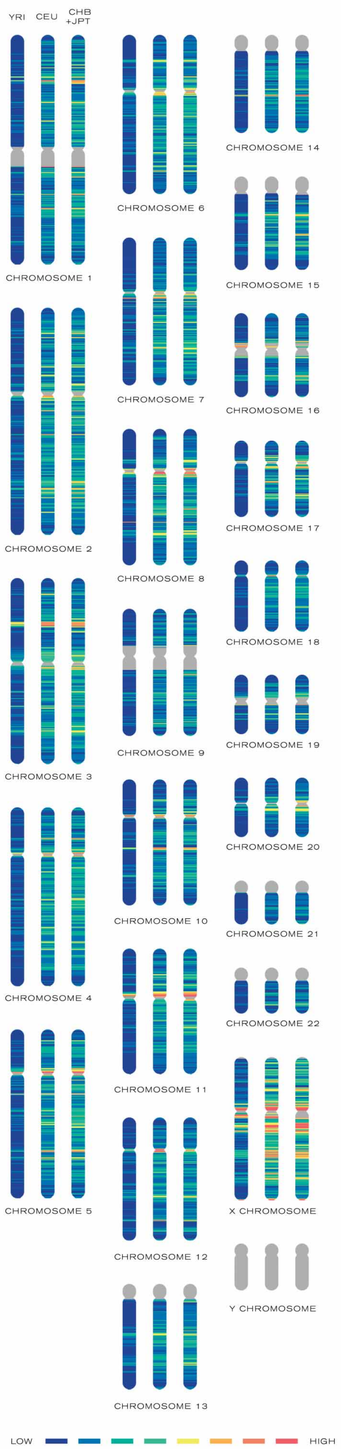
PATTERNS OF HUMAN GENOME VARIATION



The ideograms summarize the extent of linkage disequilibrium (LD) in 1 Mb windows along each chromosome. Within each window a curve was fitted for the decay of LD with distance, and regions where associations among SNPs persist for relatively long distances are shown in red, whereas regions where they persist for only short distances are shown in blue. Gray indicates chromosome regions that cannot be genotyped or analyzed this way. LD does not decay with distance for the non-recombining portions of the Y chromosome and for the mitochondrial genome. No curve was fitted in these regions, since SNPs within them show strong association irrespective of distance. The Yoruba have less LD than the other populations, which is well-known and due to the historically large population sizes of many African populations compared to the bottleneck arising from migration out of Africa for non-African populations.

This region of chromosome 15 illustrates recombination hot spots, long-range haplotypes, and complex LD patterns representative of the HapMap data. The region contains multiple genes, including the Rhesus blood group C glycoprotein (RHCG), a component of the blood antigen responsible for maternal-fetal Rh incompatibility.

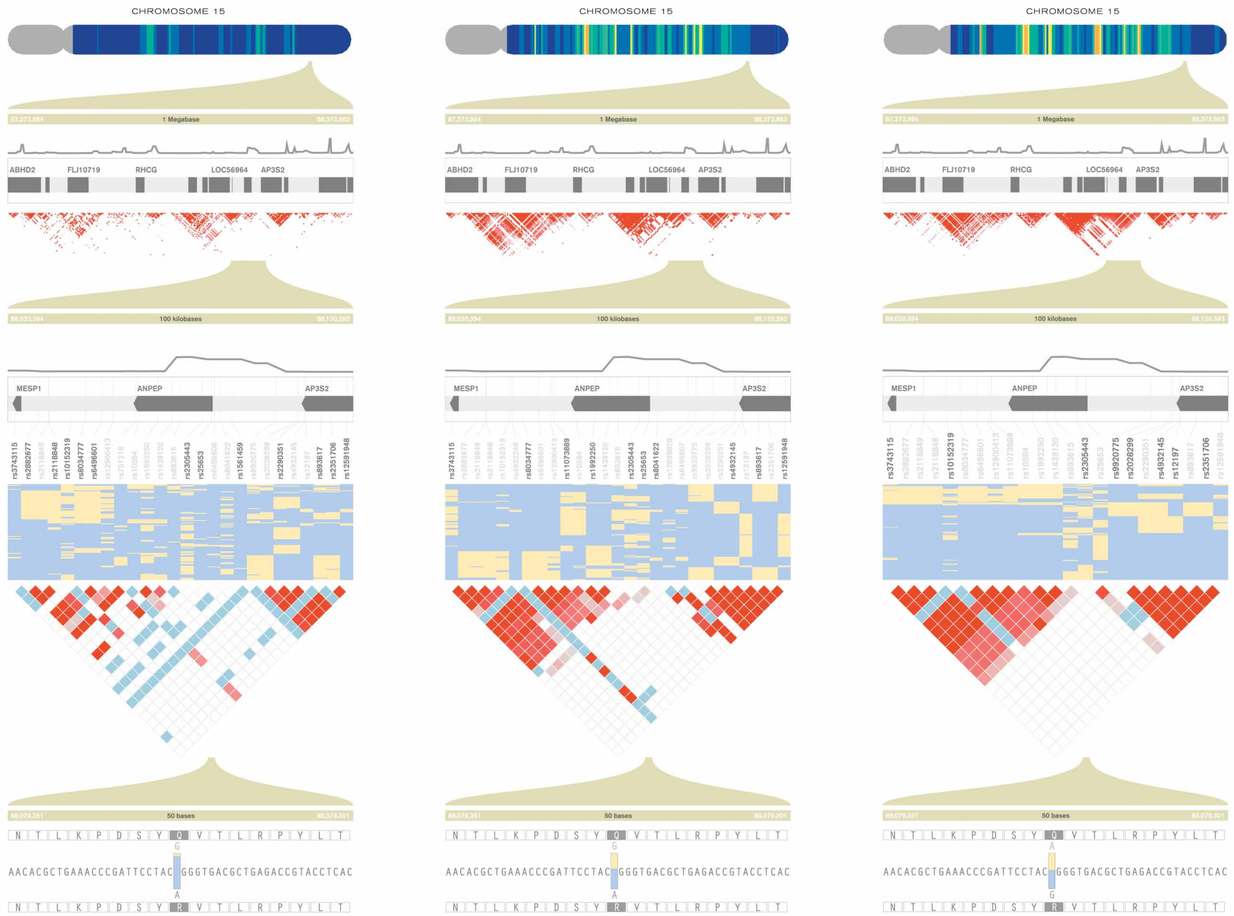
The locations on the chromosome and scale are indicated. The recombination hotspots are shown. Note that the recombination rates are estimated using data from all the populations, not for each population separately. Genes in the region are shown underneath.

