Navigating the HapMap

Michael R. Barnes

Submitted: 20th April 2006; Revised: 14th June 2006

Abstract
With the availability of the HapMap—a resource which describes common patterns of linkage disequilibrium (LD) in four different human population samples, we now have a powerful tool to help dissect the role of genetic variation in the biology of the genome. HapMap is entirely complimentary to the human genome map and so it is particularly fitting that it should be viewed in a full genomic context. However, characterization of high resolution LD across the genome can be a challenging task, owing in part to the sheer volume of data and the inherent dimensionality that its analysis entails. However, a number of tools are now available to make this task easier for researchers. This review will examine tools for viewing and analysing haplotype and LD data, enabling a number of tasks; including identification of optimal sets of haplotype tagging single nucleotide polymorphisms (SNPs); drawing links between associated SNPs and putative causal alleles; or simply viewing LD and haplotypes across a gene or region of interest. The data generated by the HapMap also has other important applications, informing, for example, on the demographic history and evidence of selection in human populations and on previously undetected regulatory relationships and gene networks. All of these properties make the HapMap no less an important resource than the human genome sequence itself and so this makes it essential viewing for all in the field of human biology.

Keywords: HapMap; haplotype; linkage disequilibrium; single nucleotide polymorphism; mutation; human selection; genome

INTRODUCTION
Systematic studies of human variation carried out in the course of the HapMap project have identified and characterized the relationships between four million single nucleotide polymorphisms (SNPs) in four human population samples [1]. This comprehensive sampling of genetic variation may form the basis of many of the phenotypic differences seen in humans and is clearly an issue that any sequence based analysis of the genome needs to take into account. Although most of the polymorphisms identified by HapMap are likely to be neutral in phenotypic effect they still have the potential to inform on nearby alleles that may play a role in diseases. The reason for this is that SNP alleles tend to be correlated together in a predictable way, known as a haplotype—this phenomenon of correlation between SNPs is mediated by linkage disequilibrium (LD) (Table 1). LD and haplotypes are a reflection of the shared ancestry of chromosomes even in outbred populations and are usually related to physical distance (with some exceptions [2]). When a disease allele enters a population by a germline mutation event, it can be subsequently identified by the specific combination of neutral ‘bystander’ alleles that surrounded it on the ‘ancestral chromosome’ where it originally arose [3]. Recombination and mutation through successive generations very slowly erode these affiliations between disease alleles and bystander alleles, generally in a distance-dependent manner until extensive correlation remains between only the closest alleles. This phenomenon underpins the study of human population genetics—in the absence of LD, only a causative allele would show any appreciable difference between case and the control groups in a population-based association study. Even if it were possible to identify all variants beforehand and include them in a study, it would not be efficient to genotype them all. However, in the presence of LD, polymorphisms that are in physical proximity to a causal polymorphism (either individually or
collectively organized as a haplotype) will also show
difference in frequency between cases and controls,
and hence an association will be seen with the trait
in question. How near these polymorphisms need to
be to the disease allele on average is still somewhat
open to debate [4**], but is generally dependent on
the population history of the sample, and the age of
the disease allele. Viewed together, haplotype and
LD information is proving critical to the scale up of
genetics, as selected variants within a haplotype can
effectively represent the variation in an entire
haplotype (these are known as haplotype tags [5*]).
This finally makes scanning of the entire genome a
practical option—which is naturally generating a lot
of excitement in the field of genetics [4**].

The implications of LD extend beyond the
immediate study of genetics and the development
of optimal marker maps. The data contained in the
HapMap is something that the whole field of biology
could benefit from considering. This data shows us
that genetic variation is often correlated, suggesting
that the functional analysis of an individual variant
may not be the best way to determine impact in a
complex system. Instead, it may be worth consider-
ing the analysis of the most commonly observed
combinations of variants in tandem. It is conceivable
that one haplotypic variant may aggravate or com-
penstate for the impact of another variant in the same
haplotype. Constrained pairings of specific variants
are not hard to imagine in the complex, three-
dimensional environments of RNAs and proteins.
Specific amino acid pairings (e.g. hydrogen-
bonded residues in parallel beta-sheets [6]) may be
critical to protein structure and function; while
ribonucleotide pairs may be functionally constrained
in RNA stem loop structures [7]. The HapMap
project is revealing an increasingly complex view of
the relationships between genomic entities based on
their patterns of inheritance. Studies in mice have
even suggested that LD appears to form scale-free
networks across large chromosomal regions or even
between chromosomes [8**]. The full complexity of
the data generated by the HapMap is far from
complete in its exploration, but now at least we
have some excellent material to start to formulate the
right questions to ask. The only outstanding
requirement, which we will try to address here,
is for the flexible tools to find the answers that are
being sought.

