Clinical reviews in allergy and immunology

Series editors: Donald Y. M. Leung, MD, PhD, and Dennis K. Ledford, MD

Application of genetic/genomic approaches to allergic disorders

Tesfaye M. Baye, PhD,a Lisa J. Martin, PhD,b and Gurjit K. Khurana Hershey, MD, PhDa,c
Cincinnati, Ohio

INFORMATION FOR CATEGORY 1 CME CREDIT

Credit can now be obtained, free for a limited time, by reading the review articles in this issue. Please note the following instructions.

Method of Physician Participation in Learning Process: The core material for these activities can be read in this issue of the Journal or online at the JACI Web site: www.jacionline.org. The accompanying tests may only be submitted online at www.jacionline.org. Fax or other copies will not be accepted.

Date of Original Release: September 2010. Credit may be obtained for these courses until August 31, 2012.

Copyright Statement: Copyright © 2010-2012. All rights reserved.

Overall Purpose/Goal: To provide excellent reviews on key aspects of allergic disease to those who research, treat, or manage allergic disease.

Target Audience: Physicians and researchers within the field of allergic disease.

Accreditation/Provider Statements and Credit Designation: The American Academy of Allergy, Asthma & Immunology (AAAAI) is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide continuing medical education for physicians. AAAAI designates these educational activities for a maximum of 1 AMA PRA Category 1 Credit™. Physicians should only claim credit commensurate with the extent of their participation in the activity.

List of Design Committee Members: Tesfaye M. Baye, PhD, Lisa J. Martin, PhD, and Gurjit K. Khurana Hershey, MD, PhD

Activity Objectives

1. To differentiate between the research methodologies currently used to study disorders that are at least in part heritable.
2. To increase awareness of current major international studies of inherited disorders.
3. To understand different types of genetic variation.

Recognition of Commercial Support: This CME activity has not received external commercial support.

Disclosure of Significant Relationships with Relevant Commercial Companies/Organizations: L. J. Martin has received research support from the National Institutes of Health. The rest of the authors have declared that they have no conflict of interest.

Completion of the human genome project and rapid progress in genetics and bioinformatics have enabled the development of large public databases, which include genetic and genomic data linked to clinical health data. With the massive amount of information available, clinicians and researchers have the unique opportunity to complement and integrate their daily practice with the existing resources to clarify the underlying cause of complex phenotypes, such as allergic diseases. The genome itself is now often used as a starting point for many studies, and multiple innovative approaches have emerged applying genetic/genomic strategies to key questions in the field of allergy and immunology. There have been several successes that have uncovered new insights into the biologic underpinnings of allergic disorders. Herein we will provide an in-depth review of genomic approaches to identifying genes and biologic networks involved in allergic diseases. We will discuss genetic and phenotypic variation, statistical approaches for gene discovery, public databases, functional genomics, clinical implications, and the challenges that remain. (J Allergy Clin Immunol 2010;126:425-36.)

Key words: Gene, allergy, database, browser, genome, common variants, rare variants, HapMap, imputation

Human genome variation encompasses all of the genetic characteristics observed within the human species. Genetic variation occurs both within and among populations and is the basis for natural selection. Insights regarding the distribution of genetic variants among human populations have recently become available.1 Interestingly, human genetic diversity decreases in native populations as the migratory distance from Africa increases, presumably because of limitations in human migration.2

Nucleotide diversity is based on single mutations called single nucleotide polymorphisms (SNPs) that occur at a rate of 1 SNP per 1,000 bp.3 Currently, there are more than 12 million SNPs deposited in GenBank, 6.5 million of which have been validated (http://www.ncbi.nih.gov/SNP). The bulk of variations at these nucleotide levels are not visible at the phenotypic level. A better understanding of the basis of genetic diversity was gained with the publication of full sequences of individuals genomes.4,5

The Human Genome Project and a parallel project by Celera Genomics yielded 2 haploid sequences; however, analysis of diploid sequences has revealed that non-SNP variation accounts for much more human genetic variation than single nucleotide diversity. Non-SNP variation includes copy number variation (CNV) and

From 1Cincinnati Children’s Hospital Medical Center, Division of Asthma Research; 2the Division of Epidemiology and Biostatistics; and 3the Division of Allergy and Immunology, Department of Pediatrics, University of Cincinnati.

Supported by National Institutes of Health grants U19AI70235 (G. K. K. H. and L. J. M.) and P30HL10133 (T. M. B.).

Received for publication March 11, 2010; revised April 28, 2010; accepted for publication May 7, 2010.

Available online July 20, 2010.

Reprint requests: Gurjit K. Khurana Hershey, MD, PhD, Division of Asthma Research, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH. E-mail: Gurjit.Khurana.Hershey@cchmc.org.

© 2010 American Academy of Allergy, Asthma & Immunology
doi:10.1016/j.jaci.2010.05.025
results from deletions, inversions, insertions and duplications. Copy number variation regions (CNVRs) have been found in 12% of the genome. CNVRs can be markedly different between populations and contain hundreds of genes, disease loci, functional elements, and segmental duplications. Taking into account this variation, as well as SNPs, human-to-human genetic variation is estimated to be approximately 0.5%. This 0.5% difference amounts to a significant number of distinct genetic traits that uniquely distinguish the genome of every person and contribute to unique and distinct risks for diseases, responses to environmental exposures (including nutrition), and responses to pharmacologic treatment.

