A HapMap harvest of insights into the genetics of common disease

Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins
National Human Genome Research Institute, Bethesda, Maryland, USA.

The International HapMap Project was designed to create a genome-wide database of patterns of human genetic variation, with the expectation that these patterns would be useful for genetic association studies of common diseases. This expectation has been amply fulfilled with just the initial output of genome-wide association studies, identifying nearly 100 loci for nearly 40 common diseases and traits. These associations provided new insights into pathophysiology, suggesting previously unsuspected etiologic pathways for common diseases that will be of use in identifying new therapeutic targets and developing targeted interventions based on genetically defined risk. In addition, HapMap-based discoveries have shed new light on the impact of evolutionary pressures on the human genome, suggesting multiple loci important for adapting to disease-causing pathogens and new environments. In this review we examine the origin, development, and current status of the HapMap; its prospects for continued evolution; and its current and potential future impact on biomedical science.

The International HapMap Project was designed to create a public, genome-wide database of patterns of common human sequence variation to guide genetic studies of human health and disease (1–3). With the publication of the draft human genome sequence in 2001 (4) and the essentially finished version in 2003 (5), the HapMap emerged as a logical next step in characterizing human genomic variation, particularly of the millions of common single-base pair differences among individuals, or SNPs (see Glossary). The HapMap was designed to determine the frequencies and patterns of association among roughly 3 million common SNPs in four populations, for use in genetic association studies. The HapMap has introduced a new paradigm into genomic research, primarily in the form of genome-wide association (GWA) studies, by making possible the cost-efficient assessment of much of the common genomic variation within an individual (1, 6). It has also provided new insights into evolutionary pressures on the human genome and has facilitated functional investigation and cross-population comparisons of candidate disease genes. In addition, it has led to important methodologic advances in imputation of untyped SNPs (that is, reliable estimation of genotypes at SNPs not typed on existing genotyping platforms based on information from typed SNPs) and in assessment of population substructure in genetic association studies. Finally, the open availability of HapMap samples (both DNA and cell lines) and the consent and consultation process through which they were collected have provided a valuable resource for continued development of genomic research methods, such as association studies of gene expression and other cellular phenotypes, and for quality assessment of genotyping data.

Genetic influences on common diseases

Most common diseases are caused by the interplay of genes and environment, with adverse environmental exposures acting on a genetically susceptible individual to produce disease (7, 8). Unlike Mendelian disorders such as sickle cell disease and cystic fibrosis, in which alterations in a single gene explain all or nearly all occurrences of disease, genes underlying common diseases are likely to be multiple, each with a relatively small effect, but act in concert with environmental influences to lead to clinical disease (Figure 1) (9).

Identifying these genetic influences would be quite difficult if the risk-associated allelic variants at a particular disease-causing locus were very rare, so that for a disease to be common there would be many different causative alleles. In contrast, the HapMap was designed to identify more common disease-causing variants based upon the “common disease, common variant” hypothesis, which suggests that genetic influences on many common diseases are attributable to a limited number of allelic variants (one or a few at each major disease locus) that are present in more than 1%–5% of the population (10–12). Evidence supporting this hypothesis was modest at the outset of the HapMap Project, and reliance on the hypothesis sparked considerable controversy (13–16). Understanding how that controversy played out and was ultimately resolved by the remarkable success of the genetic association studies enabled by the HapMap requires an understanding of genetic variation, population genetics, and the evolution of the HapMap itself.

SNPs and linkage disequilibrium

SNPs are sites in the genome sequence of 3 billion nucleotide bases where individuals differ by a single base. Roughly 10 million such sites, on average about one site per 300 bases, are estimated to exist in the human population such that both alleles have a frequency of at least 1% (3). Most SNPs are biallelic, or have only two forms, which contributes to their being relatively easy to type with automated, high-throughput genotyping methods (17). In addition, their generally low rate of recurrent mutation makes them stable markers of human evolutionary history (17).

In theory, identifying common SNPs associated with disease would involve the relatively straightforward — although time-consuming and expensive — task of typing all 10 million common SNPs in individuals with and without disease and looking for sites that differ in frequency between the groups. Such an
approach would be very expensive and would not capture rarer variants or structural variants (such as insertions, deletions, and inversions) that are not identified by genotyping of SNPs. However, the pattern of association among SNPs in the genome suggests a potential shortcut, based on haplotypes and linkage disequilibrium (LD). A haplotype is the combined set of alleles at a number of closely spaced sites on a single chromosome. Nearby SNP alleles tend to be associated with each other, or inherited together more often than expected by chance, because most arise through mutational events that each occur once on an ancestral haplotype background and are inherited with that background, rather than arising multiple times de novo on different backgrounds (18). This is because for most SNPs the rate of mutation, or novel SNP generation, is relatively low (roughly $10^{-8}$ per site per generation, or 30 new variants per haploid gamete), as are the rate of recombination occurring with each meiosis and the number of generations (roughly $10^4$) between currently living individuals and their most recent common ancestor (3). Each new allele is initially associated with the other allelic variants present on the particular stretch of ancestral DNA on which it arose, and these associations are only slowly broken down over time by recombination between SNPs and generation of new variants (Figure 2) (3).

Two polymorphic sites are said to be in LD when their specific alleles are correlated in a population. High LD means that the SNP alleles are almost always inherited together; information about

---

**Glossary**

- **Allele**, an alternative form of a gene or SNP, or another type of variant
- **Biallelic**, having only two possible alleles in a variant, typically a SNP
- **Confidence interval**, the range of values surrounding a point estimate, such as an OR, within which the true value is believed to lie with a specified degree of certainty (typically 95%)
- **Copy number variant**, a DNA sequence of hundreds to thousands of base pairs that is present a variable number of times across individuals
- **Haplotype**, a combination of alleles at multiple linked sites on a single chromosome, all of which are transmitted together
- **Haplotype block**, a region containing strongly associated SNPs
- **Light sequencing**, a process involving an average of 1–3 sequencing reads at any place in the genome
- **Linkage disequilibrium**, the nonrandom association of alleles at two or more sites on the same chromosome
- **Mendelian disorder**, a disorder caused by a single gene defect, which tends to occur in either dominant or recessive inheritance patterns
- **Minor allele frequency**, the proportion of chromosomes in the population carrying the less common variant
- **Odds**, the ratio of the probability of disease to 1 minus the probability of disease
- **Odds ratio**, the ratio of the odds of an event occurring in one group to the odds of it occurring in another group; can be presented in terms of increased risk per copy of the variant (allelic OR), of a heterozygous carrier compared with a noncarrier (heterozygote OR), or of individuals with the heterozygous and homozygous variant genotypes separately compared with those homozygous for the nonrisk allele (genotypic OR)
- **Nonsynonymous SNP**, a SNP for which each allele codes for a different amino acid in the protein sequence
- **Polymorphism**, a form of genetic variation in which each allele occurs in at least 1% of the population
- **Power**, the probability of a study detecting an association, if one exists
- **$r^2$**, LD coefficient, representing the proportion of observations in which two specific pairs of alleles occur together
- **Single nucleotide polymorphism**, a site within the genome that differs by a single nucleotide base across different individuals
- **Tag SNP**, a representative SNP in a region of the genome with high LD to other variants