The diagram below the genes illustrates how the human genome is composed of regions of strong linkage disequilibrium (LD) where the recombination rate is low, punctuated by regions of weak LD where the recombination rate is high.

Zooming in, the relationship of recombination rate and LD is more apparent. Phased haplotypes in this region are displayed for all the people studied (not including the children in trios). Each row corresponds to a chromosome from a person. SNPs that were genotyped in the region are shown at their locations (columns); tag SNPs are in darker text. The commoner allele for each SNP is blue; the rarer allele is yellow.

At this level of zoom, the level of LD between every pair of SNPs is visible in the "triangle plot." Each diamond in the plot relates the two SNPs located diagonally to its left and right such that the redder the diamond the stronger the LD between the two SNPs.

SNP allele frequencies may vary from population to population. Percentage of color illustrates the relative allele frequencies for each SNP; the commoner allele is blue, the rarer allele is yellow. The amino acids, including variants, are also shown.

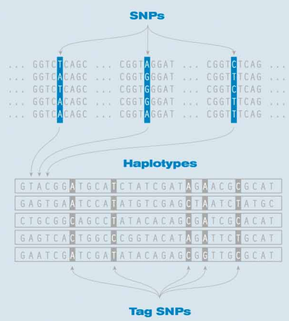


SINGLE NUCLEOTIDE POLYMORPHISMS

Single nucleotide polymorphisms, or SNPs, are sites within chromosomes where the sequence differs at one base among people. At the right, SNPs are shown among non-variable sites in a chromosome region. Each of the two copies of a person's chromosomes displays a variant, called an allele, and the set of SNP alleles across a chromosomal region is called a haplotype. Since many SNPs are highly associated with each other, most information on the patterns of genetic variation is contained in a subset of SNPs, called tag SNPs. Looking at the tag SNPs, instead of all common SNPs, will decrease the time and expense of studies designed to find gene regions associated with a particular disease or drug response. Other, less common, types of DNA sequence differences among people include insertions or deletions of stretches of DNA, and variation in the number of copies of repeated sequences.

LINKAGE DISEQUILIBRIUM

Individuals who carry a specific SNP allele at one site often carry specific alleles at other nearby SNPs, a correlation that is termed Linkage Disequilibrium (LD) because the SNP alleles occur together on a chromosome more often than expected based on their frequencies.



- **There are at least 20 million SNPs in the human genome, of which about 10 million are common (with a frequency of the rare allele of at least 1%).**
- **Approximately 250,000 to 500,000 tag SNPs are expected to provide almost as much mapping information as the 10 million common SNPs.**
- **The density of SNPs on the phase I HapMap is approximately one SNP every 3300 bases. The final version of the map will have more than one SNP every 900 bases.**
- **Most human genetic variation, about 90%, occurs within any human population; about 10% occurs as differences in allele frequencies among populations.**

REFERENCES

Edwards, A.O., Ritter III, R., Abel, K.J., Manning, A., Panhuysen, C., and Farrer, L.A. 2005. Complement factor H polymorphism and age-related macular degeneration. *Science* **308**: 421-424.