**EPIGENETIC VARIATION IN ALLERGIC DISORDERS**

Epigenetic variation does not affect the underlying DNA code but rather modifies how it is expressed through covalent modifications, including DNA methylation, histone modifications, and microRNAs. It is the structural adaptation of chromosomal regions so as to register, signal, or perpetuate altered activity states. Detailed analysis of methylation across several chromosomes has demonstrated that the promoter regions of nearly 20% of genes are methylated, many of which influence transcription. Progressive accumulation of phenotypic differences between genetically identical monozygotic twins illustrates how pollution, smoking, mold, diet, habits, or, in general, environment can shape phenotype and disease susceptibility. Monozygotic twins are epigenetically indistinguishable early in life but exhibit substantial differences with age, particularly when they have led different lifestyles and spent less of their lives together. Therefore monozygotic twin discordance for many common disorders could be interpreted as the result of external environmental factors that modulate susceptibility through a change in the profile of epigenetic modifications that ultimately determine gene function. The field of epigenetics has emerged to explain how cells and how a phenotype can be passed from one cell to its daughter cells. It is now well established that epigenetic mechanisms are important to control the pattern of gene expression during development and the cell cycle and in response to biologic or environmental changes. Unlike genetic alterations, which are permanent and usually affect all cells, epigenetic modifications are cell type specific. Epigenetic regulation of the immune system occurs at many levels, including the differentiation of T cells. Epigenetic effects on gene expression can persist even after the removal of the inducing agent and can be passed on through mitosis to subsequent cell generations, constituting a heritable epigenetic change. In a somatic cell a heritable change can generate a dysfunctional clone of cells with phenotypic consequences (e.g., a tumor). In a germ-line cell a heritable change can be transmitted to the germ cells themselves (sperm or ova) and potentially to the next generation. In this model epialleles can be in linkage disequilibrium (LD) with SNPs that are genotyped in genome-wide association studies (GWASs). The role of epigenetics in allergic disease is becoming increasingly evident. One recent study showed that epigenetic reprogramming involving aberrant DNA methylation of a 5’-CpG island in acyl-CoA synthetase long-chain family member 3 (ACSL3) was significantly associated with asthma risk in children born to mothers exposed to air pollutants, such as traffic-related combustion emissions. Another study found that neonates of allergic mothers are born with substantial changes in DNA methylation in their splenic dendritic cells and that these dendritic cells show enhanced allergen-presenting activity in vitro. Current knowledge of epigenetics in allergic diseases is limited, and novel applications of epigenetic approaches, including genome-wide approaches to allergic diseases, are necessary to uncover the role of epigenetics.

**DEFINING PHENOTYPIC VARIATION IN ALLERGIC DISEASE**

The phenotype is defined as the observable characteristics of an organism, as determined by both genetic makeup and environmental influences, including individual physical, psychosocial, and environmental exposures (Fig 1). The genotype is the descriptor of the genome, which is the set of physical DNA molecules inherited from the organism’s parents, whereas phenotype is the descriptor of the phenome, the manifest physical properties of the organism, including its physiology, morphology, and behavior. Although single-gene disorders in classical Mendelian inheritance result in direct genotype-phenotype correspondence, the relationship between genotype and phenotype in traits of multifactorial (complex) inheritance is complicated. In complex diseases with a multifaceted phenotype, such as asthma, a given genotype can result in many different phenotypes, and there are different genotypes corresponding to a given phenotype. Although a subject’s genotype is fairly stable over a lifetime, his or her phenotype is dynamic, influenced by both the environment and the underlying genotype, including interactions between them. The definition, measurement, and validity of phenotyping need to be standardized to increase the quality of research and the reproducibility of genetic studies. Indeed, recently, the National Institutes of Health launched an initiative (Consensus Measures for Phenotypes and Exposures [PhenX]) to address the need standardization of phenotype and environmental exposure measures for cross-study comparison in genetic studies. These measures do not include information for allergic diseases; however, the National Institute of Allergy and Infectious Diseases recently partnered with the National Heart, Lung, and Blood Institute; the National Institute of Environmental Health Sciences; the National Institute of Child Health and Human Development; the Agency for Healthcare Research and Quality; the Merck Childhood Asthma Network; and the Robert Wood Johnson Foundation to host an Asthma Outcomes Workshop. The objective of this workshop was to develop standardized definitions and data collection methodologies for established and validated asthma outcomes measures. The goal is that these outcomes will be broadly used in National Institutes of Health–funded studies.
There are several important variables to consider when defining a phenotype for studies of allergic disorders, including disease definition, atopic status, comorbidities, and disease outcomes. For example, severe asthma is a recognized asthma phenotype defined by receiving ongoing treatment with high-dose inhaled corticosteroids, oral corticosteroids, or both for at least 6 months, with persistent symptoms or exacerbations when the controller medications are tapered. However, "severe asthma" is not a single phenotype. Population studies have revealed differences in severe asthma that begin in childhood versus adulthood. Childhood asthma is often "allergic," whereas adult-onset asthma is more heterogeneous and often is not related to allergy but rather to other influences, including aspirin sensitivity, hormonal influences, and occupational exposures. This heterogeneity strongly supports the need for genetic studies aimed at uncovering the mechanistic bases for each distinct phenotype rather than the mixed phenotype of asthma.