---

**Figure 1**

Genetic and environmental contributions to monogenic and complex disorders. (A) Monogenic disease. A variant in a single gene is the primary determinant of a monogenic disease or trait, responsible for most of the disease risk or trait variation (dark blue sector), with possible minor contributions of modifier genes (yellow sectors) or environment (light blue sector). (B) Complex disease. Many variants of small effect (yellow sectors) contribute to disease risk or trait variation, along with many environmental factors (blue sector).
SNPs could theoretically produce an enormous number of haplotypes, but selection of the best tag SNPs requires precise mapping of the genome, and to a lesser extent across populations, so that any of SNP1–SNP3 (or SNP4–SNP6) could serve as a marker for genotyping platforms because of stronger associations with additional SNPs in the region or technical ease of genotyping. The International HapMap Project was a consortium among researchers in Canada, China, Japan, Nigeria, the United Kingdom, and the United States, organized to consider the ethical issues, develop the scientific plan, choose the populations and SNPs to be typed, carry out the genotyping and data analysis, and release the data into the public domain (1, 3). The consortium produced a human haplotype map by genotyping 270 samples, from four populations with diverse geographic ancestry, provided by people who gave consent specifically for this project and related research. These samples included 30 trios (mother, father, and adult child) from the Yoruba in Ibadan, Nigeria; 30 trios from the Centre d’Étude du Polymorphisme Humain collection of Utah residents of Northern and Western European ancestry; 45 unrelated Han Chinese in Beijing; and 45 unrelated Japanese in Tokyo (24). The Utah samples were previously collected but were reconsented for this purpose. New samples were collected from the Yoruba, Han Chinese, and Japanese after processes of community engagement (25). The newly collected samples were permanently disconnected from individual identifiers and had no associated phenotype data. Cell lines and DNA from the samples are available for research from the nonprofit Coriell Institute for Medical Research (26).

Figure 2
Breakdown of LD around a new SNP. A mutation generating a novel SNP (red circle) occurs on an existing chromosome (dark blue) with multiple preexisting SNP alleles (dark blue circles) occurring in an ancestral haplotype that spans the entire chromosomal segment shown. After multiple meioses over many generations (arrows), the chromosomal segments flanking this variant will tend to be reshuffled by recombination, as shown by different colors. Over time, therefore, the segment containing the new variant and its surrounding ancestral SNP alleles becomes shorter and occurs on a variety of haplotypes associated with different flanking SNP alleles.

Building a haplotype map of the human genome
The International HapMap Project was a consortium among researchers in Canada, China, Japan, Nigeria, the United Kingdom, and the United States, organized to consider the ethical issues, develop the scientific plan, choose the populations and SNPs to be typed, carry out the genotyping and data analysis, and release the data into the public domain (1, 3). The consortium produced a human haplotype map by genotyping 270 samples, from four populations with diverse geographic ancestry, provided by people who gave consent specifically for this project and related research. These samples included 30 trios (mother, father, and adult child) from the Yoruba in Ibadan, Nigeria; 30 trios from the Centre d’Étude du Polymorphisme Humain collection of Utah residents of Northern and Western European ancestry; 45 unrelated Han Chinese in Beijing; and 45 unrelated Japanese in Tokyo (24). The Utah samples were previously collected but were reconsented for this purpose. New samples were collected from the Yoruba, Han Chinese, and Japanese after processes of community engagement (25). The newly collected samples were permanently disconnected from individual identifiers and had no associated phenotype data. Cell lines and DNA from the samples are available for research from the nonprofit Coriell Institute for Medical Research (26).

Figure 3
Tag SNPs can define common haplotypes. Variable sites (SNPs) are shown by colored bars in this simplified example (adjacent SNPs are generally separated by longer distances). Complete independence of these 6 SNPs would predict the possibility of 2^6 or 64 different haplotypes (because n biallelic SNPs could generate 2^n haplotypes), but in reality just 4 haplotypes comprise 90% of observed chromosomes, indicating that LD is present. To be specific, SNP1, SNP2, and SNP3 are strongly correlated, and SNP4, SNP5, and SNP6 are strongly correlated, so that any of SNP1–SNP3 (or SNP4–SNP6) could serve as tags for the other 2 SNPs in each group. Specific tags may be chosen for genotyping platforms because of stronger associations with additional SNPs in the region or technical ease of genotyping.

The allele of one SNP in an individual is strongly predictive of the allele of the other SNP on that chromosome (Figure 3). The LD between many neighboring SNPs generally persists because meiotic recombination does not occur at random, but is concentrated in recombination hot spots (19). Adjacent SNPs that lack a hot spot between them are likely to be in strong LD. A commonly used measure of LD, r^2, can be interpreted as the proportion of variation in one SNP explained by another, or the proportion of observations in which two specific pairs of their alleles occur together. Two SNPs that are perfectly correlated have an r^2 of 1.0, so that allele A of SNP1 in Figure 3, for example, is always observed with allele C of SNP2, and vice versa, while an r^2 of 0 could be interpreted as an observation of allele A of SNP1 providing no information at all about which allele of SNP4 is present.

Because humans are a relatively young species, and because recombination does not occur at random, there have generally not been enough recombination events to separate a variant from the ancestral background on which it arose (20). A small number of SNPs could theoretically produce an enormous number of haplotypes if every SNP allele could occur in combination with every other SNP allele (n biallelic SNPs could generate 2^n haplotypes), but in practice, far fewer combinations make up the bulk of the haplotypes observed in a population (Figure 3) (18, 21). Because of the strong associations among the SNPs in most chromosomal regions, only a few carefully chosen SNPs (known as tag SNPs; ref. 3) need to be typed to predict the likely variants at the rest of the SNPs in each region.

The size of regions of strong LD varies dramatically across the genome, and to a lesser extent across populations, so that SNPs selected at random or even those spaced at regular intervals across the genome will not efficiently capture the bulk of genomic variation (3). The mean size of regions of strongly associated SNPs, sometimes called haplotype blocks, is estimated to be 22 kb in populations of European or Asian ancestry and 11 kb in populations of recent African ancestry (18). This difference among populations is expected based on population size and migration history; compared with the parent populations, populations with founder effects (a few relatively isolated individuals whose descendants intermarry) have larger regions with stronger associations among SNPs. It has been estimated that most of the variation in the human genome could be captured by genotyping several hundred thousand to 1 million tag SNPs, but selection of the best tag SNPs requires precise mapping of the patterns of LD (3). This was the justification for developing the human haplotype map (1, 3, 22, 23).
Within blocks also varies across populations, with the Yoruba having recent African ancestry (29). Fluctuations in estimates of allele frequency and LD because of small sample sizes also limit the transferability of HapMap-derived tag SNPs, so additional samples from the populations used to develop the HapMap as well as from seven more populations have recently been genotyped across the genome (30).

Approximately 1 million SNPs were genotyped in phase I of the project, and a description was published in 2005 (1). This was followed by the phase II HapMap of over 3 million SNPs, published in 2007 (2). Genotyping in phase II was attempted for about 4.4 million distinct SNPs, of which roughly 1.3 million either could not be typed, were not polymorphic in any of the populations, or did not pass genotyping quality control filters. Certain regions of the genome were recognized as being challenging to study, such as centromeres, telomeres, gaps in genome sequence, and segmental duplications, and only one attempt was made to develop a genotyping assay before such a region was declared to be not HapMap-able (1). All the genotype data are freely available from the HapMap Data Coordination Center (27) and dbSNP (28).

These data revealed the pattern of association among SNPs in the genome and how these patterns vary across populations. Although the four populations studied show generally similar patterns of variation, the Yoruba population has less LD overall and shorter haplotype blocks, as noted above, but the regions with higher LD are similar across the populations. The diversity of haplotypes within blocks also varies across populations, with the Yoruba having an average of 5.6 haplotypes in each block compared with 4.0 in the Japanese and Han Chinese populations (1).

Studies in additional populations have shown that the tag SNPs chosen using the HapMap are generally transferable across other populations, but there are some limitations, particularly for rarer SNPs and for populations with substantial proportions having recent African ancestry (29). Fluctuations in estimates of allele frequency and LD because of small sample sizes also limit the transferability of HapMap-derived tag SNPs, so additional samples from the populations used to develop the HapMap as well as from seven more populations have recently been genotyped across the genome (30).