Haines, J.L., Hauser, M.A., Schmidt, S., Scott, W.K., Olson, L.M., Gallins, P., Spencer, K.L., Kwan, S.Y., Noureddine, M., Gilbert, J.R., et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**: 419-421.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**: 789-796.

The International HapMap Consortium. 2004. Integrating ethics and science in the International HapMap Project. *Nat. Rev. Genet.* **5**: 467-465.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* doi:10.1038/nature04226 (in press).

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385-389.

Smith, A.V., Thomas, D.J., Munro, H., and Abecasis, G. 2005. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* **15**: 1519-1534.

WEB RESOURCES

The data used in this poster were taken from:
The HapMap Web site - www.hapmap.org
Japanese mirror site - hapmap.jst.go.jp
dbSNP - www.ncbi.nlm.nih.gov/SNP
Entrez Genes - www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene

Additional Web resources:
UCSC Genome Browser - genome.ucsc.edu
Ensembl - www.ensembl.org
National Human Genome Research Institute - www.genome.gov
The Genetic Association Database - geneticassociationdb.nih.gov

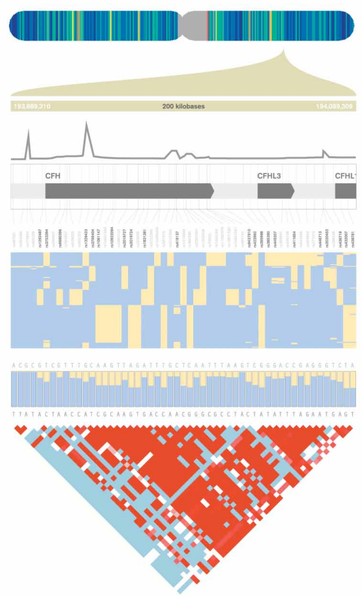
HUMAN GENOME VARIATION

The human genome consists of 3 billion bases of DNA, the sequence of which is 99.9% identical between any two unrelated people. Determining the sites where the DNA sequence varies among people is important for investigating their contribution to susceptibility to complex diseases such as diabetes, cancer, stroke, heart disease, and psychiatric disorders. The HapMap data will greatly facilitate the discovery of the DNA variants that make these genetic contributions. This will enable estimation of individual disease risk, more efficient use of drug therapy, and ultimately a better understanding of the genetic and environmental factors and their interactions that lead to disease, with the aim of improved prevention, diagnosis, and treatment.

THE HAPMAP

The goal of the International HapMap Project was to develop a haplotype map of the human genome that describes the common patterns of genetic variation, in order to find sets of tag SNPs. By describing the shared regions of variation and how they have recombined during human history, researchers can move from studying the genetic basis of health and disease in families to studying many individuals. This collaborative effort of scientists from Canada, China, Japan, Nigeria, the United Kingdom, and the United States started in 2002; by the end of 2005 the completed HapMap is expected to characterize more than 3.5 million SNPs in 270 individuals sampled from four populations: the Yoruba in Ibadan, Nigeria (YRI), Japanese in Tokyo (JPT), Han Chinese in Beijing (CHB), and a set of Utah residents with northern and western European ancestry (CEU) collected by the Centre d'Etude du Polymorphisme Humain (CEPH). The Yoruba and Utah residents provided sets of samples that were from two parents and an adult child, each called a trio. Here, the CHB and JPT data are combined in one analysis panel. Samples from additional populations will be studied in selected regions of the genome to assess how well the HapMap tag SNPs apply to other populations.

The human genome sequence and the HapMap data were used to implicate this region of chromosome 1, which includes the complement factor H (CFH) gene, in age-related macular degeneration (AMD), a progressive blinding disorder that affects about 20 to 25 million people worldwide (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005).



SNPs and LD (in CEU) in the AMD region are shown. Some SNPs are associated with a high risk for AMD and such associations allow researchers to identify regions that may be related to disease. In a region of strong LD there are many such associated SNPs and more work is needed to establish whether the increased risk is due to a SNP that changes an amino acid, a SNP in a regulatory region, or some other type of SNP or other variant.

Poster concept by Lisa Brooks (NHGRI/NIH), William Craumer (Illumina, Inc.), Mark Daly (Massachusetts General Hospital/Harvard Medical School), Sarah Murray, Karen Possemato (both of Illumina, Inc.), Albert Vernon Smith and Lincoln Stein (both of Cold Spring Harbor Laboratory). Poster designed by Ben Fry (Broad Institute of MIT & Harvard).

We are also grateful to the International HapMap Consortium for making the data freely available, to Gonçalo Abecasis, Mark Daly, Paul de Bakker, Peter Donnelly, Lalitha Krishnan, Jonathan Marchini, and Gil McVean for providing their analyses and figures, and to David Altshuler, Aravinda Chakravarti, and Richard Gibbs for their helpful input.