Age is an important factor in defining phenotypes for allergic disorders. As a population ages, it will be exposed to more environmental factors (eg, environmental tobacco smoke, diesel exhaust, and air pollution) that contribute to the pathogenesis of asthma and allergy, thus increasing sporadic (nongenetic) occurrences of these disorders. Thus when studying a cohort of adults, there will be a proportion of subjects who could be classified as having asthma because of environmental exposures without a major genetic risk. Children, on the other hand, might reduce the heterogeneity of the cause of asthma because they have had minimal time to accumulate environmental exposures, which would increase the risk of asthma. Given the risks of misclassification of asthma in the very young and heterogeneity in older groups, serious attention should be focused on the ages of participants. There has been a strong focus on powering genetic studies with very large sample sizes; however, large cohorts might not help improve our understanding of the genetic underpinnings of allergy phenotypes as much as precise phenotyping. Phenotypes can be defined through combinations of clinical information and individual biomarker and molecular data.

The phenotypic definition of control subjects is another important consideration, especially in studies of allergic disease, in which some features might overlap. For example, allergic sensitization might overlap with childhood asthma, and therefore if a study aims to identify specifically childhood asthma genes, the control group should include sensitized subjects without asthma. The selection of the control subjects should be based on the goals of the research. With the availability of genotypic and phenotypic data through public resources, such as the Database of Genotypes and Phenotypes (dbGAP; http://www.ncbi.nlm.nih.gov/gap), it is enticing to consider the recruitment of control subjects as unnecessary. However, control subjects unsolicited with respect to phenotype increase the number of participants required to obtain similar power when using control subjects who do not have the phenotype of interest. This is compounded by the fact that the publicly available control subjects are likely to be from a different population than the cases. When this situation occurs, researchers should consider applying genetic ancestry matching (discussed below) to minimize population stratification.

**STATISTICAL APPROACHES TO FINDING GENETIC VARIATION IN ALLERGIC DISORDERS**

There are 3 main statistical approaches to gene discovery: linkage, association, and admixture mapping. Linkage analysis tests to determine whether a variant cosegregates with disease in families, association analysis tests to determine whether a genetic variant occurs more often in subjects with disease than those without disease, and admixture mapping tests to determine whether there are particular regions of the genome at which inheriting DNA from ancestors from a certain region of the world
predisposes one to particular diseases. Linkage studies can be performed only in family-based studies, whereas association testing and admixture mapping can be performed in both population- and family-based studies. These approaches might appear to ask the same questions, but statistically, these are independent tests, and the strategy affects the hypotheses that can be tested.

Linkage analysis is based on the assumption that the genetic marker and the disease variant are in close proximity and transmitted intact across generations. Thus markers in close proximity to the disease-causing gene segregate with disease in families. However, the resolution of linkage is poor, with candidate regions encompassing hundreds of genes. Thus linkage analysis only identifies regions and not genes or variants. Furthermore, because linkage is statistical evidence, replication is the gold standard to minimize the risk of false-positive results.

An alternative approach is an association study, which can use population- or family-based designs. It is important to recognize that association does not equal causation. Association studies simply measure statistical dependence between 2 or more variables. Significant associations can be due to one of several misleading factors, including LD, population stratification, or random chance. Once significance is achieved, replication is required to ensure its validity.

Admixture occurs when 2 or more genetically diverse populations merge to form a new population. Localizing disease genes by using an admixed population is called admixture mapping. In human admixture studies researchers combine information about known population history with information from subjects’ measured genotypes using known ancestry-informative markers (AIMs). Studies consistently show that all the genetic disorders, such as asthma, are more common in persons of West African ancestry compared with persons of European ancestry. The African American population is an admixed population for which about 20% of the genetic material traces to European ancestry. The association between increased asthma risk and African ancestry and the admixed nature of the African American population suggest that admixture mapping might be an important asthma gene–finding strategy to study genetically heterogeneous populations.

With current technology, it is not cost-prohibitive to perform genome-wide linkage and association studies. An advantage of the genome-wide approach is that it requires no a priori evidence and thus has the ability to identify regions and variants in genes previously not implicated in allergic disorders and provide insight into the biologic underpinnings for these disorders. Researchers using genome-wide approaches must adjust the level of significance to ensure that findings did not occur by chance; with the increased numbers of statistical tests, the likelihood of obtaining a P value of .05 increases. For the current GWAS SNP chips (density of 1 million SNPs), significance thresholds of 10^-8 are required to control for multiple comparisons. Given this level of significance, the number of samples required to obtain adequate power in a GWAS is in the thousands for a gene with modest effect. By limiting the analysis to those gene regions that have promising a priori evidence of being involved with asthma, the severity of the correction for multiple testing becomes much less severe. A candidate gene study examining 1,000 SNPs will require only 60.5% of the sample size required by a GWAS study examining 1 million SNPs to obtain the same statistical power of 80%. This reduced sample requirement might permit better phenotyping and reduced heterogeneity, which will also improve the power. Thus there are benefits to both GWASs and candidate gene approaches.