**Advances in technology for high-throughput SNP genotyping**

Advances in genotyping technology have vastly increased the number of variants that can be typed and decreased the per-sample costs (31–34). These advances have made possible the dense genotyping needed to capture the majority of SNP variation within an individual at a sufficiently low cost to allow the large sample sizes needed for comparison of individuals with and without disease. When the HapMap Project began, the cost per sample per SNP was about $0.40; by 2005 the cost had dropped to about $0.01, and the current cost is about one-tenth of that for platforms typing nearly 1 million SNPs at once.

Information generated by the HapMap on LD patterns among SNPs has permitted the design of efficient and comprehensive genotyping platforms by elucidating tag SNPs that serve as proxies for the largest number of SNPs and eliminating redundant SNPs or SNPs that cannot be assayed reliably (1). Currently, genome-wide scans cost less than $1,000 per sample and include about 1 million SNPs, with more SNPs in regions with low LD than in regions with high LD. As genotyping platforms are developed to allow for an increase in the number of tag SNPs typed, they capture more variation in every population, so that even samples of recent African ancestry have most of the genome covered at high $r^2$ (Table 1) (2).

Accuracy of these platforms is paramount, because genotyping errors such as incorrectly typing some heterozygotes as homozygotes can cause spurious results and obscure the true associations, particularly if errors are differential between cases and controls (35, 36). Genotyping errors can also affect parent-offspring trio studies that are robust to other types of bias, such as differences in ancestral origin (population structure) between individuals with and without disease (37). Efforts to improve the accuracy of genotyping platforms and genotype calling algorithms are continuing and rely heavily on use of HapMap samples and data for quality assessment (32, 35, 38). An important step in evaluating the reliability of findings at present is ensuring that they are repeated on a second, independent genotyping platform (39).

Cost efficiency of genotyping platforms is a major consideration for GWA studies because of the very large sample sizes needed to detect genetic variants of modest effect. To provide the same statistical power for detecting a true association between a genetic variant and a disease, assuming such an association exists, sample sizes must increase with the following: (a) greater number of genotypes and association tests performed, and thus greater probability of spurious associations (type I error); (b) greater genotyping error or phenotypic misclassification; (c) lower size of the genetic effect (risk of disease conferred by the disease-associated allele); (d) lower frequency of the risk allele; (e) lower $r^2$ between the disease-associated SNP and the tag SNP typed on the platform; and (f) heterogeneity of the genetic association, caused by multiple genes that contribute to the disease, ancestry differences across population subsets, or gene-gene or gene-environment interactions.

The number of tests is a major factor in determining the statistical power of GWA studies, in which 10$^4$ or more association tests (at least one for each SNP) are performed. Although these tests are not strictly independent because of LD, the current convention is to apply a Bonferroni correction (which assumes independence and is thus overly conservative) by dividing the conventional P value of 0.05 by the number of tests performed (40). This requires P values in the $5 \times 10^{-7}$ to $5 \times 10^{-8}$ range to define an association, a stringent level of significance. Were one to be satisfied with a P value of 0.05, detecting a variant of 10% allele frequency conferring a 1.5-fold increased risk with 80% statistical power would require only 360 cases and 360 controls, but 50,000 potentially spurious associations would be expected by chance out of 1 million SNPs tested. Lowering the P value to $5 \times 10^{-7}$ vastly reduces the number of spurious associations but requires a more than 4-fold increase in the sample size, to roughly 1,590 cases and 1,590 controls, for the same statistical power (41). Risk associated with a variant is often

<table>
<thead>
<tr>
<th>Platform</th>
<th>HapMap population sample</th>
<th>YRI</th>
<th>CEU</th>
<th>CHB + JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix GeneChip 500K</td>
<td></td>
<td>46</td>
<td>68</td>
<td>67</td>
</tr>
<tr>
<td>Affymetrix SNP Array 6.0</td>
<td></td>
<td>66</td>
<td>82</td>
<td>81</td>
</tr>
<tr>
<td>Illumina HumanHap300</td>
<td></td>
<td>33</td>
<td>77</td>
<td>63</td>
</tr>
<tr>
<td>Illumina HumanHap550</td>
<td></td>
<td>55</td>
<td>88</td>
<td>83</td>
</tr>
<tr>
<td>Illumina HumanHap650Y</td>
<td></td>
<td>66</td>
<td>89</td>
<td>84</td>
</tr>
<tr>
<td>Perlegen 600K</td>
<td></td>
<td>47</td>
<td>92</td>
<td>84</td>
</tr>
</tbody>
</table>

Data represent percent of SNPs tagged at $r^2 \geq 0.8$. Values assume all SNPs on the platform are informative and pass quality control. YRI, Yoruba in Ibadan, Nigeria; CEU, subsample of Utah residents of Northern European ancestry selected from Centre d’Étude du Polymorphisme Humain samples; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo. From the International HapMap Consortium, 2007 (3).
assessed by the odds ratio (OR), the odds of disease in individuals with the variant divided by the odds of disease in those without the variant. ORs for many of the genetic variants believed to contribute to the risk of complex diseases are likely to fall in the 1.2–1.3 range or lower, considerably lower than the OR of 3.2, for example, for Alzheimer disease associated with the apolipoprotein E ε4 allele (42) or the OR of 4.1 for deep venous thrombosis associated with oral contraceptive use (43). Statistical power is known to decline steeply below an OR of 1.2 (44) and as minor allele frequency (MAF) falls below an OR of 1.2 (44) and as minor allele frequency (MAF) falls below an OR of 1.2 (44) and as minor allele frequency (MAF) falls below an OR of 1.2 (44) and as minor allele frequency (MAF) falls below an OR of 1.2 (44).

**Beyond SNPs: copy number variants and other structural variation**

Current generation high-throughput genotyping platforms are extraordinary efficient at genotyping SNPs, but, as stated above, they are less effective at genotyping structural variants, such as insertions, deletions, inversions, and copy number variants. Although not as common as SNPs, these variants also occur commonly in the human genome (45). The HapMap was not designed to capture these variants, although it can be used indirectly to do so, particularly for small deletions that are in LD with SNPs (46–48). Copy number variants, in which stretches of genomic sequence of roughly 1 kb to 3 Mb in size are deleted or are duplicated in varying numbers, have gained increasing attention because of their apparent ubiquity and potential dosage effect on gene expression (47, 49, 50). A variety of diseases such as DiGeorge syndrome and α-thalassemia have been shown to be caused by large deletions, insertions, and other structural variants, and the potential for structural variants to influence phenotypes in healthy individuals is now recognized (51, 52). Expansion and refinement of current genotyping platforms increasingly focus on capturing copy number variants adequately, and some success has already been achieved (38, 53). Array and sequencing methods are also being used to type structural variants using the HapMap samples for development and cross-validation of the methods (54, 55).

**Application of the HapMap to common disease: success stories**

The technological advances directly stimulated or indirectly facilitated by the HapMap have had a profound impact on the study of the genetics of common diseases, exceeding even the expectations of the project’s originators (4). Not only has the HapMap enabled a new generation of genetic association studies through the application of high-density, genome-wide genotyping to carefully characterized individuals with and without disease, but it has also stimulated the development and testing of analytic methods for reducing spurious associations (56, 57), assessing claims of replication of genotype-phenotype associations (39, 58), identifying and adjusting for ancestry differences among individuals and groups (35, 59, 60), and imputing untyped SNPs across different genotyping platforms (61).

