Schizophrenia Bulletin vol. 33 no. 1 pp. 95–104, 2007
doi:10.1093/schbul/sbl045
Advance Access publication on October 11, 2006

Statistical Genetics Concepts and Approaches in Schizophrenia and Related Neuropsychiatric Research

Nicholas J. Schork[1–6], Tiffany A. Greenwood[2], and David L. Braff[2]

[2]Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0603; [3]Department of Family and Preventive Medicine; [4]The Center for Human Genetics and Genomics; [5]The Moores UCSD Cancer Center; [6]The California Institute of Telecommunications and Information Technology, University of California, San Diego

Statistical genetics is a research field that focuses on mathematical models and statistical inference methodologies that relate genetic variations (ie, naturally occurring human DNA sequence variations or "polymorphisms") to particular traits or diseases (phenotypes) usually from data collected on large samples of families or individuals. The ultimate goal of such analysis is the identification of genes and genetic variations that influence disease susceptibility. Although of extreme interest and importance, the fact that many genes and environmental factors contribute to neuropsychiatric diseases of public health importance (eg, schizophrenia, bipolar disorder, and depression) complicates relevant studies and suggests that very sophisticated mathematical and statistical modeling may be required. In addition, large-scale contemporary human DNA sequencing and related projects, such as the Human Genome Project and the International HapMap Project, as well as the development of high-throughput DNA sequencing and genotyping technologies have provided statistical geneticists with a great deal of very relevant and appropriate information and resources. Unfortunately, the use of these resources and their interpretation are not straightforward when applied to complex, multifactorial diseases such as schizophrenia. In this brief and largely nonmathematical review of the field of statistical genetics, we describe many of the main concepts, definitions, and issues that motivate contemporary research. We also provide a discussion of the most pressing contemporary problems that demand further research if progress is to be made in the identification of genes and genetic variations that predispose to complex neuropsychiatric diseases.

Key words: genetic epidemiology/genotyping/ association studies/hapmap

Introduction

Contemporary statistical genetics research primarily focuses on the development and implementation of data analysis methodologies that facilitate the identification of genes and naturally occurring genetic variations (ie, inherited DNA sequence variations or polymorphisms) that influence phenotypic expression and disease susceptibility. The fact that there are approximately 10 million polymorphic sites in the human genome (http://www.hapmap.org),[1] any one or subset of which may contribute to disease susceptibility, complicates the identification of variations that are causally related to a particular disease. In addition, most neuropsychiatric conditions of contemporary public health concern, such as schizophrenia and bipolar disorder, are complex and multifactorial in that many genes and environmental factors, as well as their interactions, their expression.[2] This fact further complicates disease gene identification via statistical genetic analysis because the influence of any one gene or genetic variation may be obscured by the influences of other genes and/or environmental factors.

In the following, we describe some of the basic tenets and strategies in contemporary statistical genetics. We do this in a nonmathematical way, focusing on the biological phenomena exploited by relevant mathematical and statistical genetics models. In addition, we consider challenges associated with the use of information and resources provided by large-scale human genetic initiatives, such as the Human Genome Project and the International HapMap project, as well as modern high-throughput genomic technologies, in statistical genetic analysis models. The material is organized as follows. We first consider the impact of genetic variation on molecular and subclinical physiological phenotypic expression, population biology, and evolution. This discussion will motivate the descriptions of unique statistical problems and methods associated with each of these areas. We then consider the basic problems inherent in linking genetic variations with phenotypes in the absence of information about physiological links between those variations and phenotypes. We then describe the unique statistical genetic problems and potential solutions that arise as a result of the appreciation that genetic variations impact many different biological phenomena each of

[1]To whom correspondence should be addressed; tel: 858-822-5571, fax: 858-822-2113, e-mail: nschork@ucsd.edu.

which may be of relevance to a disease. We close with a brief discussion and a few concluding remarks.

The Basic Biological Influence of Genetic Variation

Inherited DNA sequence variations that ultimately contribute to disease susceptibility can be considered the most fundamental set of pathologies associated with disease. However, the relationship between those variations and the expression of a disease such as schizophrenia is often extremely complicated from a biochemical and physiological perspective, a fact which makes the identification and characterization of those relationships difficult, except in rare instances, such as Huntington disease in which a disease is influenced entirely by the presence of a single mutation.[3,4] The primary reason for this complexity concerns the manner in which genes influence physiologic (and pathophysiologic) function. Genes and their products work in combinations and within networks, which creates potential for feedback, redundancy, and the existence of compensatory mechanisms that may overcome a defect in any one gene (a point that is often overlooked in schizophrenia genetics research). In addition, the effect of a DNA sequence variation must manifest itself at many different levels of a complicated physiologic hierarchy if it is to make its mark on overt phenotypic expression. If the variation is indeed deleterious and influences the reproductive capacity of individuals who possess it, it is likely to be selected against, such that its frequency in the population at large will diminish and ultimately reach zero over time. Because schizophrenia has persisted across cultures and racial and ethnic groups across time, however, it is possible that schizophrenia vulnerability genes confer some functional advantages.[5]

