Published in final edited form as:
N Engl J Med. 2009 April 23; 360(17): 1759–1768. doi:10.1056/NEJMra0808700.

Genomewide Association Studies and Human Disease

John Hardy, Ph.D. and **Andrew Singleton, Ph.D.**
Institute of Neurology, University College London, London (J.H.); and the Laboratory of Neurogenetics, Bethesda, MD (A.S.).

For 20 years, genetic linkage combined with positional cloning has offered a rational and increasingly straightforward route to finding gene mutations that lead to monogenic disease, such as cystic fibrosis and Huntington's disease (see the Glossary). With a few important exceptions, these searches have led to mutations that alter the amino acid sequence of a protein and that enormously increase the risk of disease.

During the past few years, genomewide association studies have identified a large number of robust associations between specific chromosomal loci and complex human disease, such as type 2 diabetes and rheumatoid arthritis[1] (Fig. 1). This approach relies on the foundation of data produced by the International Human HapMap Project and the fact that genetic variance at one locus can predict with high probability genetic variance at an adjacent locus, typically over distances of 30,000 base pairs of DNA[2] in the human genome, which contains about 3×10^9 base pairs. This haplotypic structure of the human genome means that it is possible to survey the genome for common variability associated with the risk of disease simply by genotyping approximately 500,000 judiciously chosen markers in the genome of several thousand case subjects and control subjects.[3] Consequently, it is now routine to identify common, low-risk variants (i.e., those that are present in more than 5% of the population) that confer a small risk of disease, typically with odds ratios of 1.2 to 5.0.[4]

The platform that is used to genotype markers in genomewide association studies and related approaches has uncovered a startling degree of structural genomic variation. Although such variants were known to be causes of rare monogenic disorders,[5,6] the extent of structural genomic variation among persons was largely unanticipated, and there is increasing interest in understanding how such variants may confer a risk of common diseases.[7,8]

The initial contention surrounding the viability of genomewide association studies has largely subsided. However, discussion has centered on evaluating how far such studies will take us in understanding the risks and causes of disease — and thus the time and resources that should be invested in genotyping more case subjects with any one disease to garner what many see as diminishing genetic returns. These issues are discussed in three Perspective articles in this issue of the *Journal*.[9–11] Nonetheless, the current phase of rapid discovery is a remarkable change that ends a long period of frustration, when the investigation of the genetic causes of complex diseases could boast few successes. The data from genomewide association studies and emerging sequencing techniques offer a route to the dissection of genetic causes of human disease (Table 1).[12–21] Here we describe this route and some of its challenges.

Address reprint requests to Dr. Hardy at the Institute of Neurology, University College London, Queen Sq., London WC1N 3BG, United Kingdom, or at jhardy@ion.ucl.ac.uk.

No potential conflict of interest relevant to this article was reported.

Genomewide association studies identify loci and not genes per se and cannot easily identify loci at which there are many rare risk alleles in any given population.[22] Rather, this approach is designed to find loci that fit the common disease–common variant hypothesis of human disease[23,24] (Table 2). Refinement of susceptibility loci and the identification of causal variants may be achieved through fine mapping (see the Glossary).

One observation that has taken many observers by surprise is that most loci that have been discovered through genomewide association analysis do not map to amino acid changes in proteins. Indeed, many of the loci do not even map to recognizable protein open reading frames but rather may act in the RNA world by altering either transcriptional or translational efficiency. They are thus predicted to affect gene expression. Effects on expression may be quite varied and include temporal and spatial effects on gene expression that may be broadly characterized as those that alter transcript levels in a constitutive manner, those that modulate transcript expression in response to stimuli, and those that affect splicing.

Therefore, there are two clear and immediate tasks: to develop an understanding of the genetics of gene expression and to identify disease-linked variants that are too rare to be picked up by association methods and yet have risk alleles of sufficient "strength" to allow detection with the use of linkage strategies (see the Glossary for descriptions of genetic association and genetic linkage). Meeting these challenges will serve efforts to better understand environmental influences on the causes of disease and may facilitate a systems-based understanding of disease, in which we come to understand the full, molecular network that is perturbed in disease.

THE GENETICS OF GENE EXPRESSION

It is perhaps not surprising that many variants conferring a low risk of a complex disease effect a change of quantity in gene expression, because many of these diseases can be thought of as quantitative traits themselves, with disease diagnosis being made when a clinical threshold is surpassed (as is the case with hypertension and Alzheimer's disease). This frequent observation from genomewide association studies was presaged by the observation that genetic variability in the insulin-gene promoter is associated with an increased risk of type 1 diabetes.[25]

