analysis, viewing haplotypes graphically, downloading phased genotype data, and examining marker-to-marker linkage disequilibrium (LD) patterns.

Protocol 1

Browsing HapMap Data Using the Genome Browser

Research into the genetic contributions to a human disease commonly focuses on candidate genes identified from linkage and/or association studies, as well as from pathways suspected to be involved in a particular disease process. In studying candidate genes, a researcher will want to know whether there are any common SNPs in the immediate vicinity, what those SNPs' alleles are, and the relative frequencies of the alleles in the population. The researcher will also be particularly interested in coding SNPs, whose alleles change the amino acid sequence of the gene product and therefore might represent functional variations.

METHODS

Finding and Browsing to a Region of Interest

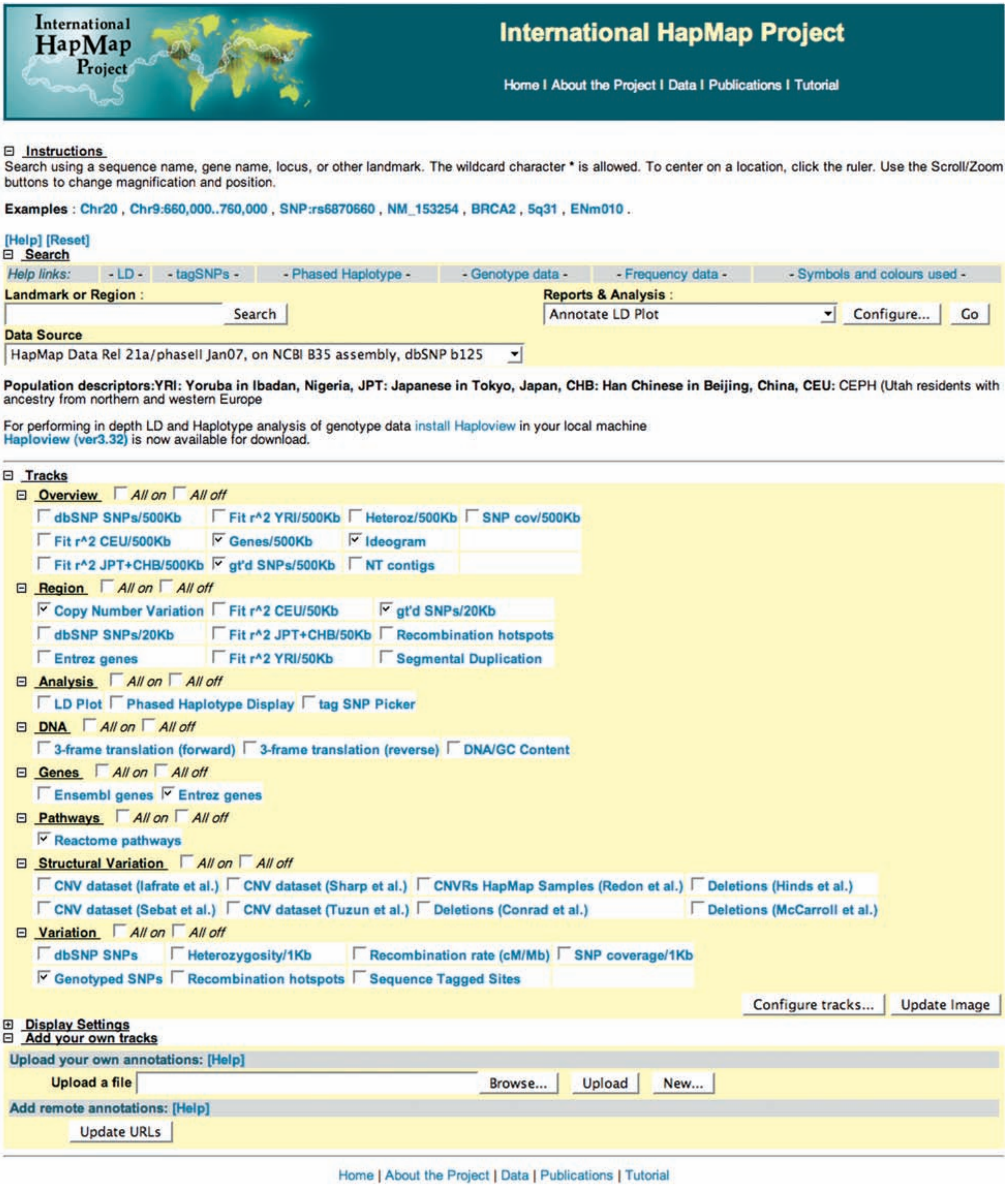
The genome browser at the HapMap Web site provides access to small to medium sized-regions of the genome for this type of interactive exploration. This basic protocol shows how to start using the genome browser.