Genetic Variation and the Flow of Biological Information

Figure 1 is meant to capture some aspects of the influence of DNA sequence variation on physiology and population biology and will serve as a mechanism for pointing out contemporary statistical genetic issues and problems. Essentially, variation in a gene can influence the level of the expression of gene or the structure of the encoded protein product. Because proteins form the "building blocks" of life, they work in combination to dictate, eg, the flow of metabolites within and across biochemical pathways, the generation and regeneration of crucial neural substrates and tissues,[6] and other phenomena at the microphysiologic level. These events in turn dictate activities at a more macrophysiologic level, which encompasses organ function, hemodynamic balance, and neural network interactions,[7] which, again, in turn influence the manifestation of subclinical (eg, "mild" schizotypal traits[8]) and overt, clinical phenotypes (eg, disease endpoints defined by the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (*DSM-IV*) and ascertained via structured clinical interviews[9]). Given that diseases are often highly deleterious, the number of individuals carrying the causative DNA sequence variations is dictated in part by the severity of the disease phenotypes they induce, and this fact bears on the frequency of those variations among individuals seen in the clinic and in the population at large. Mediating factors in the population, such as favorable environments, appropriate and early clinical, medical, and psychosocial care, the promotion of healthy behaviors, etc, can also influence the frequency of the genetic variations. Ultimately, if a genetic variation influences a disease, which affects fecundity, then that variation will be selected against. The factors that mediate the frequency of variations in the population clearly influence their presence in the genomes of individuals born in the future.

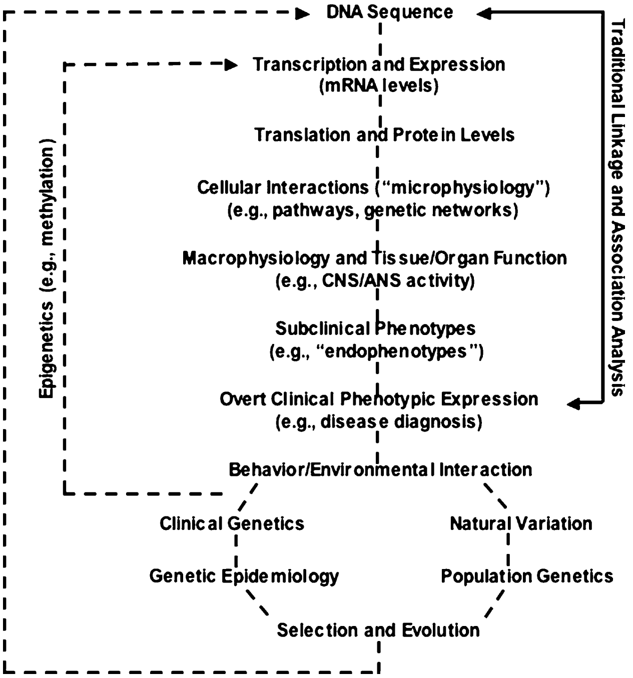
DNA Sequence
Transcription and Expression
(mRNA levels)
Translation and Protein Levels
Cellular Interactions ("microphysiology")
(e.g., pathways, genetic networks)
Macrophysiology and Tissue/Organ Function
(e.g., CNS/ANS activity)
Subclinical Phenotypes
(e.g., "endophenotypes")
Overt Clinical Phenotypic Expression
(e.g., disease diagnosis)
Behavior/Environmental Interaction
Clinical Genetics
Natural Variation
Genetic Epidemiology
Population Genetics
Selection and Evolution
Epigenetics (e.g., methylation)
Traditional Linkage and Association Analysis

Fig. 1. Diagrammatic representation of the influence of genetic variations within the physiologic, population genetic, and evolutionary hierarchy. Ultimately, genetic variation has effects at the molecular, physiological, subclinical, clinical, population, and evolutionary levels that are interdependent. Due to epigenetic programming, individual behaviors and interactions with the environment can reshape gene expression. Thus, the relationship of genetic factors to other phenomena is not unidirectional (ie, the dotted line labeled "epigenetics"). Also, traditional linkage and association analyses seek to relate phenotypes observed at the clinical level with genetic variations and avoid having to understand the connections between those genetic variations and phenotypes (solid line "traditional linkage and association").