The short history of high-density GWA scanning (i.e., about 100,000 SNPs or more) to date has demonstrated the striking success of this approach in finding genetic variants associated with disease. Variants or regions associated with nearly 40 complex diseases or traits have been identified and replicated in diverse population samples (Figure 4). Complex conditions as dissimilar as macular degeneration and exfoliative glaucoma (Table 2), diabetes (Table 3), cancer (Table 4), inflammatory bowel disease (Table 5), cardiovascular disease (Table 6), neuropsychiatric conditions (Table 7), autoimmune and infectious diseases (Table 8), and a variety of anthropometric and laboratory traits (Table 9) have recently yielded strong, convincing, replicated associations in GWA scanning. Several of these discoveries have suggested etiologic pathways not previously implicated in these diseases, such as the autophagy pathway in inflammatory bowel disease (62), the complement pathway in macular degeneration (63), and the HLA-C

**Table 2**

GWA studies in eye diseases

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size</th>
<th>Region</th>
<th>Gene</th>
<th>Strongest SNP-risk allele</th>
<th>Risk allele frequency in controls</th>
<th>P</th>
<th>OR per copy or for heterozygote (95% CI)</th>
<th>Platform manufacturer and SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration (63)</td>
<td>96 cases</td>
<td>NR</td>
<td>1q31</td>
<td>rs380390-C</td>
<td>0.70 (HapMap CEU)</td>
<td>4 × 10⁻⁴</td>
<td>4.6</td>
<td>Affymetrix: 103,611</td>
</tr>
<tr>
<td>Wet age-related macular degeneration (95)</td>
<td>96 cases</td>
<td>NR</td>
<td>10q26.13</td>
<td>rs11200638-A</td>
<td>NR</td>
<td>8 × 10⁻¹²</td>
<td>1.60</td>
<td>Affymetrix: 97,824</td>
</tr>
<tr>
<td>Exfoliation glaucoma (65)</td>
<td>75 cases</td>
<td>15q24.1</td>
<td>LOKL1</td>
<td>rs3825942-G</td>
<td>0.85</td>
<td>3 × 10⁻²¹</td>
<td>20.10</td>
<td>Illumina: 304,250</td>
</tr>
<tr>
<td></td>
<td>14,474 con</td>
<td>254 cases</td>
<td>198 con</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Only studies with at least one association significant at 9.9 × 10⁻⁷ or less are shown, regardless of number of SNPs tested. Allele frequencies, P values, and ORs derived from the largest sample size — typically a combined analysis (initial plus replication studies) — are shown; however, if this was not reported, statistics from the initial study sample are shown. Only SNPs significant at 9.9 × 10⁻⁷ in combined analysis and only one of close LD pair are shown unless there was evidence of independent association. Gene regions corresponding to SNPs were identified from the UCSC Genome Browser (104). Named genes refer to either the most plausible candidate gene(s) within the associated region or the gene nearest the (currently known) most associated SNP in a region, with the associated region being determined by the LD structure. Affymetrix (Santa Clara, California, USA) platforms include GeneChip Mapping 100K Set, Centruyon arrays, and GeneChip 500K Mapping Array Set. Illumina Inc. (San Diego, California, USA) platforms include Sentrix Human-Hap240, HumanHap300 BeadChip, Sentrix HumanHap300, Infinium II Human Hap550, Human Hap300-duo+ and Sentrix Human-1 Genotyping BeadChip.
locus in control of viral load in HIV infection (64). Note that the estimated ORs for most of these associations are relatively modest, at 2.0 or less, although smaller studies of rare diseases can give quite large ORs (and very wide confidence intervals), as in the case of the 20-fold increased risk (95% confidence interval, 10.8 to 37.4) associated with rs3825942-G in exfoliation glaucoma (65).

A major strength of the genome-wide approach facilitated by HapMap-based genotyping platforms has been its freedom from reliance on prior knowledge, imperfect as it is, of genes likely to be related to the trait of interest. Instead, GWA studies survey a risk of familial malignant melanoma (76), and it is surprising to see potentially regulatory variants of these same cell-cycle genes implicated in this additional common condition.

In addition to its pivotal role in the design of genotyping platforms, the HapMap played important roles in these discoveries, include providing better estimates of allele frequencies (64), comparing allele frequencies across the four HapMap populations (72), identifying additional variants for testing, and defining genomic regions carrying no known genes at all (69). Of considerable interest in determining pathophysiology have been variants or regions implicated in multiple diseases, such as the 8q24 region in prostate, breast, and colorectal cancer (66, 70, 71) and the PTPN2 gene in type 1 diabetes and Crohn disease (35). Notable among these are the CDKN2A/B cell-cycle variants on chromosome 9p21, which have been implicated in coronary disease (72, 73), type 2 diabetes (61, 74), and frailty (75). Prior interest in CDKN2A/B focused on the fact that germline deletions of these genes confer a risk of familial malignant melanoma (76), and it is surprising to see potentially regulatory variants of these same cell-cycle genes implicated in these additional common conditions.

In addition to its pivotal role in the design of genotyping platforms, the HapMap played important roles in these discoveries, include providing better estimates of allele frequencies (64), comparing allele frequencies across the four HapMap populations (72), identifying additional variants for testing, and defining genomic regions carrying no known genes at all (69). Of considerable interest in determining pathophysiology have been variants or regions implicated in multiple diseases, such as the 8q24 region in prostate, breast, and colorectal cancer (66, 70, 71) and the PTPN2 gene in type 1 diabetes and Crohn disease (35). Notable among these are the CDKN2A/B cell-cycle variants on chromosome 9p21, which have been implicated in coronary disease (72, 73), type 2 diabetes (61, 74), and frailty (75). Prior interest in CDKN2A/B focused on the fact that germline deletions of these genes confer a risk of familial malignant melanoma (76), and it is surprising to see potentially regulatory variants of these same cell-cycle genes implicated in these additional common conditions.

### Table 3

GWA studies in diabetes

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size Initial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1 diabetes (105)</td>
<td>1,963 cases 2,938 con</td>
</tr>
<tr>
<td>Type 1 diabetes (35)</td>
<td>1,963 cases 2,938 con</td>
</tr>
<tr>
<td>Type 1 diabetes (106)</td>
<td>563 cases 1,146 con 483 trios</td>
</tr>
<tr>
<td>Type 2 diabetes (78)</td>
<td>1,380 cases 1,323 con 2,617 cases 2,894 con</td>
</tr>
<tr>
<td>Type 2 diabetes (61)</td>
<td>1,161 cases 1,174 con 1,215 cases 1,258 con</td>
</tr>
<tr>
<td>Type 2 diabetes (90)</td>
<td>1,464 cases 1,467 con 5,065 cases 5,785 con</td>
</tr>
<tr>
<td>Type 2 diabetes (107)</td>
<td>1,399 cases 5,275 con</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1 diabetes</td>
<td>4,000 cases 5,000 con 2,997 trios</td>
</tr>
<tr>
<td>Type 1 diabetes (35)</td>
<td>See ref. 105</td>
</tr>
<tr>
<td>Type 1 diabetes (106)</td>
<td>2,350 individ in 549 families; 390 trios</td>
</tr>
<tr>
<td>Type 2 diabetes (78)</td>
<td>8q24.11</td>
</tr>
<tr>
<td>Type 2 diabetes (61)</td>
<td>3q27.2 6p22.3 9p21.3 11p12</td>
</tr>
<tr>
<td>Type 2 diabetes (90)</td>
<td>5q65 cases 5q785 con</td>
</tr>
<tr>
<td>Type 2 diabetes (107)</td>
<td>2,437 cases</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2 diabetes (74)</td>
<td>3,757 cases 5,346 con</td>
</tr>
<tr>
<td>Type 2 diabetes (35)</td>
<td>See ref. 74</td>
</tr>
</tbody>
</table>

Studies, statistics, and platforms are reported as described in Table 2. individ, individuals. *Number that passed quality control. Individuals studied were of European ancestry.

http://www.jci.org Volume 118 Number 5 May 2008
ing LD blocks and the genes contained within them. These LD patterns are critical in following up initial association findings because they help in selecting variants for cost-effective follow-up genotyping (64, 77), suggesting independence of closely located SNPs in regions of low LD, interpreting failure to detect associations with previously identified variants in populations with varying LD patterns (78), and, crucially, defining haplotypes containing the disease-associated variants (64) (see Use of HapMap data in association studies and Use of HapMap cell lines in genomic research).