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 homepage. This will take you to a genome browser based on the GBrowse package (Fig. 6-1).
3. Locate the "Landmark or Region" search box, and enter a search term. Any of the following types of search terms will work:
   - a chromosome name (e.g., "Chr19")
   - a chromosomal position in the format Chromosome:start..stop (e.g., "Chr10: 25000..300000")
   - the name of a SNP using its dbSNP "rs" name (e.g., "rs6870660")
   - a gene using its NCBI RefSeq accession number (e.g., "NM 153254")
   - a gene using its common name (e.g., "BRCA2")
   - a chromosomal band (e.g., "5q31")
4. After entering one of these landmarks, press the "Search" button (or hit "Enter"). This will return a page showing the region surrounding the requested feature (Fig. 6-2). If multiple features match, then the page will show a graphical summary (including genomic location) of all possible features and prompt you to choose one.

   By default, the genome browser goes to the most recent release of HapMap data. Previous releases are available via this interface, and the different releases can be selected under the "Data Source" menu.

   i. 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.

   ii. Below this is a Region overview, displaying 2 Mb surrounding the region of interest. Again, a red box indicates the section of chromosome.


International HapMap Project
International HapMap Project
Home | About the Project | Data | Publications | Tutorial

⊟ 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] [Reset]
⊟ Search

Help links: - LD - - tagSNPs - - Phased Haplotype - - Genotype data - - Frequency data - - Symbols and colours used -
Landmark or Region :
Search
Reports & Analysis :
Annotate LD Plot
Configure...
Go
Data Source