Epigenetics and Germ Line Manipulations

We note that figure 1 is drawn as though there are distinct, unidirectional causal relationships between the various levels depicted in the hierarchy. This is not necessarily the case, however. For example, recent discoveries concerning "epigenetic" phenomena suggest that environmental and behavioral manipulations can reshape gene expression patterns.[10,11] In addition, germline manipulations and therapies (ie, purposely altering DNA sequence in parents so that they can no longer transmit a deleterious variation to potential offspring) can clearly influence the existence of deleterious DNA sequence variations.[12] Ultimately, figure 1 should be seen as a simple organizational guide and starting point for describing contemporary statistical genetic problems facing researchers interested in understanding the genetic basis of complex neuropsychiatric diseases such as schizophrenia.

Linkage and Association

There are 2 basically complementary strategies for statistically connecting genes and genetic variations with phenotypes, whether disease related or not, that are used routinely by statistical geneticists. These strategies are linkage analysis, which involves samples of related individuals, and association analysis, which can be pursued in families or unrelated individuals. To describe the principles behind linkage and association analysis and some of the phenomena exploited in appropriate statistical genetic models, we refer the reader to figure 2. Figure 2 depicts the alleles (ie, genetic variations) that 2 parents and 2 offsprings in 6 families possess. The mother's genotypes are depicted as the upper left parent and the father's genotypes are depicted as the upper right parent. Individuals that are affected by a disease such as schizophrenia are shaded. The individual chromosomes (or haplotypes) each individual possesses are adjacent to each other with loci (ie, polymorphic sites) running from top to bottom, with the maternally derived chromosomes on the left and paternally derived chromosomes on the right. Genotypes at 3 loci are depicted. The top most locus is assumed to be a microsatellite locus with alleles 123, 125, 127, and 129. The middle locus is a single nucleotide polymorphism (SNP) with alleles A and C. The bottom locus is also a SNP with alleles A and T. Thus, the mother in family 1 has maternally derived chromosome or haplotype 123-A-T and paternally derived chromosome or haplotype 125-C-A with genotypes 123/125, A/C, and T/A at the top, middle, and bottom locus. For our purpose, it is assumed that the T allele at the bottom locus influences the expression of schizophrenia.

Linkage Analysis

Linkage analysis exploits the fact that individuals affected with a disease within the same family have likely received chromosomal material from a common ancestor who possessed a disease-causing DNA sequence variation. This chromosomal material is likely to be *marked* by variations at neighboring loci inherited together with the disease-causing variation. Note that different variations at neighboring loci may track together with the disease in different families. Thus, in family 1 of figure 2, the 123 allele at the top locus tracks along with the disease, whereas in family 2 the 127 allele tracks along with the disease. Thus, linkage analysis exploits "within-family associations" between a marker allele and a disease. When evidence is found that variations at a particular locus appear to be inherited along with a disease phenotype, then an inference can be made that the disease-causing locus is in the vicinity of the genome near the position in the sequence of the marker locus. Obviously, the closer the marker locus is to the disease-causing locus, the more likely variations at the marker locus will be inherited together with the disease-causing variant because recombination is less likely to shuffle the marker locus variations onto a different chromosome during meiosis. This fact can be exploited by geneticists and other researchers to estimate the actual location of a disease-causing variation given patterns of inheritance among family members of marker locus alleles whose positions on the genome are known (for a more technical description of linkage analysis, the reader is referred to the classic reference for human linkage analysis[13]).

Association Analysis

Association analysis essentially exploits the fact that some loci will have variations that reside on chromosomes harboring the disease-causing variation in the population at large. That is, there are likely to be variations at loci that are so near the locus harboring the disease-causing variation that they almost always appear on chromosomes harboring the disease-causing variation in the population at large. Loci that have variations that almost always appear together are said to be in "linkage disequilibrium." Such variations are likely to be observed in affected individuals in different families. Thus, association analysis proper exploits "across family associations." For example, the affected individuals (except one) in figure 2 all have the A allele at the middle locus because it is in linkage disequilibrium with the disease-causing T allele at the bottom locus. Given this fact, association analyses are often pursued with unrelated individuals, such as those with (ie, cases) and those without schizophrenia (ie, controls), as one can simply contrast the frequency of certain variations between the cases and controls and infer that those showing the greatest differences are the most likely to either be causally related to the disease or be in linkage disequilibrium with the disease-causing allele. Obviously, in this light, the best scenario for association analysis is one in which the disease-causing locus is among those genotyped on the sample of individuals studied.