### The international HapMap project—background

The HapMap project was established in 2002 to
study the LD relationships across the human genome
in four different ethnic groups [1*]. These included
a panel of 30 trios from the Yoruba, Nigeria (YRI); a
panel of 30 CEPH trios from US Utah residents
with European ancestry (CEU); and a panel of
45 unrelated Japanese individuals from Tokyo (JPT)
and 45 unrelated Han Chinese individuals from
Beijing (CHB). It is worth noting that by most
genetic measures, the Japanese and Chinese popula-
tions are very similar and so in many analyses they are
combined as a single Asian population group (JPT–
CHB). The sample sizes selected for each population
are sufficient for the immediate purpose of the
HapMap—that is to characterize LD and haplotypes
between common variants in these population
samples. However, the sample sizes are not sufficient
to be representative of the specific ‘population’

<table>
<thead>
<tr>
<th>Table I: Glossary of terms</th>
<th>One of the forms of a variant at a given locus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Allele</strong></td>
<td>Presence of different nucleotide alleles at the same locus in populations. Sometimes used loosely to include short insertion/deletion (INDEL) polymorphisms</td>
</tr>
<tr>
<td><strong>SNP (single nucleotide polymorphism)</strong></td>
<td>Alternative form of a genetic variant or locus. A single allele for each locus is inherited from each parent</td>
</tr>
<tr>
<td><strong>LD (linkage disequilibrium)</strong></td>
<td>LD exists when alleles at distinctive loci occur together more frequently than expected given the known allele frequencies and recombination fraction between the loci</td>
</tr>
<tr>
<td><strong>Haplotype</strong></td>
<td>The linear, LD ordered arrangement of alleles on a chromosome</td>
</tr>
<tr>
<td><strong>Haplotype Tag/Tag SNP/tSNP</strong></td>
<td>SNPs that represent a large number of haplotype alleles. A minimal number of tSNPs can represent a maximal number of SNPs in a haplotype or other complex variation</td>
</tr>
<tr>
<td><strong>Haplotype block</strong></td>
<td>Blocks of haplotypes that show limited genetic diversity within specific populations. For example 4–5 common haplotypes may be seen in a given population</td>
</tr>
<tr>
<td><strong>Scale-free network</strong></td>
<td>Networks with scale-free architecture, are often seen in validated biological networks, e.g. protein–protein interactions [60], this suggests that the observed associations are not random</td>
</tr>
</tbody>
</table>
from which they were collected. So the CHB sample is not representative of all Han Chinese, and it is even less representative of wider geographic populations from China. The degree of similarity between the HapMap samples and wider populations is one of the great challenges to the wider applicability of HapMap data. However, initial assessments of the portability of haplotype tagging SNPs between related populations appear relatively promising [9].

The HapMap project has been run in three phases. HapMap phase I was completed in October 2005 and involved genotyping of ~1M SNPs at an average spacing of 5 kb. Phase I of the HapMap also involved additional SNP finding in 48 samples from the original four populations across 10 specific 500 kb ENCODE regions, selected to represent a genome-wide range of evolutionary conservation and gene density [10]. Once SNP finding was completed across these regions, these and all other known variants from the dbSNP database [11] were genotyped in the 269 samples to get as close as possible to saturation coverage. This deep sampling of variation has some immediate benefits as a model of genome wide variation, which are discussed below. The phase II HapMap provided a broader sampling of genomic variation. Using the same 269 samples, a further 2.9 M SNPs were genotyped, bringing the genome-wide total of polymorphic SNPs genotyped up to 3.9 M. In the third and final HapMap phase, samples from several other populations will be genotyped across the HapMap-ENCODE regions in order to evaluate exactly how well the HapMap performs as a model of variation in different populations [12]. This phase is currently expected to include seven other population samples including: African-Americans from the southwestern US; Han-Chinese Americans from Denver, CO; Gujarati Indians from Houston, TX; Luhya from Eldoret, Kenya; Maasai from Webuye, Kenya; Mexicans from Los Angeles, CA; and Tuscans from Sesto, Italy. The HapMap-ENCODE regions are likely to become a standard for assessing genetic diversity in human populations, against which additional communities are also likely to be characterized.