HapMap Data Rel 21a/phaseII Jan07, on NCBI B35 assembly, dbSNP b125

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

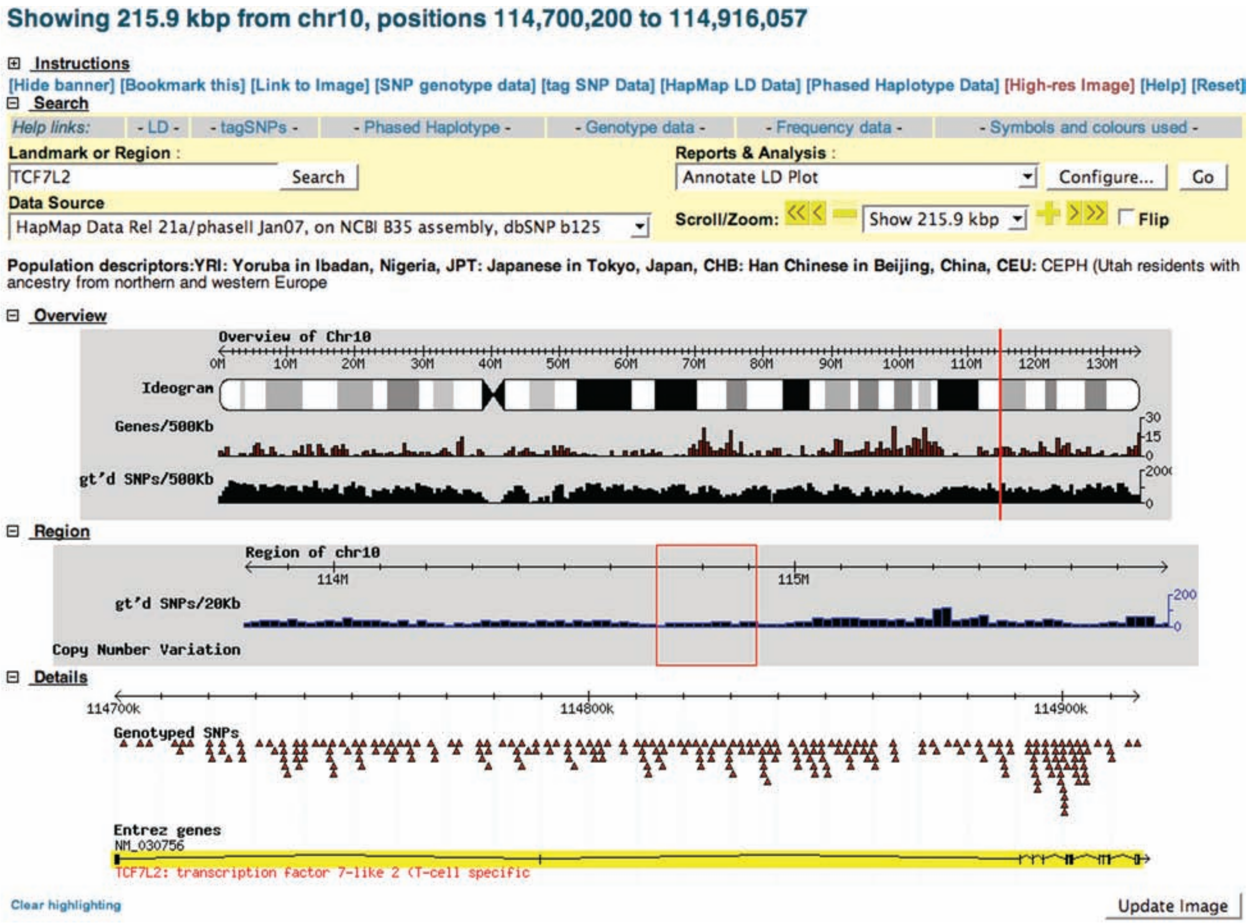
For performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
Haploview (ver3.32) is now available for download.

⊟ Tracks
⊟ Overview ☐ All on ☐ All off
☐ dbSNP SNPs/500Kb ☐ Fit r^2 YRI/500Kb ☐ Heteroz/500Kb ☐ SNP cov/500Kb
☐ Fit r^2 CEU/500Kb ☑ Genes/500Kb ☑ Ideogram
☐ Fit r^2 JPT+CHB/500Kb ☑ gt'd SNPs/500Kb ☐ NT contigs
⊟ Region ☐ All on ☐ All off
☑ Copy Number Variation ☐ Fit r^2 CEU/50Kb ☑ gt'd SNPs/20Kb
☐ dbSNP SNPs/20Kb ☐ Fit r^2 JPT+CHB/50Kb ☐ Recombination hotspots
☐ Entrez genes ☐ Fit r^2 YRI/50Kb ☐ Segmental Duplication
⊟ Analysis ☐ All on ☐ All off
☐ LD Plot ☐ Phased Haplotype Display ☐ tag SNP Picker
⊟ DNA ☐ All on ☐ All off
☐ 3-frame translation (forward) ☐ 3-frame translation (reverse) ☐ DNA/GC Content
⊟ Genes ☐ All on ☐ All off
☐ Ensembl genes ☑ Entrez genes
⊟ Pathways ☐ All on ☐ All off
☑ Reactome pathways
⊟ Structural Variation ☐ All on ☐ All off
☐ CNV dataset (Iafrate et al.) ☐ CNV dataset (Sharp et al.) ☐ CNVRs HapMap Samples (Redon et al.) ☐ Deletions (Hinds et al.)
☐ CNV dataset (Sebat et al.) ☐ CNV dataset (Tuzun et al.) ☐ Deletions (Conrad et al.) ☐ Deletions (McCarroll et al.)
⊟ Variation ☐ All on ☐ All off
☐ dbSNP SNPs ☐ Heterozygosity/1Kb ☐ Recombination rate (cM/Mb) ☐ SNP coverage/1Kb
☑ Genotyped SNPs ☐ Recombination hotspots ☐ Sequence Tagged Sites