The Transmission-Disequilibrium Test

There is a special form of association and linkage analysis that works with affected individuals and their parents

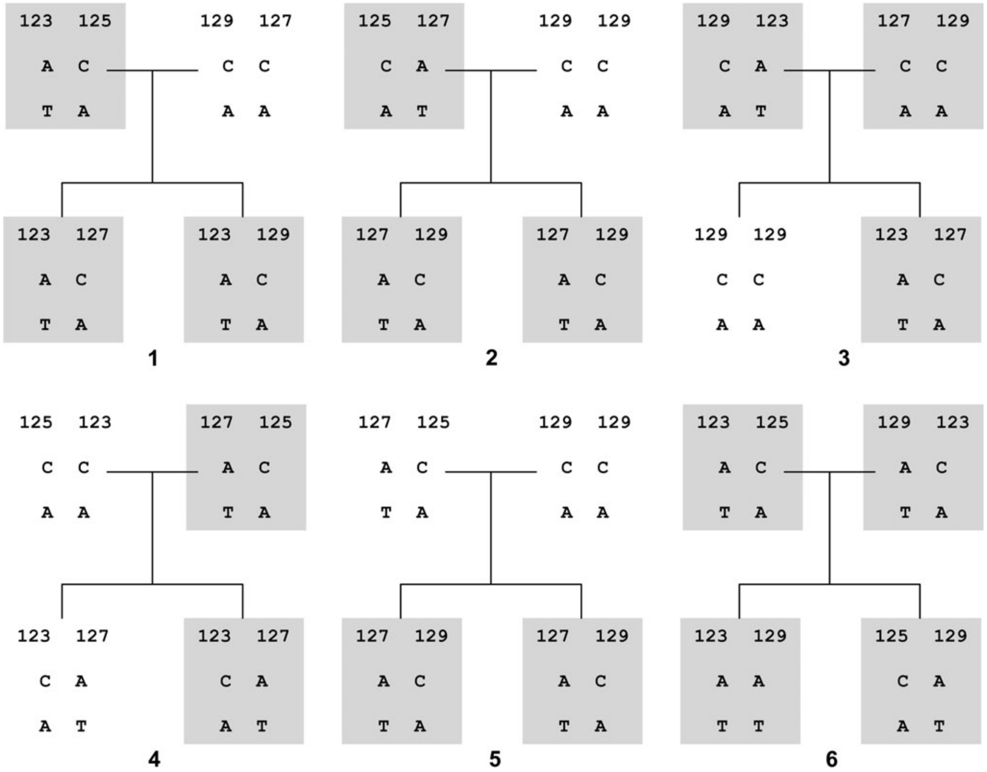

Fig. 2. Diagrammatic representation of the genetic variations at 3 loci possessed by 6 hypothetical sibling pairs and their parents. Shaded individuals manifest a disease phenotype due to the possession of the "T" allele at the bottommost locus. For each individual in a family, the genetic variations are represented at adjacent positions (loci) and should be read from top to bottom. In addition, the 2 alleles that each person possesses at each locus are positioned next to each other with maternally inherited chromosomes on the left and paternally inherited chromosomes on the right. Finally, for each 4-person family, the mother's genotypes are represented as the upper left individual, the father's genotypes as the upper right individual, and the 2 offspring are represented as the lower left and right individuals (see text for further descriptions of this figure).

known as the Transmission-Disequilibrium Test (TDT).[14,15] In this test setting, one assesses the frequency with which a particular allele is transmitted to affected offspring by heterozygous parents. If there is statistically significant evidence indicating that a particular allele has been transmitted more often than one would expect by chance alone from heterozygous parents (which would be 50% by Mendel's laws), then one can infer that the variation must be influencing the disease or is in such close proximity to the disease-causing locus as to be in very strong linkage disequilibrium with it. The advantage of this testing scenario is that one does not need a control (ie, unaffected) group of individuals because, essentially, the allele or variant that is undertransmitted by the heterozygous parents acts as a "control" allele for the allele assumed to be influencing disease susceptibility.

Statistical Complications in Linkage and Association Analysis

There are a number of phenomena that must be considered in linkage and association analyses which often require either very strict assumptions or appropriate accommodation in relevant statistical models. We briefly describe a number of these with reference to figure 2.

Multifactorial, Polygenic Disease Basis

Most neuropsychiatric diseases, such as schizophrenia, are complex and multifactorial in that many genes and environmental factors contribute to their expression. This obviously complicates statistical analysis because any one gene may have its effects obscure the effects of other genes and environmental stimuli (for detailed reviews of the genetic analysis of schizophrenia and related diseases, see Gershon and Badner,[16] Owen et al,[17] Riley et al,[18] Norton,[19] Riley and Kendler[20]). Given that genes work in tandem or in combination through networks, hypotheses have been put forward as to the genetic basis of monogenic diseases (ie, those influenced primarily by perturbations in a single gene) and complex, multifactorial, and/or polygenic diseases that are influenced by many genes.[4] Figure 3 provides a simple graphical representation of a network of genes. It is assumed that the 5 genes on the periphery of the network govern some biochemical and physiologic process that, when perturbed, causes disease. In the figure on the left, the