Three years after the launch of the project, genotyping of 4.6 million SNPs is complete and a number of tools are now offering an integrated view of LD across the human genome. A preliminary analysis of the phase I data set has been published [13**], but analysis of the HapMap is still very much ongoing and already it is helping genetics researchers to identify alleles involved in common diseases [14, 15] and drug response [16]. All the information produced by the HapMap project is freely available at the project website (www.hapmap.org).

**HapMap—ENCODE and Phase II**

**HapMap—qualitative and quantitative models of genome variation**

The deep ascertainment afforded by the HapMap—ENCODE regions and the broad ascertainment of the phase II HapMap are respectively important qualitative and quantitative data sets for analysis of variation. The intensity of SNP data across the ENCODE regions (1SNP/279 bp) is unprecedented, being 2–3 times denser than the phase II HapMap (1SNP/kb). This makes the ENCODE regions a ‘near complete’ sample of common variation (>5% minor allele frequency) and a comprehensive source of rarer variants [17**]. The ENCODE regions are some of the most intensely studied, data rich regions in the genome. Arguably on a qualitative level they are better than the entire phase II HapMap for evaluating the likely performance of genome-wide genotyping maps [18**] or the qualitative differences in variation between and within ethnic groups [19, 20]. On a quantitative level, the phase II HapMap offers impressive genome-wide sampling of variation. This can be expected to contain a very substantial proportion of common variation [18**]. In some aspects the phase II HapMap is impressively comprehensive, for example, genotyping was attempted for all known non-synonymous SNPs—a significant proportion of which may confer functional effects—17 092 of which were polymorphic in at least one population [13**].

**Robust measures of LD**

The main statistical measures of LD, used by the HapMap consortium are Dʼ and r². There are a number of assumptions underlying these measures [21, 22**]. In the case of Dʼ, if two SNPs have not been separated by recombination during the history of the sample, then Dʼ = 1. This makes Dʼ a useful measure within studies but not between different populations. The relative magnitude of Dʼ is generally not interpretable and Dʼ estimates can be strongly inflated in small sample sizes. Taking all these factors into consideration Dʼ is of limited use.
for measuring the strength of LD or for the comparison of LD between studies. A more robust measure that has been widely used for quantifying and comparing LD in the context of gene association mapping, is the correlation between alleles at two sites, $r^2$ [23*]. When two alleles are always observed together then $r^2 = 1$, this indicates perfect LD and observation of one marker provides complete information on the other marker. The $r^2$ measure has a number of advantages; it is relatively robust to allele frequency differences between markers; intermediate values of $r^2$ are interpretable and estimates of $r^2$ tend to be much less inflated in smaller sample sizes. In terms of association analysis, $r^2$ is the measure of choice because it measures how well one SNP can act as a surrogate for another in an analysis, hence this measure forms the basis of strategies for selection of ‘tag SNPs’ that capture information on most or all SNP haplotypes in a gene [23*]. Appropriate thresholds of $r^2$ are dependent on the requirements of the analysis. In simulations, Carlson et al. [23*] showed that the lowest statistically robust $r^2$ threshold was 0.5. Measures of $r^2$ below 0.5 although possibly indicative of weak LD should be interpreted with great caution. When high confidence in LD is desired a threshold of >0.8 is more often used. For a detailed overview of the principles underpinning the analysis of LD, see the review by Mueller [22**].

HapMap—some analytical caveats

The HapMap is an incredibly rich data source, however, there are some caveats which need to be considered and can hinder its effective analysis. The HapMap approach is not free from controversy, some researchers are sceptical about the fundamental theorem underpinning the project [24**]. These arguments stem from a struggle between two hypotheses—the common disease, common variant (CD/CV) hypothesis [25**] and the so-called ‘multiple rare variant’ hypothesis [26**]. The former suggests that HapMap LD relationships will effectively tag disease alleles, while the latter suggests that LD relationships between disease variants and population variants may be very different, suggesting that markers in high LD with a causal variant might not actually show association with a disease as expected, even in infinite sample sizes [24**, 27**]. These arguments, though complex, are quite reasonable, but for the optimist the HapMap approach is persuasive [28*]. Taking a pragmatic view of both arguments, it is likely that both are likely to be right some of the time, depending on study design and other factors [29*]. What is certain is that genetics is on a steep learning curve—as studies proceed using HapMap, our understanding of its strengths and limitations will improve accordingly.

