A key goal of biomedical science is to understand why individuals differ in their susceptibility to disease. Family history is among the established risk factors for most forms of cardiovascular disease, in part because inherited DNA sequence variants play a causal role in disease susceptibility. Consequently, the search for these variants has intensified over the past decade. One class of DNA sequence variants takes the form of single nucleotide changes (single nucleotide polymorphisms, or SNPs), usually with two variants or alleles for each SNP. SNPs are scattered throughout the 23 pairs of chromosomes of the human genome, and roughly 11 million common polymorphisms (i.e., those >1% frequency) are estimated to exist. A combination of SNP alleles along a chromosome is termed a haplotype.

The International Haplotype Map Project was designed to create a public genome-wide database of common SNPs and, consequently, enable systematic studies of most common SNPs for their potential role in human disease. We review the following: (1) the concept of linkage disequilibrium or allelic association, (2) the HapMap project, and (3) several examples of the utility of HapMap data in genetic mapping for cardiovascular disease phenotypes.

**Linkage Disequilibrium: Correlation Among SNPs**

Groups of SNPs across the genome are correlated with each other, a phenomenon known as linkage disequilibrium (LD) or allelic association. To understand how LD arises, one needs to recall that during meiosis, recombination occurs at multiple sites between each pair of chromosomes, thus providing for an extra source of genetic variability to pass on to offspring. This is not a random process that occurs with equal probability at every place along a chromosome; rather, there are large stretches of DNA along which there is a very low probability of recombination, punctuated by recombinaton “hotspots,” where it occurs relatively more often. The consequence is that the stretches of DNA between hotspots tend to stay together—in what are referred to as “haplotype blocks”—as they are passed along from generation to generation.

To understand how SNPs arise and become correlated with other SNPs, consider the following hypothetical example (Figure 1). At some time in the remote past, a mutation in a single individual results in a base change from “A” (adenine) to “G” (guanine). Previously, there was no variation at that site in the population, with everybody else having an “A” allele at the position in both copies of the gene (one copy on each of the paired chromosomes). There is an SNP nearby that is a “C” (cytosine) allele 50% of the time and a “T” (thymine) allele the other 50%. It so happens that the A→G mutation arose on a chromosome in which the identity of the nearby SNP is a “C” allele. If the mutation is not so harmful that natural selection would cull it out of the population, it is transmitted to many successive generations; in this example, it spreads through the population until 10% of chromosomes in the population have a “G” allele at the position.

Because the new A/G SNP and the old C/T SNP are close together with no recombination hotspots between them, resulting in essentially no recombination between them in successive generations, all chromosomes with a “G” allele at the first SNP also have a “C” allele at the second SNP. In contrast, chromosomes with an “A” allele at the first SNP have some chance of having a “C” allele at the second SNP, with the others having a “T” allele, reflecting the state of affairs before the origin of the new SNP. The C/T SNP has become correlated with the A/G SNP, and knowledge of the allele at one of the SNPs confers some information about the allele at the other SNP.

SNPs within a haplotype block and, to a lesser extent, SNPs in nearby haplotype blocks tend to remain correlated over time. The degree of correlation or LD can be quantified in two different ways, the calculated values of D’ and r². D’ measures the deviation of haplotype frequencies from linkage equilibrium and r² is a measurement of correlation between a pair of variables. r² is particularly useful in genetic mapping—when r² = 1 (the maximum value), knowing the genotypes of alleles of one SNP is perfectly predictive of the genotypes of another SNP. (Please see Wang et al11 for an expanded discussion of these concepts and the mathematical formulations.) Although any haplotype made up of n SNPs (each with two possible alleles) potentially has 2^n combinations of SNP alleles, far fewer combinations are actually seen in a population because of correlation among the SNPs. In principle, knowledge of the correlation structure among all SNPs in the genome—as represented by a vast array of
HapMap has allowed efficient design of genetic association studies. A comprehensive test of common SNPs would theoretically involve the genotyping of all 11 million common SNPs in patients with disease and individuals free of disease. However, the correlation structure among SNPs provided by HapMap allows investigators to genotype far fewer SNPs while still retaining statistical power to find regions of the genome associated with disease. Because a given SNP may be in LD with another SNP in the same region, knowledge of the genotype of the first SNP of the pair may be sufficient to infer the genotype of the other SNP, thereby acting as a “tagging” SNP for the other SNP. In this way, a single SNP can potentially serve to “tag” a number of other SNPs. A judiciously chosen panel of approximately 300,000 to 500,000 HapMap SNPs is sufficient to capture the information content of the full 3 million SNPs in HapMap individuals of European or Asian descent, whereas a panel of approximately 1.1 million SNPs is required in Yoruban individuals. Furthermore, panels of tagging SNPs chosen for each HapMap ethnicity have been shown to provide similar power for non-HapMap study populations of the same ethnicity. Greater than 60% coverage of the genome is provided by commercially available SNP “arrays” or “chips” that can interrogate several hundred thousand SNPs in a single experiment; successive generations of these chips that interrogate upward of a million SNPs will provide even better coverage, resulting in increased statistical power to find disease associations.