Because asthma is a prevalent disorder, the classic population-based sampling strategy is case-control. In this approach the researcher collects subjects with disease (cases) and unrelated subjects without disease (control subjects). This method is very efficient; compared with a random sampling design, only 35% of the total sample would be required for equivalent power (assuming an asthma frequency of 10%). Although this approach appears simple, the challenge is ensuring that the control subjects come from the same ancestrally homogeneous population as the cases. When cases and control subjects are not drawn from the same ancestral population, population stratification can result in spurious associations. For example, suppose most persons of African ancestry in a sample had brown eyes and also happened to have asthma, whereas most persons of European ancestry were blue eyed and asthma free. A naive analysis might conclude that the brown-eyes SNP is responsible for asthma, even if eye color and disease are completely unrelated; that is, the methods are likely to nab the wrong SNP suspects because of “guilt by association.” This problem becomes more pronounced in studies surveying the entire genome because of the huge number of ancestry-related SNPs being tested. Researchers can test whether cases and control subjects differ over a large number of variants not expected to be associated with disease to address this genetic mixing problem. If differences exist, adjustments can be made to minimize this effect. Currently, 3 fundamentally different methods are used to correct for confounding in allergy genetic association studies. These methods are (1) genomic control, (2) structured association, and (3) principal component analysis. Genomic control uses a set of noncandidate unlinked loci to estimate an inflation factor, I, which was caused by the population structure present and then corrects the standard \( \chi^2 \) test statistic for this inflation factor. The structured association method uses Bayesian techniques to assign subjects to “clusters,” or subpopulation classes, by using information from a set of noncandidate unlinked loci and then tests for an association within each cluster. An AIM panel can be used to control for population confounding by variations in background ancestry during structural association testing. Therefore AIMs can be termed structural informative markers. These markers exhibit differences in frequencies between population groups. Importantly, care should be taken in selecting which AIMs to use because some sets might be population specific. Principal component analysis involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. It can be used to identify and adjust for population substructure. Family-based association tests protection against stratification, a decided advantage of family-based designs.

**USE OF PUBLIC DATABASES TO INFORM GENETIC DATA**

Publicly available databanks now contain billions of nucleotide of DNA sequence data collected from over 260,000 different organisms. This proliferation of data from genome sequencing over the past decade has resulted in dramatic changes in the way the scientific community is communicating and carrying out genomic research. Once a genome-wide or candidate gene study has been performed, the investigator can readily obtain...
information about an identified SNP, including where it is located, its potential functional significance, its frequency in different populations, and whatever else might already be known (Fig 2). A summary of available public resources appears in Table I. The PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) site will provide information on whether an SNP is in a gene and whether there are reported genotypic and allelic frequencies for major population groups. The database of genomic variants (http://projects.tcag.ca/variation/) is also a useful tool. This site permits the researcher to zoom out and get a broader view of the genomic region containing the SNP of interest, including features such as newly reported genes, transcripts, and copy number variants. The Web site for the University of California, Santa Cruz, Genome Browser (http://genome.ucsc.edu) also provides excellent information about the features of the genome in a particular region. Although each of these sites is an excellent tool to examine a small number of SNPs, a large number of SNPs can be investigated efficiently by using a high-throughput method, such as the SNP and CNV annotation database (http://genemem.bsd.uchicago.edu/newscan). Once the most promising SNPs have been identified, databases are available to provide estimates of putative functionality of the SNPs. FASTSNP (http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp) evaluates all SNPs in a gene region by using the methodology proposed by Tabor and colleagues.43,44 If the SNP is nonsynonymous, then SNPeff (http://snpeffact.vib.be/search.php) can provide additional information about the molecular properties of the variant. The Genetic Association Database (http://geneticassociationdb.nih.gov/) is a useful tool to determine what is already known about a specific SNP or genes in terms of disease associations. It is an archive of genetic association studies and is searchable by both disease and gene.45 A catalog of published GWASs is regularly updated and deposited at http://www.genome.gov/GWAStudies.46 Another resource available is the relationship between SNP variants and gene expression (http://www.scandb.org).47

**HAPMAP, TAGGING SNPs, AND IMPUTATION ANALYSIS**

The International HapMap Project in 2002 initiated the construction of a genome-wide SNP database of common variation (http://www.hapmap.org) to accelerate the identification of common disease alleles. In brief, the phase I and II project has genotyped more than 3 million SNPs in 269 samples from 4 populations: 90 Utah Residents (30 parent-offspring trios) with Northern and Western European Ancestry; 45 Han Chinese subjects from Beijing, China; 44 Japanese subjects from Tokyo, Japan; and 90 Yoruban subjects (30 trios) from Ibadan, Nigeria. The average spacing of the map is 1 SNP per 1,000 bp, and this vast resource is currently being used globally as a template for both LD-based candidate gene studies and GWASs in allergic disorders. The HapMap phase III has recently released a draft version of the dataset (http://www.hapmap.org) to increase the sample size to more than 1,000 subjects in 11 populations. HapMap genotypic data, allele frequencies, LD data, phase information, and sample documentation are publicly and freely available for download from the HapMap Web site (http://www.hapmap.org).