⊞ Display Settings
⊟ Add your own tracks
Upload your own annotations: [Help]
Upload a file

Add remote annotations: [Help]

FIGURE 6-1. The initial page shown when starting to use the HapMap genome browser for the first time. Depending on your computer language settings, this page can appear in one of several languages, although this section assumes English. The page can also be reached directly at http://www.hapmap.org/cgi-perl/gbrowse/.



Showing 215.9 kbp from chr10, positions 114,700,200 to 114,916,057

Landmark or Region: TCF7L2

Data Source: HapMap Data Rel 21a/phaseII Jan07, on NCBI B35 assembly, dbSNP b125

Reports & Analysis: Annotate LD Plot

Scroll/Zoom: Show 215.9 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 6-2. The HapMap genome browser displaying a requested feature.

iii. Beneath this is a "Detail" section that has horizontal tracks showing various types of data. By default, only a small number of genomic tracks are displayed initially for the region. 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 Entrez genes track, which shows the positions and structures of human protein-coding genes.

A number of additional information tracks are available, which can particularly help with the understanding and design of association studies. A number of analyses derived from HapMap data, as well as outside data sources, are available (Table 6-1). Particularly noteworthy are a number of tracks related to structural variation in the genome, as well as links to the Reactome database (http://www.reactome.org;Vastrik et al. 2007), a curated resource of core pathways and reactions in human biology.

5. Use the controls at the top of the page to scroll left, right, or to change the magnification of the region. You can also click anywhere on the "Overview," "Region," or the scale at the top of the "Details" section in order to center the view on this position. The genotyped SNP track changes its appearance in a manner appropriate to the scale of the image:

i. At low magnifications, genotyped SNPs appear as equilateral triangles.

These colors can be customized by selecting the "Highlight SNP Properties" item in the "Reports and Analysis" menu.

TABLE 6-1. Available Tracks in the Genome Browser (as of February 2007)

Category	Track
HapMap tools	LD plot
	Phased haplotype display tag-SNP picker
Genes	Ensembl genes (Hubbard et al. 2007) Entrez genes (Wheeler et al. 2007)
Pathways	Reactome pathways (Vastrik et al. 2007)
Structural variation	Copy number variation (CNV) data sets (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005) CNVs typed in HapMap samples (Redon et al. 2006) Deletions (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006)
Variation	dbSNP SNPs (Wheeler et al. 2007) Heterozygosity/1kb Recombination rate (cM/Mb) SNP coverage/1kb genotyped SNPs Recombination hot spots Sequence tagged sites (Wheeler et al. 2007) Fit r^2 in genomic intervals (Smith et al. 2005)

ii. 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.

iii. 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 pie chart indicates the frequency of the allele that appears in the reference genome sequence. The red wedge is the frequency of the alternative allele.

The pie chart display provides the researcher with the ability to easily distinguish SNPs that are highly polymorphic in all four of the HapMap populations and, therefore, more likely to be polymorphic 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.

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

This 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.

i. Click on the hypertext link to dbSNP (http://www.ncbi.nlm.nih.gov/SNP; Wheeler et al. 2007) for more information about how the SNP was first discovered and any other population genetic information that may exist for it outside the HapMap project.

ii. Click on the link to Ensembl (http://www.ensembl.org; Hubbard et al. 2007) to reach a site where the structural impact of the SNP on coding sequence, splice sites, and other features of nearby genes can be examined.

Viewing the Extent of Linkage Disequilibrium

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 LD in the region is essential for reducing the number of SNPs that need to be genotyped across the region. If there is high LD 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 LD 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 populations characterized by the HapMap project has been one of the major goals of this project. 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.