Multimarker Tests and Imputation

Increased statistical power can also be achieved by using multimarker tests, in which haplotypes of correlated SNPs are used to tag other SNPs. This is possible because the HapMap database reveals which haplotypes are found in populations. For example, for a set of 3 SNPs for which each SNP has 2 possible alleles, there are 8 possible haplotype combinations, but only a few haplotypes may be seen in HapMap. Thus knowledge of the identity of the first SNP alone may not be sufficient to infer the identity of the third SNP, but the combination of the first and second SNPs may predict the third SNP (Figure 2). When used for tagging in this fashion, 2-marker SNP sets have been shown to significantly improve genome coverage by SNP chips—in the case of the Affymetrix 500K Mapping Array Set, from 66% to 78%. 

The HapMap Project

The International HapMap Project began in October 2002 with the purpose of identifying millions of SNPs throughout the genome, determining the allele frequencies at each SNP, and determining the correlations between SNPs. Drawing on 269 DNA samples from individuals of four different ethnicities—90 residents of Utah in the United States with Northern and Western European ancestry, 90 Yoruba people in Nigeria, 44 Japanese people in Tokyo, and 45 Han Chinese people in Beijing—HapMap has now genotyped more than 3 million SNPs in each of these populations and published the results in a public database.

Analyses of this data have yielded a number of important insights into human genetic variation. For example, although the 4 ethnic groups included in the HapMap Project share most SNPs, the allele frequencies at these SNPs can vary widely among the groups. Yoruban individuals appear to have many more rare alleles (frequency <5%) than the other groups, which may reflect the fact that European and Asian populations are “younger” (ie, descended from offshoots of an ancestral African population). Recombination hotspots are widely distributed across the genome, they are more common near telomeres (the ends) of chromosomes and more rare near the centromeres of chromosomes. SNPs in the vicinity of recombination hotspots have less correlation with surrounding SNPs compared with SNPs at some distance from hotspots.

Although these findings are of biological interest, there are other features of the HapMap data that are particularly useful for the study of human disease.

Uses of HapMap in Genetic Mapping

Coverage of the Genome

The large database of genome-wide SNPs provided by HapMap has allowed efficient design of genetic association pair-wise $D'$ and $r^2$ values and haplotype combinations—would provide a powerful tool with which to study human genetics and disease.

Figure 1. Genesis of a new SNP correlated with an old SNP. Initially there is only 1 SNP (T/C) in the region depicted. A spontaneous mutation in a single individual converts an A nucleotide into a G nucleotide. After many generations, a new A/G polymorphism has emerged, with 10% of the population having the G allele. Because no recombination between the two SNPs has occurred, all chromosomes with the G allele have a C allele at the other SNP. HapMap indicates single nucleotide polymorphism.

Figure 2. A 2-marker SNP set tags a third SNP. In this example, only SNPs 1 and 2 have been directly genotyped. Because HapMap has only 3 possible haplotypes for these SNPs (A-A-G, A-C-G, T-A-C), in all cases the identity of SNP 3 can be inferred from a multimarker test comprising SNPs 1 and 2. Note that neither SNP 1 nor SNP 2 alone can predict SNP 3. SNP indicates single nucleotide polymorphism.
This process of using genotyped SNPs to infer the identities of additional SNPs, without the need for further genotyping, is termed imputation. A validation study in which imputation was performed to predict the identities of SNPs that had also been directly genotyped found greater than 98% agreement between the results in individuals of European ancestry. Imputation is particularly useful when combining genome-wide data sets that were obtained with different SNP genotyping platforms. For example, in a recent meta-analysis of 3 genome-wide association studies with lipid traits, 2 of the studies were performed using the Affymetrix 500K Mapping Array Set, with the third using the Illumina HumanHap300 BeadChip. Although there was only a small overlap of SNPs directly genotyped by the 2 platforms (≈ 45,000 SNPs), imputation using the haplotypes in the HapMap database generated a greatly enlarged set of genotyped and imputed SNPs (≈ 2.2 million) for all individuals in the 3 studies. Combining information in this way enabled the discovery of 8 new gene regions related to low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, and/or triglycerides.

Interpreting Association Results
The HapMap database facilitates the interpretation of a genetic association result and can help arrive at an “associated” or “critical” interval, a region of the genome likely to contain the causal polymorphism. Given an index SNP with definitive statistical evidence for association with a trait or disease of interest, one can refer to the HapMap database and use the correlation structure to identify other SNPs in LD and thereby define the region in which to look for the causal variant. For example, several genome-wide association studies have highlighted an association of common noncoding SNPs on chromosome 9p21 with coronary artery disease or myocardial infarction. Given the public HapMap resource and such an association result, investigators are readily able to evaluate the patterns on SNP correlation around the index SNP(s) and delimit the region of association. Using data derived from HapMap, Schunkert et al described the correlation structure for SNPs on 9p21 (Figure 3). SNPs spanning a distance of ≈ 60 kilobases are correlated with one of the index SNPs (rs13330499) with r² of at least 0.5. The search for a causal variant for coronary artery disease has now been narrowed from the entire genome to a small span of DNA sequence.