Although whole human genome sequencing is possible,48 the costs and challenges with dealing with such a large quantity of data make this approach untenable currently. However, SNPs that are physically close to one another on the chromosome are more likely to be inherited together than SNPs farther apart. LD is a measure of this nonrandom correlation between pairs of SNPs. Thus if a causal variant is in LD with a marker SNP, then the marker will be associated with the phenotype proportional to the degree of LD between them. Furthermore, there are blocks of high LD conserved within populations.49 The coinheritance between SNP alleles showing strong LD enable most of the common genetic variations in a region to be captured by genotyping subsets of SNPs (termed haplotype-tagging SNPs) across a candidate gene or region of interest. Because redundant information can be reduced (thus reducing cost), many studies will often use the tagging

---

**FIG 2.** Public databases can be used to rapidly provide key information about putative disease-associated genetic variants.
<table>
<thead>
<tr>
<th>Categories</th>
<th>Functional description</th>
<th>Web site (URL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database/browser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Cruz</td>
<td>SNP, gene, or genomic region browser</td>
<td><a href="http://genome.ucsc.edu/cgi-bin/hgGateway">http://genome.ucsc.edu/cgi-bin/hgGateway</a></td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA databank Japan</td>
<td><a href="http://www.ddbj.nig.ac.jp">http://www.ddbj.nig.ac.jp</a></td>
</tr>
<tr>
<td>Gene expression/ontology database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ArrayExpress</td>
<td>Functional genomics data repository</td>
<td><a href="http://www.ebi.ac.uk/microarray-as/ae/">http://www.ebi.ac.uk/microarray-as/ae/</a></td>
</tr>
<tr>
<td>GO</td>
<td>Gene and gene product attributes across species</td>
<td><a href="http://www.geneontology.org">http://www.geneontology.org</a></td>
</tr>
<tr>
<td>UniProt</td>
<td>Central repository for protein sequence and function</td>
<td><a href="http://www.uniprot.org/">http://www.uniprot.org/</a></td>
</tr>
<tr>
<td>Population common/rare variances resequencing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HapMap</td>
<td>Information on 3.1 million SNPs from multiple reference/template populations</td>
<td><a href="http://www.hapmap.org/cgi-perl/gbrowse">http://www.hapmap.org/cgi-perl/gbrowse</a></td>
</tr>
<tr>
<td>1000 Genomes Project</td>
<td>Sequence variants with minor allele frequency of 1% from HapMap</td>
<td><a href="http://www.1000genomes.org/page.php">http://www.1000genomes.org/page.php</a></td>
</tr>
<tr>
<td>The Exome Project</td>
<td>High-throughput resequencing, protein-coding regions</td>
<td><a href="http://exome.gs.washington.edu">http://exome.gs.washington.edu</a></td>
</tr>
<tr>
<td>Single gene/variant browser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPper</td>
<td>Map SNPs into genes and chromosome position</td>
<td><a href="http://snpper.chip.org/bio/snpper-enter">http://snpper.chip.org/bio/snpper-enter</a></td>
</tr>
<tr>
<td>SPSmart</td>
<td>SNP allele frequency summary from multiple populations</td>
<td><a href="http://spsmart.cesga.es/hapmap">http://spsmart.cesga.es/hapmap</a>. php?dataSet=hapmap</td>
</tr>
<tr>
<td>PupaSuite</td>
<td>Explore SNPs with potential phenotypic effects</td>
<td><a href="http://pupasuite.bioinfo.cifp.es">http://pupasuite.bioinfo.cifp.es</a></td>
</tr>
<tr>
<td>Haplotter</td>
<td>Detect signature of SNP natural selection</td>
<td><a href="http://hg-wen.uchicago.edu">http://hg-wen.uchicago.edu</a></td>
</tr>
<tr>
<td>SNPedia</td>
<td>Wiki investigating the effects of variations in human DNA</td>
<td><a href="http://www.snpedia.com/index.php/SNPedia">http://www.snpedia.com/index.php/SNPedia</a></td>
</tr>
<tr>
<td>BLAST</td>
<td>Used for checking SNP assay primer designs</td>
<td><a href="http://blast.ncbi.nlm.nih.gov/Blast.cgi">http://blast.ncbi.nlm.nih.gov/Blast.cgi</a></td>
</tr>
<tr>
<td>SCAN</td>
<td>Maps SNPs into genes, identifies flanking genes, and associates with gene expression profiles (can do multiple SNPs easily)</td>
<td><a href="http://www.scandb.org/newinterface/about.html">http://www.scandb.org/newinterface/about.html</a></td>
</tr>
<tr>
<td>SNPator</td>
<td>Format (eg, creating) input files for different analysis</td>
<td><a href="http://www.snpator.org/public/new_login/index.php">http://www.snpator.org/public/new_login/index.php</a></td>
</tr>
<tr>
<td>Database Genomic Variants</td>
<td>Maps SNPs onto the genome and identifies the structural variation in the human genome</td>
<td><a href="http://projects.tcag.ca/variation">http://projects.tcag.ca/variation</a></td>
</tr>
<tr>
<td>SNAP</td>
<td>SNP Annotation and Proxy Search based on LD; excellent tool for determining SNPs in LD with the SNP of interest</td>
<td><a href="http://www.broadinstitute.org/mpg/snap">http://www.broadinstitute.org/mpg/snap</a></td>
</tr>
<tr>
<td>SNPselector</td>
<td>Selecting SNPs for genetic association studies</td>
<td><a href="http://primer.duhs.duke.edu/">http://primer.duhs.duke.edu/</a></td>
</tr>
<tr>
<td>FastSNP</td>
<td>Uses decision tree analysis to classify SNPs into putative functional effect</td>
<td><a href="http://fastsnp.ibms.sinica.edu.tw">http://fastsnp.ibms.sinica.edu.tw</a></td>
</tr>
<tr>
<td>Onto-Express</td>
<td>Identify enriched functional ontologies/expressed genes</td>
<td><a href="http://vortex.cs.wayne.edu">http://vortex.cs.wayne.edu</a></td>
</tr>
<tr>
<td>SNPeffect</td>
<td>Determines the putative functionality of SNPs in coding and regulatory regions with rs no.</td>
<td><a href="http://snpeffect.vib.be/search.php">http://snpeffect.vib.be/search.php</a></td>
</tr>
<tr>
<td>SNPs3D</td>
<td>Web site that assigns molecular functional effects of nonsynonymous SNPs based on structure and sequence analysis</td>
<td><a href="http://www.snps3d.org">http://www.snps3d.org</a></td>
</tr>
<tr>
<td>GeneSNPS</td>
<td>Graphic view of SNPs in the context of gene elements</td>
<td><a href="http://www.genome.utah.edu/gensnps">http://www.genome.utah.edu/gensnps</a></td>
</tr>
<tr>
<td>PolyPhen</td>
<td>Predicts the effect of amino acid substitution on the structure and function of proteins</td>
<td><a href="http://genetics.bwh.harvard.edu/pph">http://genetics.bwh.harvard.edu/pph</a></td>
</tr>
<tr>
<td>PolyDoms</td>
<td>Web-based application that maps synonymous and nonsynonymous SNPs onto known functional protein domains</td>
<td><a href="http://polydoms.cchmc.org/polydoms">http://polydoms.cchmc.org/polydoms</a></td>
</tr>
<tr>
<td>GVS: Genome Variation Server</td>
<td>Integration of dense, gene-centric SNP maps from dbSNPs with genomic HapMap SNPs</td>
<td><a href="http://gvs.gs.washington.edu/GVS">http://gvs.gs.washington.edu/GVS</a></td>
</tr>
<tr>
<td>Consensus Measures for Phenotypes and Exposures (PhenX)</td>
<td>Cross-study comparisons by providing standard measures of phenotype and environmental exposures</td>
<td><a href="https://www.phenxtoolkit.org">https://www.phenxtoolkit.org</a></td>
</tr>
<tr>
<td>Prioritization of candidate genes in linkage or genomic region</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SNP approach. A challenge is that tagging SNPs are not selected for their likelihood to be functional. However, recent work has shown that information from unmeasured SNPs can be imputed by using tagging SNPs.\textsuperscript{50,51} Imputation requires use of a reference population in which genotypic information is available for a large number of SNPs.\textsuperscript{52} Although some of these SNPs would overlap with the genotyped tagging SNPs in a given study, others would be untyped SNPs in LD with the genotyped SNPs. By delineating the genotype patterns in the reference set, researchers can make reasonable inferences about what genotypes are likely to be carried by subjects at untyped SNPs in their study. It is essential that the reference population is similar in ancestry to the population in which imputation will be performed. Fortunately, HapMap\textsuperscript{53} provides publicly available information on more than 3 million SNPs in 4 major ancestry groups. Once imputation is performed, imputed SNPs can be tested for association with disease in the population of interest.\textsuperscript{52} Because imputation interrogates all common variants, the likelihood of identifying biologically relevant associations (eg, with functional variants) is greater. Another advantage of imputation is that studies might not use the same SNPs in the original discovery phase. With imputation, even studies that have investigated different SNPs can be combined to determine the overall evidence for a given association.\textsuperscript{52}