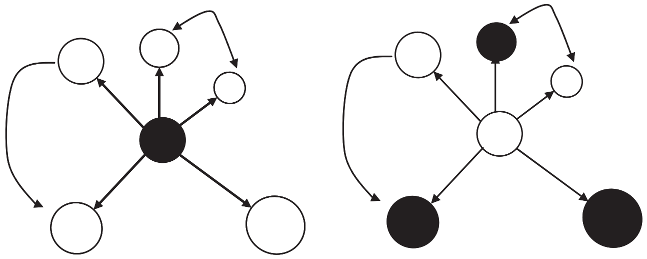

Fig. 3. A simplistic depiction of the possible origins of simple, monogenic, overtly Mendelian diseases and complex, polygenic diseases, considering the fact that genes work in networks. Arrows connect genes that influence each other and may reflect redundancy, feedback, or compensatory mechanisms within the network. In the figure on the left, the gene that plays a more central, "upstream" role in the network has been perturbed and its effects "ripple" throughout the (downstream) genes it mediates and influences. In this way, the entire system is affected by variation in a single gene. In the figure on right, the more central gene is not perturbed, and hence, in order to achieve a deleterious effect given the system governed by the network, all (or many) the downstream genes that are on the periphery of the system must be perturbed.

central or "nodal" gene is perturbed and its effect influences the functioning of the 5 peripheral genes. This scenario is consistent with monogenic disease. In the figure on the right, the more central gene has not been perturbed, so that each (or many) of the genes on the periphery of the network must be perturbed in order for the biochemical or physiologic process to fail. This scenario is consistent with polygenic disease.

Incomplete Penetrance

Given that most neuropsychiatric conditions are complex and polygenic, it is unlikely that every individual carrying a particular DNA sequence variation will manifest schizophrenia. That is, perturbations in these genes may not be sufficient, nor even necessary given heterogeneity (see below), to cause the expression of a disease such as schizophrenia. The term "incomplete penetrance" is used to describe the phenomena in which the mere presence of a specific disease allele is not enough to cause the disease.[13] The leftmost offspring in family 4 and the mother in family 5 both carry the T allele at the disease-causing (bottom) locus but do not have the disease, reflecting the incomplete penetrance of the T allele.

Quantitative Traits

Many diseases and traits are not "either/or" or binary conditions but rather show quantitative variation in the population. In fact, most traits are like this. Consider schizophrenia, depression, and anxiety: they are usually measured in degrees reflecting severity. Modeling quantitative trait expression requires sophisticated constructs in statistical genetic models.

Pleiotropy

The network nature of gene activity also provides a mechanism for a single gene to influence multiple observable phenotypes. The phenomena whereby perturbations in a single gene influence multiple clinical or observable phenotypes is termed "pleiotropy" and is likely to be one of the reasons that, eg, schizophrenia and bipolar disorder are often seen in the same families and may have *common* genetic determinants[21,22] (for an explicit genetic analysis of pleiotropy, see Zhang et al[23]).

Overt Heterogeneity

Many complex diseases may be expressed as a result of different combinations of genetic variations which work independently of other combinations. Thus, it may be the case that none or few of a set of schizophrenia-causing genes are necessary for the expression of the disease phenotype. The father in family 3 of figure 1 manifests the disease but does not carry the T allele at the disease locus, possibly due to heterogeneity (ie, he has the disease because he carries a different disease-causing variation than the T allele at the bottom locus). *Locus heterogeneity* arises when different genes influence a disease independently. *Allelic heterogeneity* arises when different variations within the same gene influence disease susceptibility.

Phenocopies

Individuals who have been diagnosed with a disease but do not carry a known disease-causing genetic variation may reflect the imprecision of the diagnostic instrument used (eg, the *DSM-IV*) and thereby complicate genetic analyses. Such individuals are termed "phenocopies." Differentiating phenocopies due to the use of a less than precise diagnostic or phenotyping instrument from individuals who manifest a disease without a particular genetic variation due to heterogeneity is problematic. The father in family 3 of figure 2 may be a phenocopy because he has been diagnosed with the disease but does not carry the T allele at the disease-causing locus.