### Use of HapMap data in association studies

Designing and refining genotyping platforms
Providing estimates of allele frequencies unbiased by identification in individuals with specific diseases
Estimating ancestry proportions and identifying population structure
Identifying other variants in regions of interest, especially variants with low LD to tag SNPs in the study population
Imputing SNPs that were typed in the HapMap but not in a particular study
Allowing integration of results from studies that used different genotyping platforms, based on imputing SNPs typed by one platform but not others
Defining LD patterns for:
  Selecting a parsimonious set of variants for follow-up genotyping
  Interpreting failure to detect previously reported robust associations
  Showing the extent of a region to be examined for functional studies
  Defining haplotypes, especially for disease-associated variants
Studying evolutionary pressures, including positive, negative, and balancing selection
Studying genomic structure, recombination rates, and mutation rates

### Table 4

**GWA studies in cancer**

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size</th>
<th>Region</th>
<th>Gene</th>
<th>Strongest SNP-risk allele</th>
<th>Risk allele frequency in controls</th>
<th>OR per copy or for heterozygote (95% CI)</th>
<th>Platform and SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate cancer (66)</td>
<td>1,452 cases</td>
<td>8q24.21</td>
<td>Intergenic</td>
<td>rs16901979-A</td>
<td>HapC</td>
<td>1.79 (1.53–2.11)</td>
<td>Illumina: 316,515</td>
</tr>
<tr>
<td>Prostate cancer (67)</td>
<td>1,172 cases</td>
<td>8q24.21</td>
<td>Intergenic</td>
<td>rs6983267-G</td>
<td></td>
<td>2.10 (1.75–2.53)</td>
<td>Illumina: 538,548</td>
</tr>
<tr>
<td>Prostate cancer (108)</td>
<td>1,501 cases</td>
<td>17q24.3</td>
<td>Geno-poor</td>
<td>rs4430796-A</td>
<td></td>
<td>1.26 (1.13–1.41)</td>
<td>Illumina: 310,520</td>
</tr>
<tr>
<td>Breast cancer (94)</td>
<td>390 cases</td>
<td>10q26.13</td>
<td>FGFR2</td>
<td>rs2981582-G</td>
<td></td>
<td>1.26 (1.23–1.30)</td>
<td>Perlegen: 205,586</td>
</tr>
<tr>
<td>Breast cancer (70)</td>
<td>1,145 cases</td>
<td>10q26.13</td>
<td>FGFR2</td>
<td>rs1219648-G</td>
<td></td>
<td>1.20 (1.16–1.24)</td>
<td>Illumina: 528,173</td>
</tr>
<tr>
<td>Breast cancer (92)</td>
<td>1,599 cases</td>
<td>2q35</td>
<td>Intergenic</td>
<td>rs13387042-A</td>
<td></td>
<td>1.08 (1.05–1.11)</td>
<td>Illumina: 311,524</td>
</tr>
<tr>
<td>Colorectal cancer (109)</td>
<td>930 cases</td>
<td>8q24.21</td>
<td>Intergenic</td>
<td>rs6983267-G</td>
<td></td>
<td>1.27 (1.16–1.39)</td>
<td>Illumina: 547,647</td>
</tr>
<tr>
<td>Colorectal cancer (71)</td>
<td>1,257 cases</td>
<td>8q24.21</td>
<td>Intergenic</td>
<td>rs13281615-T</td>
<td></td>
<td>1.27 (1.16–1.39)</td>
<td>Illumina: 547,647</td>
</tr>
<tr>
<td>Colorectal cancer (110)</td>
<td>940 cases</td>
<td>18q21.1</td>
<td>SMAD7</td>
<td>rs4939827-T</td>
<td></td>
<td>1.16 (1.09–1.27)</td>
<td>Illumina: 547,647</td>
</tr>
</tbody>
</table>

Studies, statistics, and platforms are reported as described in Table 2. *Number that passed quality control. †Allele frequencies, P values, and ORs shown for study of these individuals of European ancestry; also studied were 373 cases and 372 controls of African ancestry.
possibility of using general population-based screening, or more targeted screening of individuals with positive family histories for discoveries based on the HapMap, these findings have raised the possibility of using general population-based screening, or more targeted screening of individuals with positive family histories for these conditions, for identifying high-risk, presymptomatic subjects, determining the earliest manifestations of these conditions, and facilitating early trials of preventive therapies (79). Although the increases in risk detected in these studies are typically modest, in the 1.2- to 1.5-fold range as noted above, these associations can point the way to important therapeutic avenues and, when considered in combination, may identify individuals at substantially increased risk (61). This information can be particularly important, even in the absence of specific pharmaceutical agents targeted to such individuals, for more aggressive efforts to reduce known risk factors that can be modified, such as obesity in prediabetes and smoking in age-related macular degeneration (AMD) (80, 81). Even modest risk factors may be valuable in individualizing surveillance programs such as mammography, prostate-specific antigen (PSA) screening, or colonoscopy, although further research will be needed.

Clinical significance of HapMap success stories
As noted above, GWA studies have provided startling new insights into pathophysiology, such as the role of the complement system in macular degeneration (63) or the potential for autophagy to cause disease (62).

In addition to the pathophysiologic implications of genetic discoveries based on the HapMap, these findings have raised the possibility of using general population-based screening, or more targeted screening of individuals with positive family histories for these conditions, for identifying high-risk, presymptomatic subjects, determining the earliest manifestations of these conditions, and facilitating early trials of preventive therapies (79). Although the increases in risk detected in these studies are typically modest, in the 1.2- to 1.5-fold range as noted above, these associations can point the way to important therapeutic avenues and, when considered in combination, may identify individuals at substantially increased risk (61). This information can be particularly important, even in the absence of specific pharmaceutical agents targeted to such individuals, for more aggressive efforts to reduce known risk factors that can be modified, such as obesity in prediabetes and smoking in age-related macular degeneration (AMD) (80, 81). Even modest risk factors may be valuable in individualizing surveillance programs such as mammography, prostate-specific antigen (PSA) screening, or colonoscopy, although further research will be needed.