### RARE VARIANTS IN ALLERGIC DISORDERS

Most genetic studies, including GWASs, investigating common diseases have focused on common genetic variants on the assumption that common variants are mostly likely to contribute to common diseases (common disease/common variant hypothesis).\textsuperscript{54} There is emerging interest in association studies of rare variants, and it is hypothesized that rare variants are more likely to be functional than common variants. Furthermore, recent evidence supports that rare genetic variants can create synthetic associations that are credited to common variants.\textsuperscript{55} Although genetic association and linkage studies are well suited to find common variants for common diseases, they are not optimal for identification of rare variants.\textsuperscript{56} Rare alleles with major phenotypic effects can contribute significantly to common traits in the general population.\textsuperscript{57} Sequencing of candidate genes or entire genomes is the optimal way to identify rare variants. Unfortunately, most current studies are not designed or powered to identify, test, or both the contributions of rare SNPs to common disease. Although current approaches are not optimal to elucidate rare variants, they can identify regions of interest, which harbor rare variants; these regions can then be further analyzed by means of deep resequencing (the determination of a new genome sequence relative to a reference genome is often referred to as resequencing).

Recently, approaches have been used to study the potential health effects of private SNPs (ie, SNPs that have only been found in a given population).\textsuperscript{58} In one study investigators explored private SNPs in specific populations that might have phenotypic effects. They found that these SNPs contribute to variability in several cellular processes.\textsuperscript{59} Such variability might provide clues regarding ethnicity-specific responses to diseases or drugs. Another recent study found that in African American subjects private SNPs were associated with asthma.\textsuperscript{60} Investigation of rare and private SNPs requires deep-sequencing approaches. The 1000 Genomes Project, a deep-resequencing project aimed at providing detailed genetic variation data on more than 1,000 genomes from 11 populations around the world, will aid these efforts (www.1000genomes.org). This project will identify more than 95% of the variants with allele frequencies of greater than 1% in the human genome, substantially enhancing the HapMap data. Results from the 1000 Genomes Project will provide data to allow evaluation of the common disease/common variance hypothesis versus the common disease/more rare variants hypothesis.\textsuperscript{61}