7. To view available LD data precalculated from HapMap genotypes, browse to a region of interest (see Steps 1–4).
8. Select the "Annotate LD plot" plug-in from the "Reports and Analysis" menu.
9. 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 Fig. 6-3).

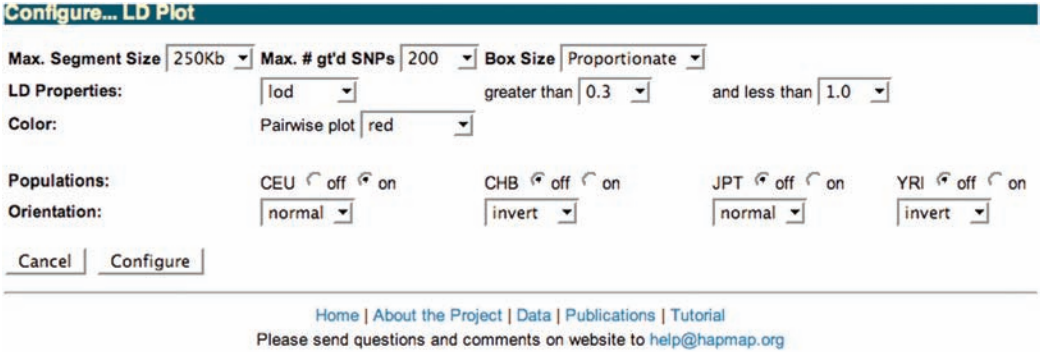
10. Click on the "Configure" button to return to the main display. The display will now show one triangle plot for each population selected (see Fig. 6-4).

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

11. The traditional D′ and r^2 metrics reflect the degree of pair-wise 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 in the HapMap Web site display is described in Daly et al. (2001).

Picking and Viewing tag-SNPs

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 above, 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



Configure... LD Plot

Max. Segment Size 250Kb | Max. # gt'd SNPs 200 | Box Size Proportionate

LD Properties: lod | greater than 0.3 | and less than 1.0

Color: Pairwise plot red

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

FIGURE 6-3. The configuration page of the HapMap genome browser allows the user to customize numerous style features of the data display.