Bilineality

When both parents in a family possess a disease-causing variation that can be (or has been) transmitted to their offspring, then the family is termed "bilineal" (eg, family 6 of figure 2). Bilineality can cause problems for statistical genetic analyses because one can not easily trace the inheritance of potential disease variations through a single line of descent.[24] Therefore, eg, the ascertainment scheme used by the Consortium on the Genetics of Schizophrenia excludes families with evidence of bilineal transmission of schizophrenia.[9]

Stratification

One of the most vexing problems in the analysis of case-control–based genetic association studies concerns situations in which the cases are sampled (knowingly or unknowingly) from one population (eg, Australia) and controls are sampled from another (eg, Japan). Because the 2 populations are likely to have very different origins and gene pools, one might observe many different genetic variations providing evidence for association with the disease-bearing individuals (ie, greater frequency in cases), not because of a causal relationship between those variations and the disease but rather because those variations are simply more frequent in the population from which the cases were sampled.[4,25] Although it will rarely be the case that sampling of cases and controls is pursued (consciously) from populations as different as, eg, Australia and Japan, more subtle differences can occur if there is any population "substructure" within the geographic locations from which the individuals have been sampled. The "stratification" problem, as it is known, can be overcome through the use of TDT analysis or the use of clever statistical analysis strategies which assess and control for stratification in an association analysis.[26,27] Ultimately, stratification does not have to occur as an overt genome-wide allele frequency differences between cases and controls but can rather be more cryptic in the sense that many, but not all, cases are sampled from one population, as are the controls, creating subgroups among the cases and controls, that could lead to false-positive (and false negative) results.[28,29]

Admixture Mapping

A special form of stratification or genetic background differences can be exploited in combined linkage and association analyses. Individuals that are admixed (ie, have parents, grandparents, etc, who were from different racial or population subgroups known to differ in allele frequencies at many loci as well as disease rates from the population his or her other parent, grandparent, etc, was from). Some of these admixed individuals will have schizophrenia because they have been transmitted a genetic variation that is more likely to have emanated from a parent, grandparent, etc, of a particular subgroup. These individuals can be compared with unaffected individuals to see which regions of the genome or alleles the diseased individuals have in common that are more frequent in the population with the higher disease rates. The idea is that those shared genomic regions and alleles are likely to reflect the variations that contribute to the higher disease rate in the one population and hence are responsible for the disease in the affected subjects.[30,31]

Epistasis and Gene × Environment Interactions

Modeling and testing gene × gene and gene × environment interactions, if such interactions contribute to disease susceptibility, can be daunting, given the number of potential combinations that can be tested. Despite this fact, recent articles have shown that, in certain instances, such testing can be quite powerful and informative.[32–34] It has also been shown that, despite the large number of tests that would be performed, the analyses of 2 or 3 locus interactions can result in statistical significant results.[32]

Parametric vs Nonparametric Tests

Geneticists often make assumptions about the mode of inheritance of a trait or disease (eg, it is caused by a dominant allele that is fully penetrant) and then incorporate these assumptions into appropriate statistical models. This type of analysis assumes some "parametric" form (ie, the values of certain parameters, such as penetrance and allele frequency, are assumed). Nonparametric statistical genetic analyses do not require as many assumptions. For example, the classic affected sibling pair design in linkage analysis settings merely assesses the degree to which affected siblings (eg, figure 2' families 1, 2, 5, and 6) share alleles in a manner that cannot be attributed to chance. Nonparametric tests are notoriously "underpowered" (ie, they require huge sample sizes in order to detect an effect). Parametric analyses, on the other hand, obviously, assume that one has incorporated the correct values of certain parameters in the model, which can be hard to know a priori. The distinction between parametric and nonparametric models is most pronounced in linkage analysis settings, as opposed to association analysis settings, because linkage analysis modeling of the relationship between allele sharing and phenotypic similarity is more complex and subtle than association analysis modeling of the relationship between particular variations and a phenotypes.

Multiple Comparisons and False-Positive Results

When there is no a priori reason to believe that variations in a particular gene contribute to disease susceptibility, researchers are forced to sequentially test hundreds to millions of variations for association or linkage with a trait. Multiple testing of this sort creates enormous potential for false positives if very stringent criteria for declaring statistical significance are not used. Although many guidelines and methods for assessing statistical significance have been proposed for both linkage and association studies,[35,36] more work is needed in this area, especially in the context of assessing the biological significance of a potential association. One particularly useful strategy for accommodating multiple comparisons involves the notion of the "false discovery rate" (FDR).[37–39] The FDR is used to assess the probability that a large number of statistical tests have produced some test statistics or *P* values that are not likely to have occurred by chance given the number of tests performed.

Modeling the Influence of Genetic Variation

Linkage and association analyses have been pursued for virtually every neuropsychiatric condition of contemporary importance. Historically, most of these analyses have considered the relationship between DNA sequence variation and overt, clinical phenotypes such as the diagnosis of schizophrenia (see figure 2). However, modern statistical geneticists, armed with insights provided by many novel molecular phenotyping technologies and evolutionary studies, are now considering the pursuit of linkage and association studies at all levels of the "physiological hierarchy" (figure 2), and each of these activities presents its own set of statistical analysis challenges. In the following, we describe some of these statistical genetic challenges by working through figure 2˒ starting with DNA sequence variation.