Genetic variability in gene expression may occur at many stages: transcription, messenger RNA (mRNA) stability, and splicing or translation efficiency. In each of these instances, the underlying variability would be expected to occur in different DNA elements that may have element-specific sequence motifs and should be distinguishable by their effects on cellular RNA species. Tissue-specific genomewide analyses of gene expression offer a starting point for parsing the various possibilities.[26] This approach has been particularly helpful in understanding the effect of susceptibility variants in immune-mediated disease, such as asthma, because the lymphocyte (which is pivotal to the pathological analysis of such disease) is easily accessible,[27] although human fat tissue[28] and human brain tissue[29] obtained at autopsy have also been used. Three examples of this approach illustrate its power: a haplotype associated with asthma also shows an association with lymphoblastoid expression of the proteins ORMDL3 and GSDML,[27,30] genetic variants that are associated with obesity are also associated with the expression of their cognate mRNAs in adipose tissue,[28] and a variant of *MAPT* (encoding tau) that is associated with progressive supranuclear palsy is also associated with *MAPT* mRNA expression.[29,31,32]

Genetic variability can also result in differences in translational efficiency through changes in the mRNA sequence or in the level or sequence of regulatory RNAs.[33] Both modes can be queried through high-throughput transcriptomic sequencing, which enumerates the number of times that any individual RNA species is present in preparations from that tissue.

Unlike chip technologies, such sequencing does not depend on the relevant RNA being represented on an array; it can also provide a survey of all RNA species, not just mRNA. The eventual goal of the recently announced Genotype-Tissue Expression (GTEx) project is to create a whole-body map of haplotypic expression so that any risk haplotype for any disease can be easily checked for its effect on genomewide and tissuewide RNA expression (Fig. 2).

RARE HIGH-RISK CODING VARIANTS

The present array techniques do not enable hypothesis-free means of identifying high-risk variants with one exception: that of structural genomic variation.[5] The process for reaching the goal of systematic identification of rare high-risk variants is clear, both from candidate-gene studies and from emerging techniques of high-throughput sequencing, which will soon permit the routine and complete sequencing of the human genome.[19] Existing but imperfect intermediate techniques toward that goal are transcriptome sequencing and exome sequencing.[34] The latter uses array techniques to pull exonic DNA from genomic DNA, which is then sequenced to give full representation of the coding genome. All coding polymorphisms in a subject will therefore be identifiable. Candidate-gene studies have already suggested the power of this type of approach. For example, Cohen and colleagues[35] sequenced several genes encoding cholesterol-metabolizing proteins in patients with low plasma levels of high-density lipoprotein (HDL) cholesterol and found that rare variants were more common in case subjects who had low levels of HDL cholesterol than in control subjects. This example, in which a limited number of candidate genes were sequenced in a large number of subjects (an approach called deep resequencing), shows the power of testing a specific hypothesis. One can imagine testing a hypothesis with the use of genomewide data, in which case the usual criterion of having sufficient power to overcome the limitation of multiple testing and the low prevalence of rare variants would apply.

THE POWER OF THE PATHWAY

There is increasing focus on the idea of networks that are composed of genes and proteins. Although the complex interplay of macromolecules is a certainty, there is benefit in taking a reductionist approach when envisioning common molecular routes toward disease. Indeed, for many diseases, different genetic loci must impinge on a common pathway to pathogenesis. This means that as a risk allele at a genetic locus comes into focus, it provides clues to other risk loci and mechanisms by which variability at the same locus or on the same pathway can contribute to disease.

This point is well illustrated in the case of coronary artery disease, in which cholesterol metabolism has long been thought to be a pathogenic pathway to disease.[36] Although few pathogenic pathways are as well delineated as cholesterol metabolism, huge amounts of data pertaining to protein and pathway interactions have been obtained with the use of yeast, roundworms (*Caenorhabditis elegans*), and fruit flies (*Drosophila melanogaster*). Studies of these creatures have informed and continue to inform human genetic studies. For example, two of the genes that are involved in recessive parkinsonism, *PARK2* and *PINK1*, have recently been shown to be involved in the same mitochondrial pathway through work in drosophila.[37,38]

MOVING FROM DICHOTOMOUS TO GRADED GENETIC RISK

The vast majority of success in defining genetic risk in disease has been a result of traditional gene-hunting efforts to find mutations that underlie monogenic diseases. In this approach, our understanding of disease revolves around the idea of normal and abnormal variation, with the latter greatly increasing the risk of disease. In considering the genetics of complex disease and particularly the role of common variants that affect expression, a more nuanced perspective is useful. The difference in genetic effect between rare high-risk variants and common low-risk variants is quantitative and not qualitative, as illustrated in Parkinson's disease: point mutations within the α-synuclein gene[39] and genomic multiplications containing this gene[6] lead to monogenic disease, whereas a common haplotype of the α-synuclein gene moderates the risk of sporadic disease.[40]