Another such example involves genetic variation on chromosome 1p13 associated with both low-density lipoprotein cholesterol and coronary artery disease, with multiple genome-wide association studies identifying rs599839 as an index SNP for these phenotypes. On interrogation of this SNP in HapMap, it is evident that the set of SNPs in strong LD with rs599839 span a region roughly 100 kilobases in size. In this region lie at least 4 genes—CELSR2, PCSRC1, MYBPHL, and SORT1—and any of these may represent the gene influencing both low-density lipoprotein cholesterol and coronary artery disease. These genes may now be prioritized for the next set of studies (ie, deep sequencing of the 100-kilobase region in humans and manipulation of these 4 positional candidate genes in cells or mice).

HapMap data may also facilitate “fine mapping” of an initial association result. In fine mapping, additional SNPs (beyond the index SNP) within an associated interval are tested to see if they provide stronger evidence for association. As an example, genome-wide association mapping for triglyceride levels identified an SNP in the glucokinase regulatory protein gene (GCKR) as being highly associated with triglyceride levels. The index GCKR SNP was intronic (rs780094) and the associated interval spanned ≈ 400 kilobases and contained 17 genes. To fine-map across the associated interval, an additional 120 SNPs were selected from HapMap to tag the associated interval. With fine mapping, a common missense SNP in GCKR (rs1260326) that changes the amino acid 446 of the protein from proline to leucine emerged as the strongest association signal. These results now raise the next testable hypothesis, that the coding variant affects the function of GCKR (possibly by altering binding to glucokinase) and thereby alters triglyceride and glucose levels.

Limitations of HapMap
A major limitation of the HapMap project is that low-frequency SNPs (ie, with minor allele frequencies between 0.5% and 5%) are incompletely captured in the database. Rare SNPs (<0.5% frequency) are even more underrepresented. As it is likely that an important fraction of disease-causing variants are of low frequency or rare, these will be difficult to identify through the use of tagging SNPs selected from the HapMap database.

An additional limitation is that genotypes are only available for individuals from 4 ethnic groups (European descent in Utah, Yoruban, Japanese, and Han Chinese) at the time of the second phase of HapMap. Although it has been shown that the correlation structures in each of these groups remains valid in other cohorts of the same ethnicity, this may not hold true for ethnicities not represented in HapMap.

Both of these shortcomings are to be squarely addressed by new projects that are now underway. The third phase of HapMap will include genotyping of SNPs in individuals of additional ethnicities beyond the original 4 and thus will extend the utility of HapMap to a wider variety of populations under study worldwide. On an even larger scale, the 1000 Genomes Project, launched in January 2008, aims to fully sequence the genomes of at least 1000 individuals from 11 ethnic/regional groups (including individuals from the original HapMap Project). This effort will markedly increase the number of low-frequency SNPs available for study and, with integration into the existing HapMap database, allow for an extension of the correlation structure to these low-frequency SNPs.

Conclusion
HapMap is a public resource that has critically enabled genome-wide association mapping using common DNA sequence variants. These genetic mapping studies have proven useful in identifying novel contributors to cardiovascular traits including myocardial infarction, atrial fibrillation, and...
lipid levels, diabetes mellitus, statin-induced myopathy, electrocardiographic QT interval, and abdominal aortic aneurysm. Further application of tools such as HapMap should clarify the full spectrum of DNA sequence differences that confer susceptibility to cardiovascular disease.

Acknowledgments
The authors thank Dr Mark J. Daly, who provided the illustration on which Figure 2 is based.

Sources of Funding
Dr Kathiresan is supported by a Doris Duke Charitable Foundation Clinical Scientist Development Award, a charitable gift from the Fannie E. Rippel Foundation, the Donovan Family Foundation, and a K23 career development award from the United States National Institutes of Health. Dr Musunuru is supported by a T32 grant in Cell and Molecular Training for Cardiovascular Biology from the National Institutes of Health.

Disclosures
Dr Musunuru has received consulting fees from Alnylam Pharmaceuticals and honoraria from the American College of Cardiology Foundation within the last year. Dr Kathiresan reports no potential conflicts.

References


**KEY WORDS:** cardiovascular diseases / genes / mapping
HapMap and Mapping Genes for Cardiovascular Disease
Kiran Musunuru and Sekar Kathiresan

doi: 10.1161/CIRCGENETICS.108.813675
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 1942-325X. Online ISSN: 1942-3268

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circgenetics.ahajournals.org/content/1/1/66

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Genetics can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Genetics is online at:
http://circgenetics.ahajournals.org/subscriptions/