Haplotyping

Modern genotyping technologies typically only provide which combination of alleles an individual possesses at a particular locus (ie, their genotype) and not which alleles at adjacent loci have been transmitted together on maternally derived and paternally derived inherited chromosomes. Thus, analysis methods for assigning "phase" (ie, which alleles were transmitted together on a parental chromosome) are crucial for many genetic analysis. Salem et al[40] provide a comprehensive review of haplotyping methods and resources. Hennah et al[41] consider haplotype analysis in a large study of schizophrenia. As has been previously discussed, haplotyping is an important analytical strategy because association studies using haplotypes are more powerful than single-locus tests under certain circumstances.[42–44]

Phenotype Issues

Mapping Expression Quantitative Trait Loci and Protein Quantitative Trait Loci

Genetic variation that has some physiologic or phenotypic effect, such as neurocognitive or neurophysiological dysfunctions,[3,45,46] clearly must influence the expression or structure of the protein encoded by the gene in question. Thus, the most basic phenotypes are those associated with, eg, the expression levels of a gene, the amount of protein produced, the structure of the encoded protein, etc. Researchers have begun to pursue linkage and association analyses aimed at the identification of genetic variations that influence the expression levels of genes (termed "expression quantitative trait loci" [eQTLs]) as well as the amount of protein produced (termed "protein quantitative trait loci" or [pQTLs]). These studies are typically pursued using microarray technologies that can interrogate the expression levels of thousands of genes or proteins simultaneously, which creates enormous multiple comparisons problems, because a researcher may test each of thousands of genetic variations for association with thousands of gene expression levels. In addition, much of this research has been pursued on accessible human tissues, immortalized cell lines, or model organisms.[47,48] It is important to note that, even though one may be able to identify genetic variations that contribute to the regulatory circuitry or the network genes and proteins that participate in, eg, the cortico-stimato-pallido-thalamic neural circuit that mediates prepulse inhibition (PPI) in mammals,[3,49,50] this does not necessarily suggest how such variation might be of relevance to human disease susceptibility. Finally, one very important concern with the identification of DNA sequence variations that influence the expression or protein levels associated with a particular gene has to do with the tissue from which the expression or protein levels have been assayed. Obviously, assaying tissues such as brain from living humans is problematic and prevents appropriate analysis in all but very special circumstances.

Imaging for Subclinical Phenotyping

Capturing genetically mediated physiological processes and phenotypes at the micro and macro levels that are of relevance to schizophrenia and other neuropsychiatric disease is difficult at best, given the complexities surrounding cognitive processes[45] and the measurement of neurophysiological functions[46] in the brains of living individuals. However, recently developed and extended imaging technologies have the potential to overcome some of these obstacles and have been applied in genetic studies of neuropsychiatric diseases. For example, Ohnishi and colleagues[51] recently examined morphological features of schizophrenic patients' brains that could have a genetic component using magnetic resonance imaging (MRI) technologies. Ho and colleagues used a combination of MRI and positron emission tomography or "PET" technologies to consider brain blood flow differences between schizophrenia and nonschizophrenic patients that may have a genetic basis;[52] and Raemaekers et al used functional MRI (ie, fMRI) to investigate the genetic basis of differences in activation patterns in schizophrenia and nonschizophrenic subjects' brains during cognitive challenges.[53] The combination of phenotypes derived from imaging technologies and genetic linkage and association studies creates statistical genetic problems not unlike those discussed in the context of eQTL mapping. Because tens of thousands of "voxels" (or activation points) might be assessed on the brain, any subsets may be defective and show association with particular genetic variations. These problems are receiving attention, however, among the statistical genetics community and will likely receive greater attention as the technologies are refined (see, eg, http://www.imaginggenetics.uci.edu/index.htm).

Neurocognitive Endophenotypes

Subclinical neurocognitive endophenotypes, such as the PPI of the startle response and working memory, have been studied in schizophrenia.[3,54,55] Many of these endophenotypes have been shown to be heritable and as such may be amenable to genetic association and linkage studies.[3,56,57] One of the biggest issues in the analysis of multiple endophenotypes in schizophrenia research involves pleiotropy and the identification of sets of endophenotypes that appear to be influenced by the same sets of genetic defects and the identification of sets of endophenotypes that may cluster together independently and thereby provide insight into the clinical and etiologic heterogeneity of a disease such as schizophrenia.

Clinical Psychometrics

Psychometric scales have been used for years in neuropsychiatric and psychological research. Typically, a questionnaire containing items of relevance to a diagnosis is administered to a subject. The questions are then converted to a score or scale which provides information that could lead to a diagnosis. Because there are many scales that are used to measure different traits (schizotypy, psychosis proneness, anxiety, depression, etc) that are often given to subjects, it makes sense to analyze them together to look for patterns that may reveal insights into the genetic basis of a disease or phenotype.[58] In addition, it may make sense to analyze psychometric data in a way that considers the individual items (or questions) themselves and not some aggregate score derived from them.