Parsimony would suggest that there is probably a graded influence of genetic variation in gene expression because for any gene many elements contribute to the control of its expression, and genetic variability in any one of such genes is likely to result in a change in expression. In this model, at any locus there are multiple variants, which can exist across a single haplotype block or in multiple haplotype blocks proximal to the affected transcript. Thus, there is no single haplotype for disease risk and no single protective haplotype but, rather, a collection of haplotypes that confer a graded risk of disease. The variant with the highest population attributable risk (a combination of allele frequency and relative risk) is likely to be the first at the locus to be detected as a risk factor, and further dissection of the same locus will yield other risk alleles of smaller effect. Although such dissection is proving to be a tough task, there are already examples of success. After the identification of a risk allele for macular degeneration, a polymorphism that causes the substitution of tyrosine for histidine at position 402 in complement factor H (CFH),[41–43] several additional and independent risk variants, including noncoding alterations, have been detected in and around the *CFH* gene, and none of these variants in isolation account for all the risk attributed to this locus.[44]

CENTRALIZED RESOURCES

The evolution of genetic analysis of traits has revealed the power of testing markers across the whole genome to identify novel factors involved in disease and has shown that large samples are required to determine true biologic associations. This, in turn, has underscored the desirability of accessible resources and data, such as the human genome sequence and the haplotype map from the HapMap project, for these and future techniques. The generation of population-control data for genomewide association studies by the Wellcome Trust and other groups, while initially expensive, has been useful to many independent research groups and proved to be an economical approach. A similarly useful resource will be the 1000 Genomes Project, a large international effort that aims to identify all single-nucleotide polymorphisms (SNPs) with a prevalence of 1% or more in the human genome. This effort will focus on resequencing samples from the initial and extended HapMap populations from around the world. Even with new sequencing techniques, this is a monumental effort. However, it is still likely to be only a first installment. To reliably determine the pathogenicity of rare variants as they are identified, we will probably need reference sequences from tens or hundreds of thousands of subjects, coupled with a better understanding of the biologic effects of SNPs. Housing and making accessible such data will be a considerable challenge, especially when one considers that the data will include variants pertaining to both SNPs and structural genomic variability.

ENVIRONMENTAL EFFECT

To state that most complex diseases are caused by an interaction between genome and environment is a cliché. Such interactions, while likely, have for the most part not been demonstrated, and we should be cautious about universally subscribing to this belief without evidence. Since the quantification of environmental influences is notoriously difficult, it is likely that such a demonstration will remain a formidable challenge. At least, the definition of gene-based pathways for disease will provide a framework for the systematic investigation of exogenous influences. This is one of the goals of the recently announced Genes, Environment, and Health Initiative of the National Institutes of Health. There is increasing interest in genomewide assessments of epigenetic modification brought about by a greater understanding of the ubiquitous nature of such modifications and the availability of genome-scale sequences, which makes such investigation tenable from a practical perspective. It is hoped that greater understanding of the epigenome, particularly in the context of genetic variation and gene expression, will offer a direct and quantifiable link between putative environmental influences and pathways relevant to pathogenesis.

The jigsaw puzzle of understanding the causes of disease lies before us: we now have the edges and corners in place. The identification of monogenic disease loci and the common genetic variability that contributes to disease risk is now a tractable problem. The techniques that are necessary for genomewide identification of such rare variants that contributes to disease risk are quickly being refined. There is an enormous amount of filling in to do (including the dissection of the interactions among different genes), and there are formidable challenges, which increased bioinformatic data will help to address. Undoubtedly, there will be surprises, but the boundaries of the task ahead have already been drawn.

Acknowledgments

Supported in part by the Intramural Research Program of the National Institute on Aging of the National Institutes of Health (project number, 1-Z01-AG000949-02) and by the Medical Research Council.

We thank Dr. Katrina Gwinn for her constructive comments on the manuscript.

Glossary

Common disease–common variant hypothesis
A theory that many common diseases are caused by common alleles that individually have little effect but in concert confer a high risk.

Complex disease
A disorder in which the cause is considered to be a combination of genetic effects and environmental influences.

Deep resequencing
A technique for sequencing a gene in several thousand subjects, typically with the use of high-throughput sequencing.

Epigenetics
The study of heritable changes to DNA structure that do not alter the underlying sequence; well-known examples are DNA methylation and histone modification.

Exome
All the expressed messenger RNA sequences in any tissue.

Fine mapping
The precise mapping of a locus after it has been identified by genetic linkage or association. The initial localization is determined within megabases of DNA in genetic linkage studies and within tens of kilobases in genetic association studies. In genetic association studies, fine mapping implies finding all the variants at the locus and trying to determine which changes may be related to pathogenesis with the use of statistical, functional, or bioinformatic methods.

Genes, Environment, and Health Initiative (GEI)
A project funded by the National Institutes of Health to determine the relationships between genetic factors and disease. A proportion of the funding supports research of systematic ways to quantify environmental exposures.

Genetic association A relationship that is defined by the nonrandom occurrence of a genetic marker with a trait, which suggests an association between the genetic marker (or a marker close to it) and disease pathogenesis.