The HapMap SNP ascertainment strategy is another issue that has generated some debate. Phase I and II HapMap SNPs were prioritized for selection primarily on the basis of prior validation, failing this they were also considered validated if they matched a variant in chimpanzee sequence data [13**]. This means that the phase I, and to a lesser extent the phase II, data set show some significant ascertainment bias towards ancestral (generally common) alleles [30]. The impact of this is complex and dependent on the specific analysis undertaken, but essentially it means that the power to detect rare variants is reduced, while power to detect common variants is increased [29*].

Another issue worth highlighting is the general concept of the ‘haplotype block’. Early descriptions of genome-wide haplotype structure tended towards descriptions of the genome as a rather fixed arrangement of haplotype blocks [31]. This is problematic as the defined limits of haplotype blocks can be dependent on the density of markers used to determine them [32*], and the underlying recombination rate of the genomic region in which they are located [33]. The concept of the ‘block’ also fails to recognize the extensive LD between markers that may extend beyond a block boundary defined by localized recombination [2*]. Perhaps a more accurate definition of genome-wide haplotype structure is one of the block-like regions of lower recombination showing limited haplotype diversity, punctuated by recombination hotspots [34*]. This is apparent across most regions when LD is viewed (mentioned subsequently).

ACCESSING HAPMAP DATA

Given the potential impact of variation on the regulation and function of genes—tools for visualizing, analyzing and interpreting LD should be considered a core part of the bioinformatics analysis process. HapMap data is now available in an integrated form in a number of tools at varying levels of complexity (Table 2). Also significantly, a number of resources are also beginning to present non-LD analyses based on HapMap data, focusing
for example on recombination hotspots or signatures of natural selection. More than ever, integration of this data across genomic, biological and chemical domains is the challenge where bioinformatics approaches could make great advances which are quite complementary to those in the genetics domain.

The international HapMap project website
The first stop for HapMap data should be the project website (www.hapmap.org) [35]. This website provides simple but powerful access to all HapMap data on three levels: the simplest access is by the HapMap genome browser; for the more advanced user, specific subsets of data can be accessed by HapMart, a DataMart application; and for the database user, there is a bulk data download capability. These different tools all access the same core data sets, both raw and processed. Among the raw data are genotypes, allele frequencies and assay information. Among the processed are LD data sets calculated across the genome in 250 kb windows. Phased haplotypes calculated from trio data in CEU and YRI populations and recombination hotspots—short (~2 kb) regions over which recombination rates rise dramatically. The website itself is also well worth browsing, as it provides pertinent updates and announcements on all issues relating to the project and also contains some very thorough tutorials on most of the common HapMap analysis activities.

HapMap genome browser
The HapMap genome browser is the simplest access point to HapMap data and can be used quite intuitively to view LD and haplotypes around a gene or region of interest, to select tagging SNPs, or to export genotypes or LD data in single or multiple populations. The browser, which can be reached by following the ‘browse project data’ link on the HapMap homepage, can be searched using a gene, genomic region or SNP ID. The browser is fully configurable to display LD, haplotypes, recombination hotspots and tag SNPs alongside genes across the selected region.

To view different features, select the ‘reports & analysis’ menu and select the desired feature. For example, to view haplotypes, select ‘annotate phased haplotype display’ and click the ‘configure’ button. Figure 1 shows an example of output from the tool. HapMap LD data can be viewed alongside a variety of other features including known genes, recombination hotspots and phased haplotypes. SNP frequency of HapMap genotyped SNPs is also shown as a population-specific SNP frequency pie chart. It is also possible to upload custom annotations to view your own data in the context of the data presented in the tool. There are some limitations to the viewer.