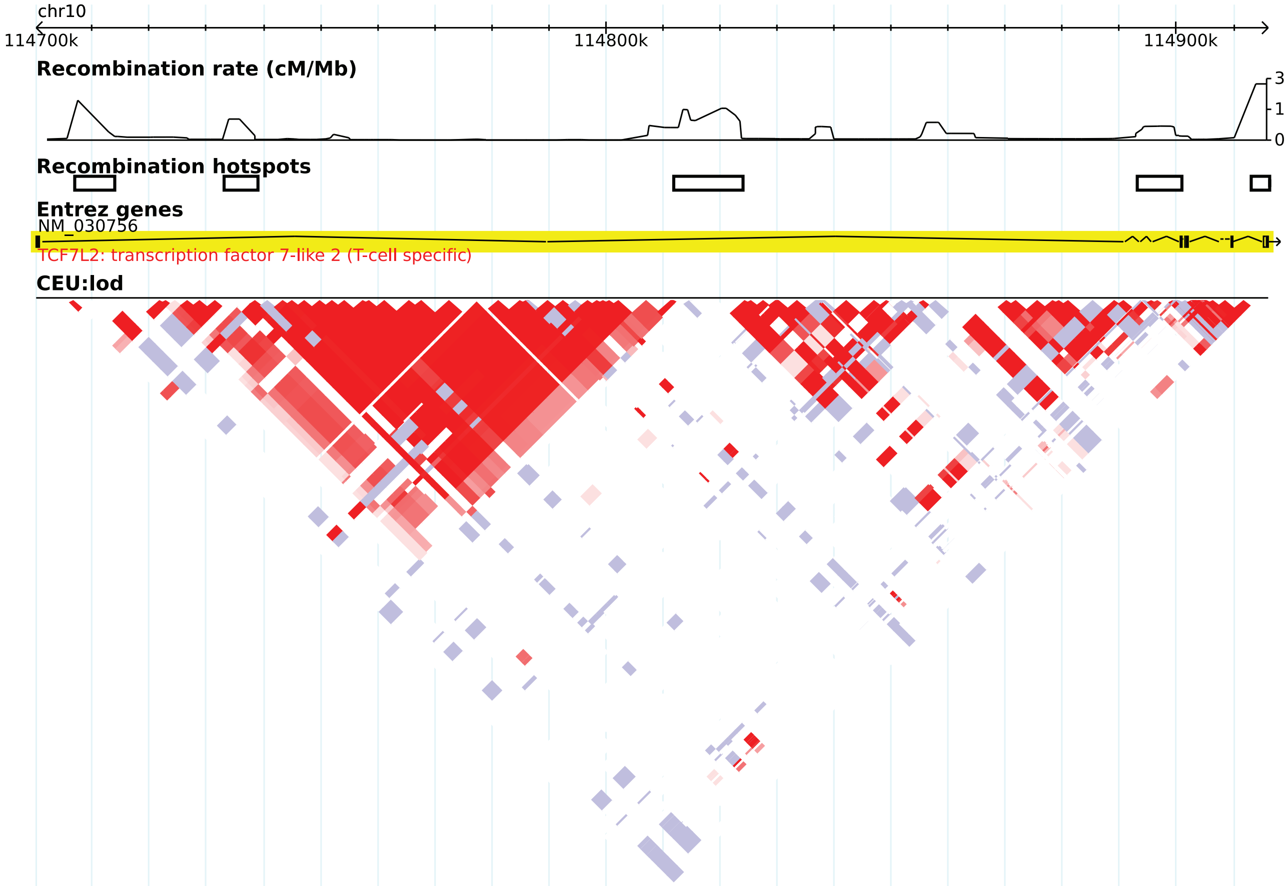


FIGURE 6-4. The HapMap genome browser displaying a triangle plot of LD values for multiple populations. A typical region of LD demonstrating "patches" of high LD separated by relatively well-defined boundaries of low LD is shown. The triangle plot is constructed by connecting every pair of SNPs along lines at 45 degrees to the horizontal track line. The color of the diamond at the position where two SNPs intersect indicates the amount of LD; more intense colors indicate higher LD. A gray diamond indicates that data are missing.

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. The tag-SNP lists are generated from algorithms in the Tagger program (http://www.broad.mit.edu/mpg/tagger/; de Bakker et al. 2005).

12. Navigate to a region of interest (see Steps 1–4).
13. Under the "Reports and Analysis" menu, select the "Annotate tag SNP Picker" option.
14. Press "Configure" to select the desired options for tag-SNP selection (see Fig. 6-5). 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
    - selecting cutoffs for minimum acceptable LD value and allele frequency for SNPs to be included in the set