Genetic linkage A relationship that is defined by the coinheritance of a genetic marker with disease in a family with multiple disease-affected members.

Genomewide association study A test of the association between markers, called single-nucleotide polymorphisms (SNPs), across the genome and disease, usually involving 300,000 or more markers that are reasonably polymorphic and are spread across the genome fairly evenly. This approach is hypothesis free (i.e., there is no existing hypothesis about a particular gene or locus but the null hypothesis that no detectable association exists).

Genotype-Tissue Expression (GTEx) A project funded by the National Institutes of Health that aims to study and map the relationship between human gene expression and genetic variation. The project, which is in a pilot phase, will analyze dense genotyping and expression data collected from multiple human tissues and will correlate genetic variation and gene expression, thus producing a list of genetic regions associated with expression of specific transcripts.

Haplotype A series of polymorphisms that are close together in the genome. The distribution of alleles at each polymorphic site is nonrandom: the base at one position predicts with some accuracy the base at the adjacent position. Persons sharing a haplotype are related, often very distantly. Haplotypes in Europeans are generally of the order of tens of kilobases long; older populations, such as those of West Africa, tend to have shorter haplotypes, since a longer period of evolutionary time means more meiotic events and a greater chance of population admixture, both of which result in shorter haplotypes.

Haplotypic structure The general underlying segmentation of the genome. As a result of recombination events occurring throughout the history of a population, contiguous segments of DNA are shared by persons within a population. Chromosomes can thus be broken down into contiguous segments, containing haplotypes common to members of particular populations.

HapMap A catalogue of common genetic variation in humans compiled by an international partnership of scientists and funding agencies. Its goal was to determine the identity and length of haplotypes across the genome in different human populations. Stage 1 of the process, which was completed in 2005, yielded haplotype maps from SNPs present in at least 5% of chromosomes of each of three groups defined by ancestry: Yoruban, Northern and Western European, and Asian (Chinese and Japanese). Stage 2 involves determining haplotypes made up of SNPs with a lower prevalence (at least 1% of chromosomes) in these three groups and also in the Luhya and Maasai from Kenya, Toscani from Italy, Gujarati Indians, persons of Mexican ancestry, and persons of mixed African ancestry.

High-throughput sequencing Several new techniques that since 2005 have increased the speed and decreased the cost of DNA sequencing by two orders of magnitude.

Human Genome Project A coordinated international effort that led to the consensus sequence of the human genome.

Linkage disequilibrium The nonrandom association of genetic markers; a set of markers in a haplotype are said to be in linkage disequilibrium.

Monogenic disease A disorder caused by a mutation in a single gene (also called a mendelian disease).

Positional cloning An approach for determining the position of a gene that, when mutated, causes monogenic disease. In families with disease, genetic markers from every chromosome are typed in both affected and unaffected members. Markers that are coinherited with disease indicate the chromosomal position of the genetic defect, and then genes at that position are sequenced to find the pathogenic mutation, which in turn indicates the causative gene.

Sequence motif DNA sequences whose functions can be inferred because they are similar to sequences whose function has been biologically determined.

Structural genomic variation Variation within the genome that results from deletion or duplication (both referred to as copy-number variation) or from inversion of genomic segments. Although common large variants (of more than one kilobase) exist, the majority of such variants are rare.

Transcriptome A description of all DNA that is transcribed into RNA (messenger RNA, transfer RNA, microRNA, and other RNA species). The prevalence of a specific RNA sequence in a particular tissue may be proportionate to the relevance of that RNA species in the tissue.

1000 Genomes Project A whole-genome resequencing of 1000 subjects from the original and extended HapMap populations, which was started in 2008, with funding from an international research consortium.

REFERENCES

1. Hunter DJ, Kraft P. Drinking from the fire hose — statistical issues in genome-wide association studies. N Engl J Med. 2007; 357:436–439. [PubMed: 17634446]
2. International HapMap Consortium. A haplomap type of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]
3. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet. 2005; 6:109–118. [PubMed: 15716907]
4. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008; 118:1590–1605. [PubMed: 18451988]
5. Lupski JR. Structural variation in the human genome. N Engl J Med. 2007; 356:1169–1171. [PubMed: 17360997]
6. Singleton AB, Farrer M, Johnson J, et al. α-Synuclein locus triplication causes Parkinson's disease. Science. 2003; 302:841. [PubMed: 14593171]
7. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. Science. 2007; 316:445–449. [PubMed: 17363630]
8. Need AC, Ge D, Weale ME, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet. 2009; 5(2) e1000373.