Table 2: Tools for LD and haplotype analysis

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD and haplotype analysis</td>
<td><a href="http://www.hapmap.org">www.hapmap.org</a></td>
<td>HapMap project website including genome browser, HapMart and all data for bulk download</td>
</tr>
<tr>
<td>HapMap genome browser</td>
<td>genome.ucsc.edu</td>
<td>Genome browser incorporating data on LD, recombination, positive selection and detailed ENCODE characterization</td>
</tr>
<tr>
<td>UCSC genome browser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other applications of HapMap data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensembl</td>
<td><a href="http://www.ensembl.org">www.ensembl.org</a></td>
<td>Genome browser incorporating LD (in SNPview)</td>
</tr>
<tr>
<td>Haplotter</td>
<td>hg-wen.uchicago.edu/selection/</td>
<td>Web tool for querying selection data published by Voight et al. [45]</td>
</tr>
<tr>
<td>Sweep</td>
<td><a href="http://www.broad.mit.edu/mpg/sweep/">www.broad.mit.edu/mpg/sweep/</a></td>
<td>Java applet for visualisation of haplotype selection data</td>
</tr>
</tbody>
</table>
Figure 1: The HapMap genome browser. HapMap genome browser view across the 3' portion of the CARD4 gene. Tracks display genotyped SNPs with allele frequencies represented by a pie in each population, known genes, recombination hotspot information, HapMap $r^2$ LD information, tag SNPs selected by the tagger program and haplotypes computed by the PHASE program.
The detailed view of genomic features is limited to 5 Mb, while views of LD are limited to 1 Mb.

Aside from visualization functions, the HapMap genome browser also serves as an effective data mining and bioinformatics analysis tool. The tool allows the user to export data on LD, SNP frequency and genotypes from the current window being visualized. The browser is closely integrated with the HaploView software package [36], allowing the user to pipeline genotype data from a region directly into the HaploView package, which gives the user more flexibility to recalculate LD and haplotypes and select tag SNPs. In a similar way, the user can also pipeline genotypes into Tagger [18**] to allow selection of tag SNPs, which capture haplotype diversity across a region in a maximally efficient manner. This pipelining capability makes the HapMap browser a remarkably flexible web tool, which effectively provides the web user with a view of LD and haplotypes that would otherwise require hours of data loading, manipulation and database expertise.

HapMart

While the HapMap genome browser is useful for exporting LD data across a defined region or gene, the HapMart tool is the only practical tool for exporting LD data from specific populations across a range of chromosomes or genes based on different filter criteria. These filters can include lists of query SNPs, a minor allele frequency cutoff, a region or by gene (Figure 2). Powerful queries may be performed with these filters. For example, filtering by SNP region, it is possible to retrieve all HapMap SNPs that are in LD with non-synonymous SNPs (nsSNPs). Alternatively, one might use the SNP filter function to highlight other SNPs that are in LD with a specific list of SNPs of interest (e.g. SNPs associated in a study). This enables the user to identify all SNPs in LD with an associated SNP to allow functional analysis to identify putative causal SNPs for further analysis in the lab [37*]. The output of such queries can be easily formatted into a custom genome browser track for display in tools such as the University of California Santa Cruz (UCSC) human genome browser (see subsequent text and Figure 3 for an example). HapMart is also a useful tool for retrieving raw data for further analysis, e.g. genotypes for a submitted list of genes [based on human genome organisation (HUGO) gene IDs]. These are powerful queries as they allow the user to extract specific data on genes of interest from the wider HapMap data set, which is otherwise quite intractable outside of a database environment.

Downloading HapMap data in bulk

The final HapMap data access option that is open to the user is bulk data download. Although it is known that LD can occasionally extend across mega bases—this is not easy to calculate due to the high dimensionality of comparing thousands of data points against each other. Pre-computed sets of HapMap LD calculated within a sliding 250 kb window are available within the bulk data download area. However, pre-computed data sets will not suit all use cases, so in some cases raw genotype data will need to be processed in a database environment, using PERL and SQL query tools in combination with LD analysis algorithms [38, 39]. All genotypes (pre- and post-QC), SNP frequencies, haplotypes, recombination rates and even the genome browser code are available for download. Downloading all this data without a very good reason might be foolhardy as the pre-computed LD alone amounts to more than 24 gigabytes of data. It is probably fair to say that the combined flexibility of the HapMap genome browser, HapMart and other tools like HaploView can carry out all but the most demanding analysis objectives, so barring whole genome analysis it is probably worth investigating these options before hitting the download button.