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size</th>
<th>Replication</th>
<th>Region</th>
<th>Gene</th>
<th>Strongest SNP-risk allele</th>
<th>Risk allele frequency in controls</th>
<th>$P$</th>
<th>OR per copy or for heterozygote (95% CI)</th>
<th>Platform manufacturer and SNPs$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBD (96)</td>
<td>547 cases</td>
<td>401 cases$^6$</td>
<td>1p31</td>
<td>$IL23R$</td>
<td>rs11209026-A</td>
<td>$4 \times 10^{-11}$</td>
<td>3.84 (2.33–6.66)$^6$</td>
<td>2.22 (1.37–3.70)$^6$</td>
<td>Illumina: 308,332</td>
</tr>
<tr>
<td>Crohn disease (68)</td>
<td>547 cases</td>
<td>1,266 cases</td>
<td>5p13.1</td>
<td>Intergenic</td>
<td>rs1373962-?</td>
<td>0.59</td>
<td>2 \times 10^{-12}</td>
<td>1.46 (NR)</td>
<td>Illumina: 302,451</td>
</tr>
<tr>
<td>Crohn disease (62)</td>
<td>946 cases</td>
<td>353 cases</td>
<td>2q37.1</td>
<td>$ATG16L1$</td>
<td>rs2241880-G</td>
<td>0.55</td>
<td>1 \times 10^{-13}</td>
<td>1.45 (1.27–1.64)</td>
<td>Illumina: 304,413</td>
</tr>
<tr>
<td>Crohn disease (111)</td>
<td>1,748 cases</td>
<td>1,182 cases</td>
<td>1q24.3</td>
<td>Intergenic</td>
<td>rs1000113-T</td>
<td>0.08</td>
<td>3 \times 10^{-8}</td>
<td>1.47 (1.22–1.76)</td>
<td>Affymetrix: 469,557</td>
</tr>
<tr>
<td>Crohn disease (35)</td>
<td>1,748 cases</td>
<td>See ref. 111</td>
<td>3p21.31 Many genes</td>
<td>rs9858542-?</td>
<td>0.23</td>
<td>3 \times 10^{-8}</td>
<td>1.14 (1.14–1.31)</td>
<td>Affymetrix: 469,557</td>
<td></td>
</tr>
<tr>
<td>Crohn disease (112)</td>
<td>382 trios</td>
<td>521 trios</td>
<td>1p31.3</td>
<td>$IL23R$</td>
<td>rs11209026-A</td>
<td>$4 \times 10^{-11}$</td>
<td>3.84 (2.33–6.66)$^6$</td>
<td>2.22 (1.37–3.70)$^6$</td>
<td>Illumina: 308,332</td>
</tr>
<tr>
<td>Celiac disease (113)</td>
<td>778 cases</td>
<td>991 cases</td>
<td>4q27</td>
<td>$KIA1109$, $TENR$, $IL2$, $IL21$</td>
<td>rs8822844-G</td>
<td>0.82</td>
<td>1 \times 10^{-14}</td>
<td>1.59 (1.41–1.75)</td>
<td>Illumina: 310,605</td>
</tr>
<tr>
<td>Gallstones (114)</td>
<td>280 cases</td>
<td>2,000 cases</td>
<td>2p24.2</td>
<td>$ABCG8$</td>
<td>rs11887534-C</td>
<td>0.10</td>
<td>1 \times 10^{-14}</td>
<td>2.20 (1.80–2.60)</td>
<td>Affymetrix: 382,492</td>
</tr>
</tbody>
</table>

Studies, statistics, and platforms are reported as described in Table 2. IBD, inflammatory bowel disease; ?, risk allele not indicated. $^4$Number that passed quality control. $^6$Also studied were 883 families and 1,119 affected offspring. $^{1-3}$Non-Jewish. $^5$Jewish.
to explore the effectiveness of such approaches. To the degree that they determine treatment response, genetic variants may also be useful in tailoring pharmacologic therapy to individuals most likely to respond — and not react adversely — to specific treatments (82).

In the long run, the greatest contribution of genetic discoveries facilitated by the HapMap may be in the identification of new therapeutic targets. Such treatments may well be effective in individuals without the specific genetic variant that led to the discovery of these targets. Perhaps the best example is the development of HMG-CoA reductase inhibitors that effectively lower cholesterol levels in nearly everyone who takes them — except, ironically, in individuals with homozygous absence of LDL-receptors who were instrumental in identifying this key metabolic pathway (83). Even variants with very modest ORs may provide clues to key drug tar-
Inferences about population genetics from the HapMap

An important use of HapMap data is to test for the presence of population structure, or allele frequency differences related to geographic (and presumably ancestral) differences within and across study populations, even in relatively homogeneous groups, such as the Britons studied in the Wellcome Trust Case Control Consortium (WTCCC; ref. 35). Thirteen genomic regions were found to differ significantly among geographic areas of Great Britain once samples of non-European ancestry were omitted (based on estimates of the genetic distance of individual WTCCC samples from the three original HapMap analysis panels, another key use of HapMap data), but this divergence had little impact on the genetic associations identified with the seven common diseases studied by that consortium. Although some geography-based differences may be just the result of population drift or founder effects, they do provide tantalizing clues to possible selection pressures on populations ancestral to those now residing in the United Kingdom (35).

In fact, HapMap data have provided critical evidence in support of recent positive selection, or selection in favor of new alleles, for genes related to response to infectious agents such as malaria, dietary factors such as disaccharides and fatty acids, and pigmentation differences that confer advantages at different latitudes (86, 87). Such analyses rely on the fact that under strong positive selection, an allele may rise to high frequency so rapidly that associations extend for substantial distances along chromosomes (the long-range haplotype; ref. 88) because there has not been time for them to be broken down by recombination (89). Regions of unusually low diversity suggestive of such selective sweeps have sometimes been detected in three or all four HapMap population samples, but are more commonly found in one or two populations, presumably because of local environmental selective pressures. They are also easier to detect in a single population by comparison with the other populations (88). Investigation of such loci may yield valuable insight into pathways governing responses to environmental pathogens and other functional effects as yet unsuspected.

HapMap data have also been crucial in facilitating the pooling and comparison of association data across populations, so that differences in ancestral background can now be adjusted for in a continuous fashion without loss of data through exclusion or loss of statistical power through stratification (90). Allele frequency differences in HapMap populations have been used to suggest reasons for differences in associations in individuals of varying

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size</th>
<th>Region</th>
<th>Gene</th>
<th>Strongest SNP-risk allele</th>
<th>Risk allele frequency in controls</th>
<th>P</th>
<th>OR per copy or for heterozygote (95% CI)</th>
<th>Platform manufacturer and SNPs a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sporadic amyotrophic lateral sclerosis (121)</td>
<td>276 cases 271 con</td>
<td>10q26.13</td>
<td>Intergenic</td>
<td>rs4363506-C</td>
<td>7 × 10⁻²</td>
<td>1.90 (1.50–2.40)</td>
<td>Illumina: 549,062</td>
<td></td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis (122)</td>
<td>737 cases 721 con</td>
<td>7q36.2</td>
<td>DPP6</td>
<td>rs10260404-C</td>
<td>5 × 10⁻⁴</td>
<td>1.30 (1.18–1.43)</td>
<td>Illumina: 311,946</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis (97)</td>
<td>931 trios 2,431 con</td>
<td>10p15.1</td>
<td>IL2RA</td>
<td>rs12722489-C</td>
<td>3 × 10⁻⁴</td>
<td>1.25 (1.11–1.36)</td>
<td>Affymetrix: 334,923</td>
<td></td>
</tr>
<tr>
<td>Restless legs syndrome (123)</td>
<td>306 cases 15,664 con</td>
<td>2p12.1</td>
<td>BTBD9</td>
<td>rs3928309-A</td>
<td>1 × 10⁻⁷</td>
<td>1.90 (1.50–2.20)</td>
<td>Illumina: 306,937</td>
<td></td>
</tr>
<tr>
<td>Restless legs syndrome (124)</td>
<td>401 cases 1,644 con</td>
<td>6p21.2</td>
<td>BTBD9</td>
<td>rs2304748-G</td>
<td>3 × 10⁻²</td>
<td>1.74 (1.57–1.92)</td>
<td>Affymetrix: 236,758</td>
<td></td>
</tr>
<tr>
<td>APOE*4 carriers with late-onset Alzheimer disease (125)</td>
<td>446 cases 290 con</td>
<td>11q14.1</td>
<td>GAB2</td>
<td>rs2373115-G</td>
<td>1 × 10⁻¹⁵</td>
<td>4.06 (2.81–14.69)</td>
<td>Illumina: 312,316</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia (126)</td>
<td>178 cases 144 con</td>
<td>Xp22.33/Yp11.32</td>
<td>CSF2RA</td>
<td>rs4129148-C</td>
<td>4 × 10⁻²</td>
<td>3.23 (2.04–5.15)</td>
<td>Affymetrix: 439,511</td>
<td></td>
</tr>
<tr>
<td>Bipolar disorder (127)</td>
<td>461 cases 563 con</td>
<td>13q14.11</td>
<td>DGKH</td>
<td>rs1012053-A</td>
<td>2 × 10⁻⁴</td>
<td>1.59 (1.35–1.87)</td>
<td>Illumina: 555,235</td>
<td></td>
</tr>
<tr>
<td>Bipolar disorder (35)</td>
<td>1,868 cases 2,938 con</td>
<td>16p12.1</td>
<td>PALB2, NDUFAB1, DCTN5</td>
<td>rs420259-A</td>
<td>6 × 10⁻⁴</td>
<td>2.08 (1.60–2.71)</td>
<td>Affymetrix: 469,557</td>
<td></td>
</tr>
</tbody>
</table>