9. Kraft P, Hunter DJ. Genetic risk prediction — are we there yet? N Engl J Med. 2009; 360:1701–1703. [PubMed: 19369656]

10. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009; 360:1696–1698. [PubMed: 19369660]

11. Hirschhorn JN. Genomewide association studies — illuminating biologic pathways. N Engl J Med. 2009; 360:1699–1701. [PubMed: 19369661]

12. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [Erratum, Nature 2001;412:565.]. [PubMed: 11237011]

13. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science. 2001; 291:1304–1351. [Erratum, Science 2001;292:1838.]. [PubMed: 11181995]

14. Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A. 2002; 99:4465–4470. [PubMed: 11904358]

15. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007; 445:881–885. [PubMed: 17293876]

16. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

17. Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000; 18:630–634. [Erratum, Nat Technol 2000;18:1021.]. [PubMed: 10835600]

18. Olson M. Enrichment of super-sized resequencing targets from the human genome. Nat Methods. 2007; 4:891–892. [PubMed: 17971778]

19. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–876. [PubMed: 18421352]

20. 1000 Genomes: a deep catalog of human genetic variation. at http://www.1000genomes.org/page.php.

21. NIH Roadmap for Medical Research. Bethesda, MD: National Institutes of Health; Genotype-Tissue Expression (GTEx) project. at nihroadmap.nih.gov/GTEx/.

22. Terwilliger JD, Hiekkalinna T. An utter refutation of the "Fundamental Theorem of the HapMap.". Eur J Hum Genet. 2006; 14:426–437. [PubMed: 16479260]

23. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273:1516–1517. [PubMed: 8801636]

24. Hardy J, Singleton A. The future of genetic analysis of neurological disorders. Neurobiol Dis. 2000; 7:65–69. [PubMed: 10783291]

25. Kennedy GC, German MS, Rutter WJ. The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. Nat Genet. 1995; 9:293–298. [PubMed: 7773292]

26. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005; 437:1365–1369. [PubMed: 16251966]

27. Moffatt MF, Kabesch M, Liang L, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007; 448:470–473. [PubMed: 17611496]

28. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452:423–428. [PubMed: 18344981]

29. Myers AJ, Gibbs JR, Webster JA, et al. A survey of genetic human cortical gene expression. Nat Genet. 2007; 39:1494–1499. [PubMed: 17982457]

30. Bouzigon E, Corda E, Aschard H, et al. Effect of 17q21 variants and smoking in early-onset asthma. N Engl J Med. 2008; 359:1985–1993. [PubMed: 18923164]

31. Melquist S, Craig DW, Huentelman MJ, et al. Identification of a novel risk locus for progressive supranuclear palsy by a pooled genomewide scan of 500,288 single-nucleotide polymorphisms. Am J Hum Genet. 2007; 80:769–778. [PubMed: 17357082]

32. Myers AJ, Pittman AM, Zhao AS, et al. The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. Neurobiol Dis. 2007; 25:561–570. [PubMed: 17174556]

33. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. Trends Genet. 2008; 24:489–497. [PubMed: 18778868]
34. Ng PC, Levy S, Huang J, et al. Genetic variation in an individual human exome. PLoS Genet. 2008; 4(8) e1000160.
35. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–872. [PubMed: 15297675]
36. Goldstein JL, Brown MS. Molecular medicine: the cholesterol quartet. Science. 2001; 292:1310–1312. [PubMed: 11360986]
37. Clark IE, Dodson MW, Jiang C, et al. Drosophila pink1 is required for mitochondrial function and interacts genetically with parkin. Nature. 2006; 441:1162–1166. [PubMed: 16672981]
38. Park J, Lee SB, Lee S, et al. Mitochondrial dysfunction in Drosophila PINK1 mutants is complemented by parkin. Nature. 2006; 441:1157–1161. [PubMed: 16672980]
39. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science. 1997; 276:2045–2047. [PubMed: 9197268]
40. Farrer M, Maraganore DM, Lockhart P, et al. α-Synuclein gene haplotypes are associated with Parkinson's disease. Hum Mol Genet. 2001; 10:1847–1851. [PubMed: 11532993]
41. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389. [PubMed: 15761122]
42. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science. 2005; 308:419–421. [PubMed: 15761120]
43. Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. Science. 2005; 308:421–424. [PubMed: 15761121]
44. Li M, Atmaca-Sonmez P, Othman M, et al. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. Nat Genet. 2006; 38:1049–1054. [PubMed: 16936733]