PLACING LD INTO A GENOMIC CONTEXT

While the tools on the HapMap website provide state-of-the-art access to HapMap data and analysis, they only offer limited information on the wider genomic context of a region. To fully understand the biological and functional ramifications of human variation, it is important to put LD and haplotype information in a full genomic context. The UCSC human genome browser [40] and Ensembl [41] are two key tools that are available to achieve this and both have now integrated HapMap LD data to visualize along with other genomic information.

Although both tools have many similarities, each contains distinct information and data interpretation and so it usually pays to consult both viewers, if only for a second opinion (both viewers provide reciprocal links). The UCSC genome browser has one great advantage over both the HapMap genome browser and Ensembl as it allows visualization of LD across
regions of greater than 1 Mb or even whole chromosomes. This robust LD visualization really makes the UCSC browser an exceptional tool for integrated LD/genomic visualization [42]. Figure 3 shows a 1.5 Mb region containing the lactase gene (LCT). This shows clear differences in LD between the CEU, the YRI and the JPT–CHB populations, it also displays recombination rates calculated from HapMap data (which correlate well with LD block boundaries) and evidence of positive selection in different ethnic groups based on calculation of Tajima’s D from SNP genotype data (see subsequent text for details). The LD and haplotype information is also placed in context with known genes and vertebrate genome conservation. Descriptive information for each UCSC data set can be reached by pressing the grey button to the left of each track. A great deal of configurable extra information is also available but not shown here for brevity.

The UCSC browser is also very effective for detailed analysis of the genomic context of LD data. A causal SNP is unlikely to be tested directly in a genome scan, but it may be in LD with markers that are tested. Using the output of a HapMart query, it is relatively simple (e.g. using Microsoft Excel) to create a UCSC custom data track (see UCSC documentation for details) based on SNPs that show evidence of LD ($r^2 > 0.5$) with an associated SNP. Figure 4, shows an example of such an analysis. By loading the associated SNP location and the locations of the SNPs showing evidence of LD as custom tracks, alignment with genomic features is
much easier than using the LD block view. Genomic features and SNPs can be directly aligned to assess overlap. The entire DNA sequence can also be exported with different tracks annotated on the sequence by using the ‘DNA’ link at the top of the browser. Beyond visual inspection, custom tracks created for the UCSC browser also have another incredibly powerful application—they can be queried using the UCSC table browser [43**]. Table browser, which is accessed by the ‘Tables’ link in the main browser is an excellent tool that effectively allows the user to perform complex queries between data sets, including custom tracks loaded by the user. Using the intersection filter,

Figure 3: The UCSC genome browser. The Lactase (LCT) gene locus is visualized. Tracks describe known genes, conservation across a range of vertebrate genomes, SNPs, LD in CEU, YRI and JPT–CHB populations. Tajima’s D values present evidence of selection in populations of African, European and Asian descent. Negative Tajima’s D values seen in Caucasians represent strong evidence of positive selection across the Lactase locus. The bottom two tracks show recombination hotspots independently identified in the HapMap and Perlegen data sets. These are clearly seen to disrupt haplotype blocks across the locus in all three populations.
Figure 4: Using custom tracks to get a detailed view of genomic context in the UCSC genome browser. Genomic location of an associated SNP close to the CARD4 gene. Custom tracks in the UCSC genome browser show the location of SNPs showing LD with an $r^2 > 0.5$ threshold in the context of genome annotation. Additional tracks describe known genes, human spliced expressed sequence tags (ESTs), unspliced ESTs, predicted regulatory potential and conserved transcription factor binding sites. HapMap LD information below is for the CEU individuals and suggests that there are two conserved haplotype clusters in this region. These clusters are divided by a recombination hotspot independently identified in the HapMap and Perlegen data sets.
for example, it is possible to identify all SNPs previously identified by LD (your custom track) that overlap with conserved transcription factor binding sites or known promoter elements or micro RNA target sites, etc. This highly flexible functionality makes the UCSC Table browser a singularly powerful tool which is highly recommended to the reader.