### FUNCTIONAL GENOMICS

Once a genetic study has been performed and allergy-causing variants have been identified, the investigator can gain information to unify the biologic function of gene products. Several groups have reported that genes involved in predisposing to a given polygenetic disease tend to share more commonalities (annotated by similar Gene Ontology [GO] terms) in their molecular function or biologic pathway than genes chosen at random or genes not involved in the same disease.\textsuperscript{52-69} GO (http://www.geneontology.org) can be used to identify commonalities...
between gene products in the form of an agreed ontology. It provides a controlled vocabulary about genes and gene products based on known or predicted molecular function, cellular location, and biologic process. Because of the existing homologies between proteins among different taxa, the GO terms provide researchers with a powerful way to query and analyze functional genomic information in a way that is independent of species. Once genetic analyses determine which genes (among the thousands analyzed) might be related to the phenotypes, functional genomics experiments allow the scaling of the classical functional experiments to a genomic level. The GO analysis could potentially be used to reduce the number of targets of a large group of correlated genes and to find biologic functions potentially affected by multiple genes. In summary, GO annotation terms are enriched among genes linked to the trait, and such commonalities are often sufficient to narrow the list of candidate genes.

INTEGRATION OF GENE EXPRESSION AND SEQUENCE VARIATION APPROACHES IN ALLERGIC DISORDERS

Both coding and noncoding variability contribute to genetic variation. Novel approaches to capture human genetic variation have integrated global gene expression arrays, DNA sequence variation arrays, and public databases (Fig 3). This strategy has been successfully applied to asthma. In association studies the investigators found markers on chromosome 17q21 to be reproducibly associated with childhood asthma. They then evaluated the relationships between the markers and transcript levels of genes in cell lines derived from children in the association study. The SNPs associated with childhood asthma were associated with transcript levels of ORM1-like protein 3 gene (ORMDL3), suggesting that genetic variants regulating ORMDL3 expression are determinants of susceptibility to childhood asthma. Thus gene expression data informed the genetic data and provided insights regarding the biologic mechanisms that might be involved. Gene expression arrays can also be used in a discovery approach to identify dysregulated genes and pathways. The gene expression profiles can be used to identify key regulatory networks, to identify novel potential candidate genes, and to define phenotypes, which can then serve as quantitative traits for genetic studies. Variation in gene expression is an important mechanism underlying susceptibility to complex disease. An integrated genetic/genomic approach allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression quantitative trait loci). The major public data repositories, ArrayExpress and Gene Expression Omnibus, house raw microarray data and serve as warehouses for processed experimental data, facilitating gene-based queries of multiple expression profiles. ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae) is a public repository for experimental microarray data that is able to be queried based on a range of gene annotations, including gene symbols, GO terms, and disease associations. Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) is a public repository that archives and freely distributes microarrays, next-generation sequencing, and other forms of high-throughput functional genomic data.

SUCCESSES AND CLINICAL IMPLICATIONS

By using a candidate gene approach, common mutations in the filaggrin gene (FLG, 1q21) have been implicated in the causation
of ichthyosis vulgaris. Filaggrin (filament aggregation protein) is a major epidermal protein involved in maintaining the skin barrier, and previous studies have demonstrated that filaggrin was absent or reduced in the skin cells of subjects with ichthyosis vulgaris. Several independent replication studies have now provided convincing evidence of an association of FLG mutations with atopic dermatitis (AD). The estimated penetrance varies from 42% to 79%, that is, between 42% and 79% of subjects with 1 or more FLG null mutations are likely to have AD. The discovery that null mutations in FLG are associated with atopic eczema represents the single most significant breakthrough in understanding the genetic basis of this complex disorder. In addition, this association has yielded important insights into the biologic underpinnings of AD and support for the hypothesis that a barrier defect might be a contributory mechanism for the pathogenesis of AD and related atopic disorders. The exact contribution of FLG to atopic disorders remains to be delineated. The identification of patients with these FLG mutations might facilitate the targeting of novel therapies to repair or replace the defective epidermal barrier.

GWASs have also yielded successes. As discussed above, the association of ORMDL3 with asthma was first identified by means of a GWAS. Since the initial report, multiple groups have replicated the association between ORMDL3 variants and asthma. Furthermore, these variants have recently been found to associate not only with ORMDL3 expression but also with transcripts of multiple genes in this region. Increased expression of ORMDL3 has been associated with the unfolded-protein response. There is still much work to be done in this area, but it further illustrates how genetic/genomic approaches can provide insights into novel biologic networks and potential disease mechanisms.

MISSING HERITABILITY AND FUTURE DIRECTIONS

Genetic associations, including GWASs, have identified hundreds of genetic variants associated with complex human diseases, including 43 replicated genes for asthma. Most variants identified thus far confer relatively small increments in risk and explain only a small proportion of disease heritability. This has led to considerable speculation regarding the sources of the remaining “missing heritability.” Much of the speculation has focused on the possible contribution of rare variants (minor allele frequency, 0.5% to 5.0%). Such variants are not sufficiently frequent to be captured by current genotyping arrays nor do they carry sufficiently large effect sizes to be detected by current studies. With the completion of the human genome, more focus has gone into dense resequencing of regions. Because the cost of sequencing is still high, researchers often sequence DNA pools to identify variants that can be explored with additional genotyping. The pooled samples reliably detect variants at a frequency of 1% or greater with as little as 287 samples. Furthermore, if overlapping pools are used, these samples can be used to estimate allele frequencies. Once variants are identified, the next challenge is how to proceed. Much larger samples are needed for the identification of associations with variants than those needed for the detection of the variants themselves. One technique that has been used is to group rare variants such that the presence of any one of a number of rare variants is examined for disease association. However, this is complicated by the fact that the rare variants might have disparate effects on phenotype, making this approach uninterpretable.