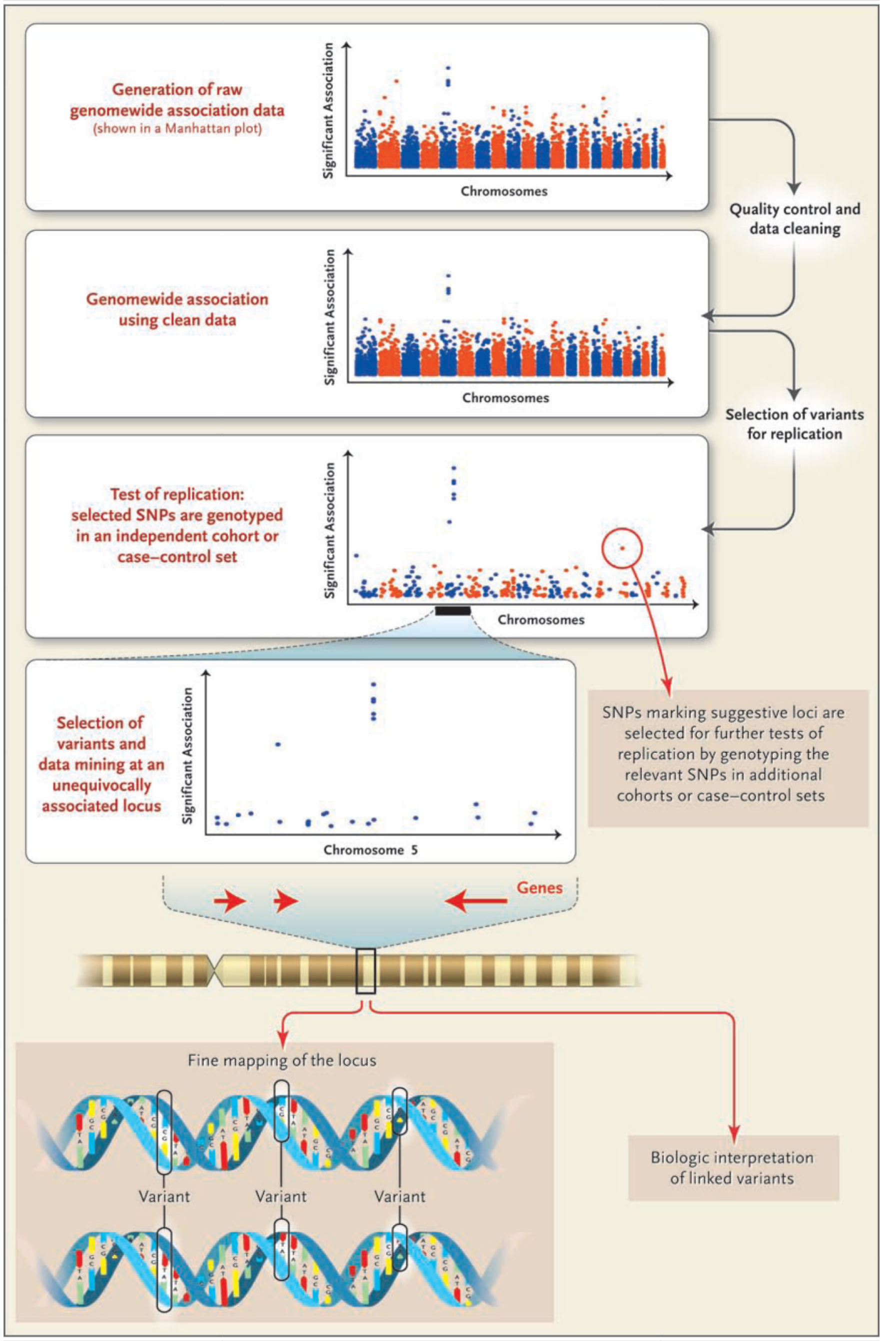
Generation of raw genomewide association data (shown in a Manhattan plot)
Significant Association
Chromosomes
Quality control and data cleaning
Genomewide association using clean data
Significant Association
Chromosomes
Selection of variants for replication
Test of replication: selected SNPs are genotyped in an independent cohort or case–control set
Significant Association
Chromosomes
Selection of variants and data mining at an unequivocally associated locus
Significant Association
Chromosome 5
SNPs marking suggestive loci are selected for further tests of replication by genotyping the relevant SNPs in additional cohorts or case–control sets
Genes
Fine mapping of the locus
Variant
Variant
Variant
Biologic interpretation of linked variants