**HAPMAP—FROM GENOTYPE TO GENOME FUNCTION**

Although initiated as a fundamental resource to support genome-wide association studies, the HapMap data set is in itself an informative genome scan. For example, it can illuminate the history of human populations [44] and the past and present selection pressures that operate on these populations [45**]. It can inform on the physical nature of the genome, by highlighting structural variation [46] and recombination hotspots and coldspots [4]. This data can in turn be used to differentiate functional constraint and low mutation rates, a key step in the determination of possible function [47]. HapMap data can also inform on functional relationships between genes, variants and regulatory elements, by highlighting selectively constrained relationships between variants, e.g. between a group of genes or a gene and cis-regulatory elements [48]. The sections below highlight just a few of the key advances that the HapMap is facilitating. Other valuable applications no doubt exist but remain to be elucidated.

**HapMap insights into selective pressure in human populations**

When a new mutation arises, which confers a selective advantage, it is likely to increase in frequency in the population by natural selection [49]. This event will also influence the standing variation neighbouring the mutation, as the pattern of variation in the individual in which the mutation arose sweeps away other variation in the selected locus in the population. This leads to a reduction in haplotype diversity, increased LD and a skewed pattern of mostly low allele frequency variants in the selected locus. This chain of events is known as a ‘selective sweep’ [49]. Already a number of signals of very strong and recent population specific selection have been identified in human genes, for example, in response to malaria [50]. Researchers have used SNP genotype data to detect these signatures across the genome, using a variety of statistical measures. The results are fascinating and may highlight links between historic selection and disease susceptibility in contemporary human populations [45**, 51, 52]. Some of the outputs of these analyses are readily available on the web. Voight et al. [45**] scanned phase I HapMap SNP data in the CEU, the YRI and the JPT–CHB populations using a novel haplotype-based test, iHS (integrated haplotype score), and found recent evidence of positive selection in all three population samples. They identified at least nominally significant evidence of positive selection in at least one population in 2532 genes. The results of these analyses are available to query any phase I HapMap SNP, gene or region for evidence of selection in a stand-alone application—Haplotter (http://hg-wen.uchicago.edu/selection/). In an earlier study, Carlson et al. [52] used another high-density SNP data set (1.5 M SNPs described by Hinds et al. [53]) to carry out a similar analysis using the Tajima’s D test statistic. This analysis is available in the ‘Tajima’s D’ track in the UCSC genome browser. Figure 3 shows an example of the Tajima’s D data generated by Carlson et al. across the LCT gene locus. This gene shows evidence of strong positive selection in Caucasian populations (represented by strongly negative Tajima’s D values), while no strong evidence of positive selection is evident in African and Chinese populations. This is believed to be due to the selective advantage that alleles conferring lactose tolerance afforded Caucasian dairy farming populations [54*]. The value of both of these data sets should not be underestimated, as both represent compelling evidence of population specific shifts in function across genes and their regulatory regions. Narrowing down a selective sweep signal to the putative allele undergoing selection is obviously non-trivial as the allele undergoing selection may not be in the HapMap data set. However tools such as the HapMap genome browser help this search, for example, by allowing the user to visually review allele frequencies in all populations across a region showing selection [e.g. by using the population-specific SNP frequency pie charts (Figure 1)]. Other tools such as HapMart can be used to export allele frequency data in bulk to evaluate population-specific differences. Like any other large-scale data analysis, approaches using HapMap data may also miss many loci undergoing selection, and falsely identify others depending on the mode of positive
selection and the demographic history of the population (see Teshima et al. [55*] for a review of these issues).

**Using HapMap to evaluate cis- and trans-regulatory relationships in the genome**

The full complexity of relationships between genes and their regulatory regions are still very poorly understood. The conventional view expects that cis-regulatory elements are located in the immediate 5' region within 500 bp–2 kb of the transcriptional start site of the gene undergoing regulation [56]. However, some studies are beginning to demonstrate cis-regulation of genes over great genomic distances (over 1 Mb) [57*, 58*]. Alternatively, evidence is emerging of trans-regulation of genes from different chromosomes, probably indirectly by downstream regulatory proteins [57*, 58*]. Analysis of HapMap data can give clues to these relationships by identifying alleles that are in strong LD, sometimes across great distances or even between different chromosomes. Several groups have carried out HapMap-based association studies to identify regions in which gene cis- and trans-regulatory expression variation segregates with genotype [57*, 58*]. This hypothesis-free approach for the genome-wide analysis of sequence elements that influence gene regulation is extremely promising, and should bring new insight into the understanding of gene regulation. However, some problems still need to be overcome, such as the potential for false-positive associations due to multiple testing [59], and the limited sample sizes currently available for these studies [48].