Studies, statistics, and platforms are reported as described in Table 2. a, risk allele not indicated. b, Number that passed quality control. c, Homozygote. d, Pooled genotyping.
Key lessons from HapMap applications to common disease

Important conclusions from these preliminary successes include the relatively modest effect sizes observed for genetic variants associated with common diseases and the consequent need for very large sample sizes to detect them. Sample sizes for detecting and confirming variants related to diabetes, obesity, and breast cancer have been in the many tens of thousands (61, 93, 94). These contrast sharply with the success of early AMD studies, in which only 100–200 cases were needed, but the identified variants conferred a much greater risk of disease (63, 95). Estimates of residual heritability after accounting for the variants found in this first round of analyses suggest that numerous other variants of modest effect, undiscovered structural variants, or less common variants of large effect remain to be found for most of these diseases (61, 93, 94). It is important to note that the most robust findings, those that survive multiple rounds of replication in an initial study and are subsequently replicated in other studies, are often not the most statistically significant associations in the initial GWA scan, and may not even be in the top few hundred associations (70, 97).

Another important lesson from these studies has been that variants in noncoding regions — rather than nonsynonymous coding SNPs, which code for different amino acids in the resulting protein — are likely to be causative in most instances (35). That regulation of the protein products, rather than differences in the structure or function of the protein, may be most important for disease risk was suspected before the advent of the HapMap and GWA studies, but the relative importance of each to disease risk was unknown, demonstrating the value of an agnostic approach to genome-wide interrogation (39, 98).

Continued evolution of HapMap-based research

Several studies have shown that tag SNPs chosen on the basis of the data from the four populations included in HapMap phases I and II apply well to other populations (2, 29). Still, to allow better choice of tag SNPs and more detailed analyses for various populations, additional samples were collected from the same four initial HapMap populations and from seven additional populations.
Table 9
GWA studies in various traits

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>Sample size</th>
<th>Region</th>
<th>Gene</th>
<th>Strongest SNP-risk allele</th>
<th>Risk allele frequency in controls</th>
<th>P</th>
<th>OR per copy or for heterozygote (95% CI)</th>
<th>Platform manufacturer and SNPs⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body mass index (93)</td>
<td>10,657 adults, 19,424 adults, 10,172 children</td>
<td>16q12.2</td>
<td>FTO</td>
<td>rs9999609-A</td>
<td>0.39</td>
<td>2 × 10⁻²⁰</td>
<td>0.36 (NR)&lt;sup&gt;³&lt;/sup&gt; ~0.4 (NR)&lt;sup&gt;²&lt;/sup&gt;</td>
<td>Affymetrix: 490,032</td>
</tr>
<tr>
<td>Height (132)</td>
<td>4,921 studied, 29,098 studied&lt;sup&gt;⁶&lt;/sup&gt;</td>
<td>12q14.3</td>
<td>RMG42</td>
<td>rs1042725-C</td>
<td>0.51</td>
<td>6 × 10⁻¹⁶</td>
<td>0.4 (NR)&lt;sup&gt;⁶&lt;/sup&gt;</td>
<td>Affymetrix: 364,301</td>
</tr>
<tr>
<td>Height (133)</td>
<td>6,669 studied, 28,801 studied</td>
<td>20q11.22</td>
<td>BFCZB</td>
<td>rs6606369-C</td>
<td>0.44</td>
<td>2 × 10⁻¹⁶</td>
<td>0.44 (NR)&lt;sup&gt;⁷&lt;/sup&gt;</td>
<td>Illumina and Affymetrix&lt;sup&gt;⁶&lt;/sup&gt;</td>
</tr>
<tr>
<td>Skin pigmentation by reflectance spectroscopy (154)</td>
<td>363 maxL&lt;sup&gt;⁺&lt;/sup&gt; &lt;56⁴</td>
<td>116 low maxL&lt;sup&gt;⁺&lt;/sup&gt;</td>
<td>SLC24A5</td>
<td>rs1834640-G</td>
<td>0.30</td>
<td>1 × 10⁻⁵</td>
<td>12.5 (8.33–20.0)</td>
<td>Perlegen: 1,502,205&lt;sup&gt;⁵&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>374 maxL&lt;sup&gt;⁺&lt;/sup&gt; &gt;63⁴</td>
<td>116 high maxL&lt;sup&gt;⁺&lt;/sup&gt;</td>
<td>TYR, SLC45A2</td>
<td>rs1042602-C, rs16891982-C</td>
<td>0.90</td>
<td>4 × 10⁻¹⁰</td>
<td>4.36 (2.64–7.20)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs6606369-C</td>
<td>0.97</td>
<td>3 × 10⁻¹¹</td>
<td>4.86 (2.88–8.21)</td>
</tr>
<tr>
<td>Freckles (135)</td>
<td>2,986 studied, 3,932 studied</td>
<td>6p25.3</td>
<td>between SECSL1 and IRF4</td>
<td>rs1540711-A</td>
<td>0.50</td>
<td>2 × 10⁻⁸</td>
<td>1.40 (1.26–1.57)</td>
<td>Illumina: 317,511</td>
</tr>
<tr>
<td>Blond vs. brown hair (135)</td>
<td>2,986 studied, 3,932 studied</td>
<td>12q21.33</td>
<td>KITLG</td>
<td>rs12821256-C, rs4904868+C</td>
<td>0.15</td>
<td>2 × 10⁻¹⁴</td>
<td>2.32 (1.86–2.92)</td>
<td>Illumina: 317,511</td>
</tr>
<tr>
<td>Blue vs. green eyes (135)</td>
<td>2,986 studied, 3,932 studied</td>
<td>14q32.12</td>
<td>SLC24A4</td>
<td>rs4904868+C, rs2402130-A</td>
<td>0.60</td>
<td>9 × 10⁻⁹</td>
<td>2.56 (2.12–3.09)</td>
<td>Illumina: 317,511</td>
</tr>
<tr>
<td>F cell distribution (136)</td>
<td>179 adults&lt;sup&gt;¹&lt;/sup&gt;, 90 adults</td>
<td>2p16.1</td>
<td>BCL11A</td>
<td>rs1427407-?</td>
<td>0.14</td>
<td>6 × 10⁻¹¹</td>
<td>13.1% (NR)&lt;sup&gt;⁶&lt;/sup&gt;</td>
<td>Illumina: 317,511</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6q23.3</td>
<td>Intergenic</td>
<td>rs9399137-?</td>
<td>0.23</td>
<td>3 × 10⁻⁶</td>
<td>15.8% (NR)&lt;sup&gt;⁶&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Serum uric acid levels (137)</td>
<td>4,305 Sardinian, 1,301 Tuscan</td>
<td>4p16.1</td>
<td>GLUT9</td>
<td>rs6855911-A</td>
<td>0.74</td>
<td>2 × 10⁻¹⁶</td>
<td>0.32 (NR)&lt;sup&gt;⁷&lt;/sup&gt;</td>
<td>Illumina: 362,129</td>
</tr>
<tr>
<td>Serum urate (117)</td>
<td>1,955 hypertensive individ</td>
<td>2p33 individ</td>
<td>519 families; 1,461 twins&lt;sup&gt;⁸&lt;/sup&gt;</td>
<td>SLC24A9</td>
<td>0.79</td>
<td>2 × 10⁻¹⁵</td>
<td>0.024 (0.018–0.030)&lt;sup&gt;⁹&lt;/sup&gt;</td>
<td>Affymetrix: 400,496</td>
</tr>
<tr>
<td>Recombination rate (138)&lt;sup&gt;⁶&lt;/sup&gt;</td>
<td>1,887 men, 1,702 women</td>
<td>4p16.3</td>
<td>RNF212</td>
<td>rs3796619-T</td>
<td>0.33&lt;sup&gt;³&lt;/sup&gt;</td>
<td>3 × 10⁻²⁴</td>
<td>70.7 (84.3–57.1)&lt;sup&gt;¹⁰&lt;/sup&gt;</td>
<td>Illumina: 309,241</td>
</tr>
</tbody>
</table>
| Study characteristics        | Studies, statistics, and platforms are reported as described in Table 2. individ, individuals; maxL<sup>⁺</sup>, maximum reflectance; ?, risk allele not indicated. | | | | | | | |<sup>¹</sup>Number that passed quality control, <sup>²</sup>kg/m² per copy in adults. <sup>³</sup>kg/m² per copy in children. <sup>⁴</sup>Of which 19,064 were adults. <sup>⁵</sup>cm per copy in adult height. <sup>⁶</sup>Height increase in cm. <sup>⁷</sup>About 2,261,000 imputed. <sup>⁸</sup>20% tails of distribution. <sup>⁹</sup>Pooled genotyping. <sup>¹⁰</sup>Percent of variance explained. <sup>¹¹</sup>Increase in mg/dl. <sup>¹²</sup>One twin selected randomly. <sup>¹³</sup>Increase in mmol/l. <sup>¹⁴</sup>Values for men shown in the top row; values for women shown in the bottom row. <sup>¹⁵</sup>Men and women combined. <sup>¹⁶</sup>Decrease in cM. <sup>¹⁷</sup>Increase in cM.