Figure 1 (facing page). Stages of a Genomewide Association Study

Although genomewide association studies are increasingly popular, they present formidable logistical and technical challenges. The primary challenge lies in selecting a disease or a trait suitable for analysis. A successful analysis is more likely when the phenotype of interest can be sensitively and specifically diagnosed or measured. For such studies, extremely large sample series are required, involving thousands of case subjects and control subjects. This process usually mandates collaboration among groups that were previously competitors, which in itself presents a formidable challenge to success. In the first stage, single-nucleotide polymorphisms (SNPs) across the genome are genotyped, almost exclusively on chip-based products generated by one of two companies, Illumina or Affymetrix. The genotyping content of these products differs, but recent advances allow the imputation of ungenotyped SNPs from those that have been genotyped, which facilitates collaboration and comparison among groups that have used different techniques. Second, after the generation of SNP data, the data are subjected to quality control and cleaning procedures, such as ensuring that the genotyped sex (based on X and Y genotypes) matches the reported sex for individual samples, measuring how well the samples are matched as a group, and identifying individual outliers (all based on general patterns of genetic variability). This step allows the removal of samples from ethnically distant subjects and adjustment for any systematic differences between or within cohorts. Third, each SNP that survives quality control and cleaning is then tested for association with a disease or trait. Shown is a Manhattan plot, which is typically used in genomewide association studies and plots the negative log of the P value against chromosomal position. Because of the number of statistical tests that are performed, there is a high false positive rate. Therefore, depending on the study design, genomewide statistical significance is set at P values of approximately 1.0×10^{-8} or less at this stage of the analysis. The models of risk that are most typically tested are dominant, recessive, genotypic, allelic, and additive (with the additive model, which assumes that the presence of one risk allele confers an intermediate risk between having no allele and having two alleles, most frequently tested). Fourth, SNPs or loci are selected for replication in an independent sample set, ideally of the same or larger size than the sample analyzed in the genomewide association. The selection of loci may be based on statistical significance alone or a combination of statistical significance and biologic plausibility; the number of SNPs that are selected for testing may be as few as 10 or as many as 20,000, depending on the initial study design and resources available. Fifth, replication experiments lead to any combination of three results: selected loci show clear and unequivocal association with disease, show no association signal whatsoever, or show an association with disease that is not of sufficient magnitude to pass a predetermined statistical threshold. Sixth, additional genotyping is performed in independent replication cohorts to determine whether an association with a disease is genuine or not. Seventh, data mining at unequivocally associated loci reveals transcripts in and around this locus, in addition to the mapping of all known genetic variation within the region. Further fine mapping of the locus is performed by a combination of deep-resequencing methods to discover new variants and genotyping of untyped variants to determine which are most significantly associated with disease. Further analysis of the region is performed to determine the most critical variants, the pathologically relevant gene, and the likely biologic effect.