**The Future?—What can LD tell us about gene and genome function?**

It is clear that the data contained in the HapMap has potential to inform on the fundamental organization of genes and their functional networks. However, our current ability to analyse HapMap data is limited by the sheer volume of the data. For example, the HapMap LD analysis was calculated across 250 kb windows. There is evidence of LD over much greater distances [2*]; however, we have not yet tested the limits of LD on a genome-wide scale in humans. Studies of less genetically complex organisms such as inbred mouse strains highlight possible future directions for human analysis. In a study of 60 inbred mouse strains, Petkov et al. [8**] found that intense selection during inbreeding favoured the coinheritance of optimal sets of alleles, resulting in extensive LD, both across and between chromosomes. Recombination that disrupted the preferred combinations of alleles reduced the ability of the offspring to survive further inbreeding. Analysis of this data suggested that a quarter or more of the mammalian genome consists of chromosome regions containing LD clusters of functionally related genes. The fact that, LD was also seen between markers on separate chromosomes, forming networks with scale-free architecture, which is often seen in validated biological networks [60], suggests that the observed associations are not random. The networks they observed were large, with nodes corresponding to genes and connections corresponding to metabolic or physical interactions. Combining LD data with Gene Ontologies [61] and genome annotation, they identified shared biological functions underlying several domains and networks. The potential dimensionality of calculating LD between chromosomes in humans, severely limits evaluation of this phenomenon in humans; however, given the strong conservation of gene order among mammals, Petkov et al. [8**] proposed that these domains and networks are unlikely to be unique to rodents and probably characterize all mammals, including humans.

It may seem a little far-fetched to extrapolate from findings in inbred mouse strains to diverse human populations. But perhaps what this study does illustrate is that HapMap genotype data, just like other genomic data, such as transcriptome data, can help to identify novel, high-level, organizational networks between genes. This may not have been the prime objective of the HapMap project, but it is a layer of complexity that needs to be resolved if genome function, phenotype and disease are to become continuous in our understanding.

**Key Points**

- The HapMap is a powerful tool enabling the dissection of the role of genetic variation in human disease.
- Characterization of LD and haplotype data in a full genomic context is being enabled by a number of public domain applications described in this article.
- HapMap has applications beyond the immediate field of human genetics, showing potential to illuminate diverse areas of genome function, including recombination, population-specific selection pressure, gene regulation and function.
- Effective use of HapMap data by bioinformatics researchers may be key to the understanding of the function of the human genome.
**References**

Papers of particular interest, published within the annual period of review, have been highlighted as:
* of special interest
** of outstanding interest

   *The first major publication from the HapMap group laying out the foundation of the project and describing its aims and objectives.*

   *An interesting paper which reports an unexpected observation that many SNPs are in complete LD with other SNPs hundreds of kilobases away, separated by SNPs that show limited or no LD. This illustrates the problem with the 'LD block' concept as the definitive definition of LD.*


   *An excellent review of issues that are faced by the genetics research community in running well-designed, well-powered genome scans. The appendix contains a particularly valuable review of the state of the art in this area.*

   *One of the first publications describing the use of a minimal number of SNPs (haplotype tags) to reflect the full diversity of SNP haplotypes.*


   *A fascinating analysis of LD in inbred mouse strains which may give an insight into scale-free LD networks in humans.*


   *A major publication from the HapMap consortium describing analysis of the phase I Hapmap data.*


   *See annotation to [18**].

   *These papers demonstrate the value of HapMap data in evaluating the genome-wide coverage and likely performance of fixed marker genotyping panels in disease gene hunting.*


   *A very good review of the measures of LD and their use cases.*

   *A widely used method for selection of informative sets of markers.*

   *An eloquent, but somewhat outspoken critique of the HapMap approach to disease gene hunting.*

   *One of the early papers proposing a common disease: common variant (CDCV) hypothesis.*

   *An interesting paper proposing a ‘multiple rare variant’ hypothesis in complex disease.*

   *Another good contribution to the debate, this time against the CDCV hypothesis.*

   *An upbeat appraisal of the potential of the HapMap as a resource for genetic association studies.*

A good review of some of the issues of study design based on HapMap data.