Luhya inWEBuye, Kenya; Maasai in Kinyawa, Kenya; Tuscan inItaly; Gujarati Indians in Houston, Texas; Chinese in metropolitan Denver, Colorado; Mexican ancestry in Los Angeles, California; and African ancestry in the southwestern United States. These 1,301 extended HapMap samples, now available from the Coriell Institute, have been genotyped on the Affymetrix 6.0 platform and the Illumina 1 million SNP chip, and genome-wide sequencing of these samples will be soon. As with the initial HapMap samples, these will become a standard research resource for many additional studies and will be particularly useful in providing additional information on rare variants.

The GWA studies described above have shown substantial early promise, and new applications of genome-wide technology to well-characterized population samples are continuing. However, important limitations of GWA studies should be kept in mind, including their lack of statistical power for identifying associations with rare sequence variants, because these are poorly represented on current genotyping platforms, and the need for very large numbers of samples (see Limitations of HapMap-based GWA studies). The benefits of collaboration across multiple GWA studies for replicating initial associations and developing common methods have been amply demonstrated by the pioneering efforts of the WTCCC study of seven complex diseases and common controls (35). Several other collaborative programs are currently in the pipeline (Table 10).

Data from many of these GWA studies are released to the scientific community through the Database of Genotype and Phenotype (dbGaP) of the National Center for Biotechnology Information (99). Study descriptions, protocols, data summaries, and other group-level data are available in the open-access portion of the dbGaP website, while individual-level data are provided through a controlled-access process consistent with the informed consent provided by study participants, as described in the recently finalized policy for sharing of data obtained in NIH-supported or -conducted GWA studies (100). This commitment to rapid data release builds on the now-well-established ethic in genomic community research projects of maximizing data access.
Expression, as recently demonstrated for asthma-associated variants. Fine-mapping of candidate regions with specific genotype-phenotype association findings to the search of human genetic variation and limit follow-up sequencing of targeted cell lines. Successful GWA studies are the most visible and exciting outcome of HapMap to date, with the large number of robust and highly replicated genetic associations with common diseases providing novel and unexpected insights into the pathophysiology of disease. The HapMap has also been invaluable in developing genotyping and analytic methods, expanding our understanding of evolutionary pressures and natural selection, defining genetic relationships across populations, and providing samples for validation of variation detection methods and standardization of laboratory processes. Application of these association findings is expected to produce new advances in the prevention and treatment of common diseases.

**Note added in proof**. Updated information on GWA studies and SNP associations is available online (139).

**Acknowledgments**

The authors express their sincere appreciation to Mia Diggs, Lucia Hindorff, Heather Junkins, Darryl Leja, and Lisa McNeil for assistance in preparation of the manuscript.

Address correspondence to: Teri Manolio, National Human Genome Research Institute, 31 Center Drive, Room 4B-09, Bethesda, Maryland 20892-2154, USA. Phone: (301) 402-2915; Fax: (301) 402-0837; E-mail: manolio@nhgri.gov.

---

### Table 10

<table>
<thead>
<tr>
<th>GAIN</th>
<th>GEI</th>
<th>STAMPEED</th>
<th>CGEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>Type 2 diabetes</td>
<td>Early-onset myocardial infarction</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>Maternal metabolism and birth weight</td>
<td>Asthma</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>Bipolar I disorder</td>
<td>Preterm birth</td>
<td>Platelet phenotypes</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Oral clefts</td>
<td>CHD and other heart, lung and blood disorders</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>Type I diabetic nephropathy</td>
<td>Dental caries</td>
<td>Childhood respiratory outcomes</td>
<td>Bladder cancer</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>Coronary disease</td>
<td>Hematopoietic cell transplant outcome</td>
<td>Renal cancer</td>
</tr>
<tr>
<td></td>
<td>Lung cancer</td>
<td>Arteriosclerosis in hypertensives</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Addiction</td>
<td>Asthma and lung function</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cardiovascular risk factors</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Atherosclerosis pathway genes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CV events</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Early coronary artery disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phenotypic variability in sickle cell anemia</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Centenarians</td>
<td></td>
</tr>
</tbody>
</table>

2. International HapMap Consortium. 2007. A sec-
3. International HapMap Consortium. 2003. The Inter-
4. International Human Genome Sequencing Con-
5. International Human Genome Sequencing Con-
6. Hirschhorn, J.N., and Daly, M.J. 2005. Genome-
wide association studies for common diseases and
enment, health, and disease: facing up to com-
15. van den Oord, E.J., and Neale, B.M. 2004. Will haplo-
20. Christensen, K., and Murray, J.C. 2007. What Genome-wide association studies can do for medi-
26. Corell Institute for Medical Research, et al. Inter-
edu/Sections/Collelections/NHGRI/hapmap.
asp?pg=266&col=Genome Sequencing.
www.hapmap.org.
28. NCBI Reference Assembly. Single nucleotide poly-
32. Steemers, F.J., et al. 2006. Whole-genome genoty-
ping with the single-base extension assay. Nat. Meth.
ods. 3:31–33.
35. Welcome Trust Case Control Consortium. 2007. Large-scale association study of 1,400,000 cases of seven common diseases and 3,000 shared controls. Nature. 447:661–678.
ferential genotyping error rate on the type I error probability of case-control studies. Hum. Hered. 61:55–64.
38. Manolio, T.A., et al. 2007. New models of collabo-
43. Sidney, S., et al. 2004. Venous thromboembolic disease in users of low-estrogen combinedestro-
46. Komura, D., et al. 2006. Genome-wide detection of human copy number variations using high-
density DNA oligonucleotide arrays. Genome Res. 16:1575–1584.
48. Hinds, D.A., Kloeck, A.P., Jen, M., Chen, X., and Fra-
50. Stranger, B.E., et al. 2007. Relative impact of nucle-