Structural variants, including copy number variants (including insertions and deletions) and copy neutral variation (including inversions and translocations), might account for some of the unexplained heritability. Although the variation affecting large chromosomal regions can result in large phenotypic perturbations, small/regional CNVs can have minimal-to-severe effects on phenotype. In 2006, the first comprehensive CNV map of the human genome was published. Since then, CNVs have been associated with many different diseases, including asthma. The challenge for copy number variants is detection. Furthermore, in a recent study 2 copy number algorithms resulted in poor agreement. Thus although CNV analysis offers promise, the technical and statistical assessment of CNVs is still evolving.

The modest size of genetic effects detected thus far confirms the multifactorial cause of these complex disorders. The next frontier of genetic studies will require innovative approaches to look for the sources of missing heritability. This will include application of whole-genome sequencing to persons with extreme phenotypes, use of expanded genome variation data provided by the 1000 Genomes project, development of novel methods to detect additional sources of variation, improved phenotyping and use of expression quantitative trait loci, expanded efforts in epigenetics and identification of epigenetic variation, rigorous assessment of environmental influences and gene-environment interactions, assessment of gene-gene interactions, and the design of meta-studies with well-defined consistent phenotypes spanning across large population sets.

What do we know?

- Genetic variation plays a large role in asthma and allergic disease risk.
- Non-SNP variation accounts for much more human genetic variation than single nucleotide diversity. CNVRs (copy number variation) have been found in 12% of the genome.
- Whole-genome information and high-throughput tools are now available for high-resolution mapping.
- Gene-environment interactions play an important role in allergic diseases and have been relatively well studied in model organisms.
- Epigenetic effects on gene expression can persist even after removal of the inducing agent and can be passed on, through mitosis, to subsequent cell generations, constituting a heritable epigenetic change.
- There are 3 main statistical approaches to identify disease-associated genes: linkage, association, and admixture mapping.
- Recent evidence has revealed that rare alleles with major phenotypic effects can contribute significantly to common traits in the general population. Sequencing of candidate genes or entire genomes is currently the optimal way to identify rare variants.
- Recent evidence has revealed that rare/private SNPs can contribute significantly to common traits in the general population. Although genetic association and linkage studies are well suited to find common variants for common diseases, they are not optimal for identification of rare variants. Sequencing of candidate genes or entire genomes is currently the optimal way to identify rare variants.
Novel approaches to capture human genetic variation have integrated global gene expression arrays, DNA sequence variation arrays, and public databases. Variation in gene expression is an important mechanism underlying susceptibility to complex disease. An integrated genetic/genomic approach allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression quantitative trait loci).

What is still unknown?

- Identified variants account for a small proportion of disease, and the factors that contribute to the majority of the heritability of allergic diseases are still unknown.
- The effect of structural variation (including CNV) on asthma and allergic disease is unclear. Furthermore, the technical and statistical assessment of CNVs is still evolving.
- Linkage of genetic variation to phenotypic variation and to translation into biologic function is still in its infancy.
- Rigorous quantitative assessment of environmental influences will be necessary to elucidate gene-environment interactions in human subjects.
- Approaches to efficiently dissecting the role of gene-gene and gene-environment interactions, epigenetics, and imprinting are lacking.
- A positive association does not imply causality or a direct effect on gene expression or protein function.
- The role of rare variants is unclear. Furthermore, although genetic association and linkage studies are well suited to find common variants for common diseases, they are not optimal for identification of rare variants.
- Although rare and private SNPs are largely unknown, the 1000 Genomes Project, a deep-resequencing project, will provide detailed genetic variation data on more than 1,000 genomes from 11 populations around the world.
- Genetic studies have identified hundreds of genetic variants associated with complex human diseases, including 43 replicated genes for asthma. The variants identified thus far confer relatively small increments in risk and explain only a small proportion of disease heritability. The clinical implications (i.e., the contribution of the genetic variation to asthma subphenotypes, variations in treatment response, and different disease outcomes) remain largely undetermined.

REFERENCES

65. Huang D, Wei P, Pan W. Combining gene annotations and gene expression data in
64. Baye TM, Wilke RA, Olivier M. Genomic and geographic distribution of private
63. Visscher PM, Andrew T, Nyholt DR. Genome-wide association studies of quan-
58. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying com-
56. Salmela E, Lappalainen T, Fransson I, Anderson PM, Dahlman-Wright K, Fiebig
54. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage dis-
52. Baye TM, Wilke RA, Olivier M. Genomic and geographic distribution of private
44. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG. Systematic analysis,
42. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG. Systematic analysis,
41. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage dis-
40. Myles S, Stoneking M, Timpson N. An assessment of the portability of ancestry in-
39. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying com-
37. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Prin-
34. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG. Systematic analysis,