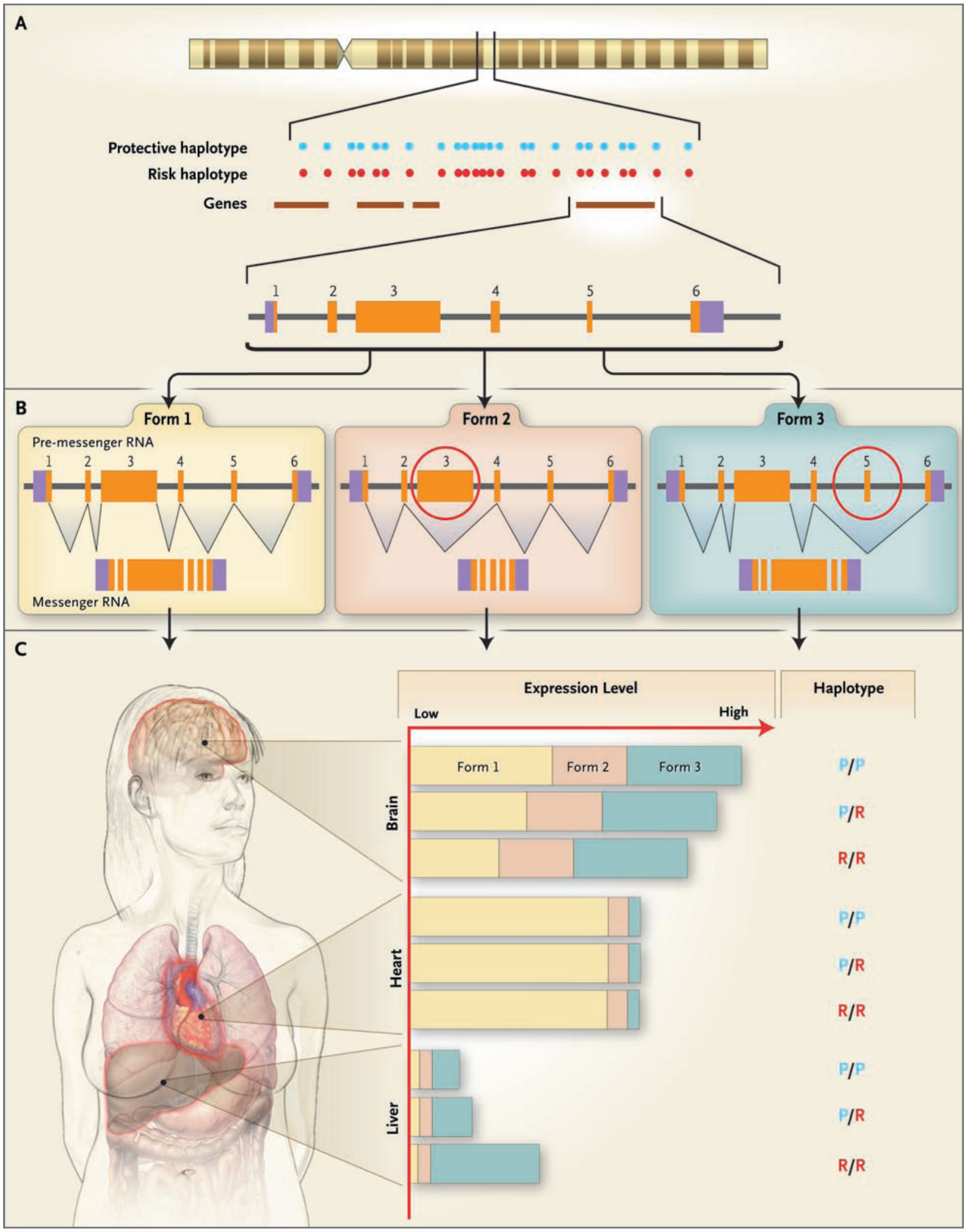

Figure 2 (facing page). Genetic Control of Gene Expression in Various Tissues

After the identification of a genetic locus associated with disease, the next step is to determine whether this variant alters the expression of transcripts within the region. In Panel A, a disease-associated region of the genome contains four genes. Although it is clear that genetic variability at a locus may affect distal genes (and even those on other chromosomes), in most instances the most proximal genes are investigated. In this case, expression of all four genes would be assessed, but one is shown for clarity. The single-nucleotide polymorphisms (SNPs) across this region form haplotypes that confer risk (red) or protection (blue) against disease. In Panel B, three splice forms are known to exist for the gene of interest: forms 1, 2, and 3, which differ according to their inclusion or removal of exons 3 and 5. In Panel C, genotyping of the risk variants in human tissues and expression analysis of the three splice forms in the same tissues allow a test of association between the genotype or haplotype and expression level. In this example, the risk haplotype is associated with decreasing levels of form 1 of the gene in the brain, has no measurable effect in the heart, and increases expression of form 3 in the liver. If no genotype-expression association were observed in the other three genes in the region, it would be reasonable to suggest that this is the pathogenically relevant transcript; if the disease of interest is neurologic, it would be reasonable to hypothesize that the risk is mediated by reduced levels of form 1 messenger RNA.

Table 1

Genetic Progress through Technology.*

Scientific Advance	Technological Platform	Explanation	Reference
Sequencing of the human genome	Whole-genome expression arrays	Allows the expression of all genes to be determined by hybridization	Lander et al.,[12] Venter et al.,[13] Su et al.[14]
Human HapMap			
SNP technology		Demonstrates that individual SNPs predict adjacent SNPs and therefore suggests that genotyping of <500,000 SNPs may allow a nearly complete survey of all common genetic variability	The International HapMap Consortium[2]
Genome genotyping	Whole-genome SNP genotyping arrays	Allows whole-genome associations to be performed for common diseases, the commercial consequence of the HapMap	Sladek et al.[15]
High-throughput analysis	High-throughput sequencing techniques	Allows DNA sequencing that is faster and cheaper than conventional sequencing	Margulies et al.[16]
		Allows the expression of all RNA species, including different splice forms to be assessed in any tissue	Brenner et al.[17]
		Allows individual full-coding genome sequencing, together with whole-genome arrays that hybridize and bind to all exons	Olson[18]
Sequencing of the individual genome		Opens the way for personal genome sequencing	Wheeler et al.[19]
1000 Genomes Project		Allows the identification of comparatively rare polymorphic changes by placing the full genome sequences of 1000 anonymous subjects into the public domain	1000 Genomes Project[20]
Genotype-Tissue Expression (GTEx) project		Allows the creation of haplotypic gene-expression databases for many human tissues	NIH Roadmap for Medical Research[21]

*NIH denotes National Institutes of Health, and SNP single-nucleotide polymorphism.

Table 2

Benefits, Misconceptions, and Limitations of the Genomewide Association Study.

Benefits
Does not require an initial hypothesis
Uses digital and additive data that can be mined and augmented without data degradation
Encourages the formation of collaborative consortia, which tend to continue their collaboration for subsequent analyses
Rules out specific genetic associations (e.g., by showing that no common alleles, other than *APOE*, are associated with Alzheimer's disease with a relative risk of more than 2)
Provides data on the ancestry of each subject, which assists in matching case subjects with control subjects
Provides data on both sequence and copy-number variations
Misconceptions
Thought to provide data on all genetic variability associated with disease, when in reality only common alleles with large effects are identified
Thought to screen out alleles with a small effect size, when in reality such findings may still be very useful in determining pathogenic biochemical pathways, even though low-risk alleles may be of little predictive value
Limitations
Requires samples from a large number of case subjects and control subjects and therefore can be challenging to organize
Finds loci, not genes, which can complicate the identification of pathogenic changes on an associated haplotype
Detects only alleles that are common (>5%) in a population
Requires replication in a similarly large number of samples