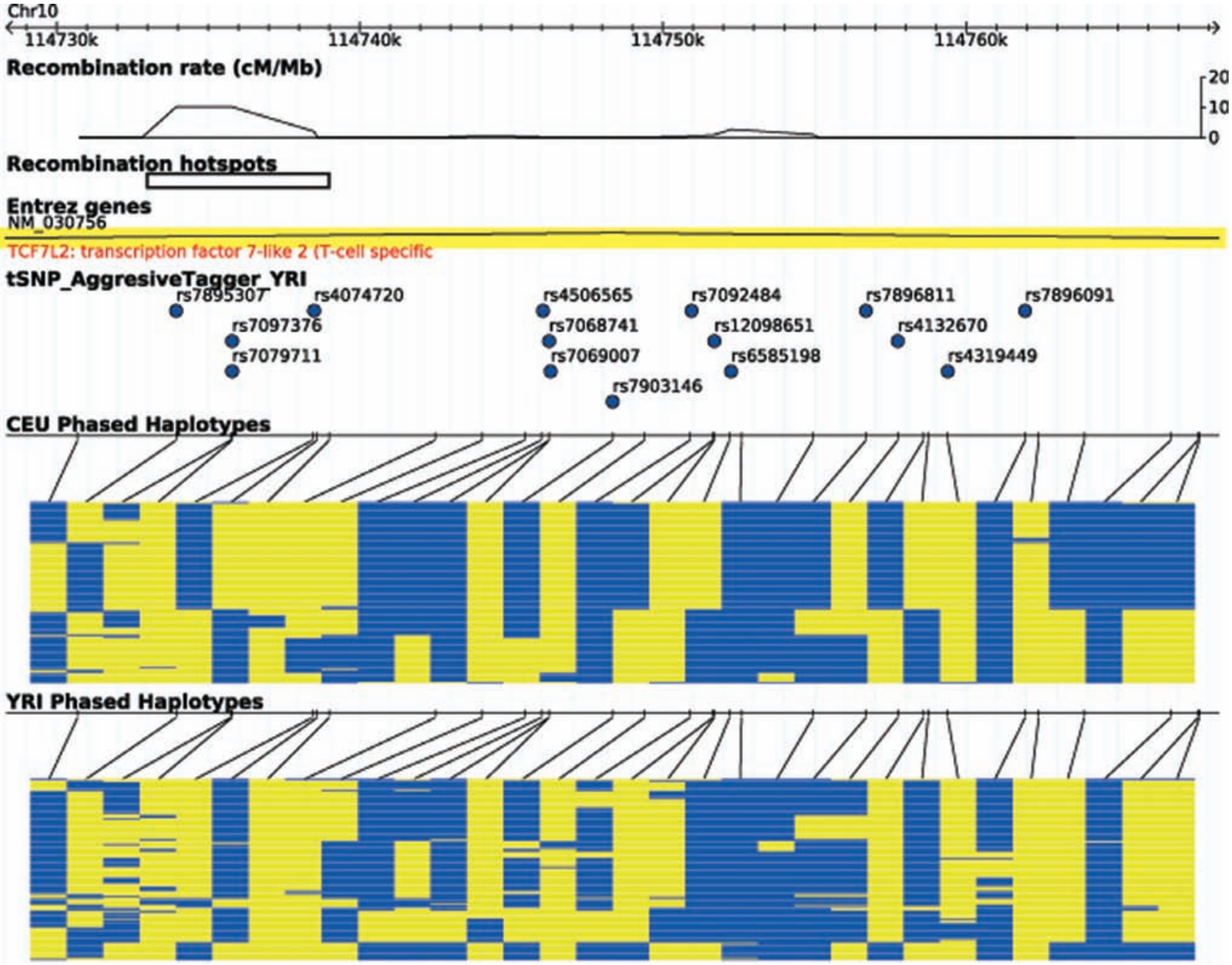


FIGURE 6-5. The HapMap genome browser graphically displaying tag-SNPs, as well as phased haplotypes.

15. Click the "Configure" button to run the analysis and return to the main display. Results are shown on a new feature track (see Fig. 6-5).

    As with the LD display above, settings are stored in a browser cookie, and the plug-in track can be turned off when it is not needed.

Viewing Phased Haplotypes

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 pair-wise LD and tag-SNP tracks on simultaneously (Steps 7–11 and 12–15, respectively). An alternative, however, is to activate a track that displays the phased haplotypes themselves. The phased haplotype data described in this section were generated by the International HapMap Project Consortium using the program PHASE version 2.1 (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 that share alleles across multiple chromosomes, 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.

16. Navigate to a region of interest (see Steps 1–4).
17. Select "Annotate Phased Haplotype Display" from the "Reports and Analysis" menu.
18. Press "Configure" to set options for Haplotype display.

    The options give you the ability to select the population for which to display haplotype information.

19. 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 (Fig. 6-5).

    The order of chromosomes is determined by a fast hierarchical clustering methodology, which places chromosomes that share similar haplotypes together.

    The advantage of this display over the pair-wise LD "triangle display" 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 disequilibrium among SNPs that are not adjacent to one another.

20. 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 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 that seen in the graphical track.