Systems Biology

Given the fact that genetic variation impacts molecular, micro- and macrophysiologic phenomena, endophenotypes, and clinical phenotypes, it is important to consider how one can identify the various connections and relationships these phenotypes and phenomena have to the genetic variations in mediating disease susceptibility. This task has been taken up by practitioners of "systems biology" approaches to multiparameter biological systems and disease pathology. For example, geneticists may consider assessing the aggregate impact of genetic variations in genes known to be involved in the same biochemical network or pathway on a clinical phenotype or disease using system biology-like approaches.[59] This analysis approach can also be used to make sense of associations involving different genes and other experimental results via metaanalyses.[60–62]

Gene × Environment Interactions and Population Risk

Identifying genetic variations that are associated or linked with schizophrenia and other neuropsychiatric disorders is not the end of statistical genetic analyses. Once genes have been identified as associated with a particular condition, researchers can seek to quantify their contribution to disease susceptibility in the population at large as well as their interactions with environmental factors that mediate disease outcomes. This kind of research is in the realm of clinical and genetic epidemiology and applied population genetics.[33,34,63] Relevant statistical genetic analyses are complicated and rarely pursued primarily due to cost reasons. For example, if one really wanted to estimate the "risk" of developing a disease given that an individual carries a certain genetic variation, then one would have to study individuals' pre- and postdisease manifestation in order to determine those "rates" at which carries develop disease from a nondisease state. Longitudinal studies of this sort are extremely costly. However, their value and need has been recognized to the point that position articles in leading journals have been published justifying their pursuit.[64]

Evolutionary Analysis

Obvious questions arise as to the origins of disease-causing genetic variations that are of interest to statistical geneticists. Although much of this inquiry can go well beyond studies seeking to relate variations with actual diseases in order to discover disease-causing variations, there are aspects of evolutionary analyses that can greatly facilitate linkage and association studies. For example, one can consider the evolution or phylogeny of the chromosomes harboring disease-causing variations within the human species.[65] Such analyses can shed light on groups of haplotypes or chromosomes that should be analyzed together in, eg, an association analysis setting.[66,67]

Conclusions

The research field of statistical genetics in neuropsychiatric and other disorders is likely to grow in the next few decades for a number of reasons. First, as this article has tried to make clear, there are myriad issues that one must consider when evaluating the evidence that a particular genetic variation influences the expression of a particular phenotype. Given that most neuropsychiatric diseases are complex and multifactorial, these issues are pronounced, suggesting the need for more powerful and appropriate statistical genetic analysis models and tools. Second, the mere derivation of a statistical genetic analysis model is not sufficient to answer many key questions because relevant data must be painstakingly collected in order for the analysis model to be implemented and used.[9] Thus, efficient study designs for collecting relevant data are needed. This review did not focus on study design, but rather on the mechanics behind statistical genetic analysis, even though robust study designs are absolutely crucial for reducing costs and permitting valid inferences to be drawn for relevant data. Third, genomic resources that statistical geneticists can take advantage of are increasing exponentially. This includes publicly available databases harboring information on genetic variations (such as the International HapMap database; http://www.hapmap.org) and more efficient DNA sequencing and genotyping technologies, molecular genetic technologies, such as DNA microarrays, and phenotyping instruments. In this light, it is fair to say that genetics researchers' ability to generate and collate data is much further ahead than the ability to analyze and draw compelling inferences from those data, as we enter a new era of utilizing endophenotypes and other new information in order to understand the genetic basis of schizophrenia and other complex neuropsychiatric disorders.

Acknowledgments

Nicholas J. Schork and his laboratory are supported in part by the following research grants: The National Heart, Lung, and Blood Institute Family Blood Pressure Program (U01 HL064777-06); the National Institute of Aging Longevity Consortium (U19 AG023122-01); the National Institute of Mental Health Consortium on the Genetics of Schizophrenia (NIMHCOGS; 5 R01 HLMH065571-02); National Institutes of Health R01s: HL074730-02 and HL070137-01; and the Donald W. Reynolds Foundation (Helen Hobbs, Principal Investigator). David L. Braff and his laboratory are supported by the following research grants: the NIMHCOGS (R01 MH065571), the NIMH Translational Study (R01 MH042228), the National Alliance for Research on Schizophrenia and Depression, Bowman Family Foundation, and the Niederhoffer Family Foundation.

References

1. International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437:1299–1320.
2. Cowan WM, Kopnisky KL, Hyman SE. The human genome project and its impact on psychiatry. *Annu Rev Neurosci*. 2002;25:1–50.
3. Braff DL, Freedman R, Schork NJ, Gottesman II. Deconstructing schizophrenia: an overview of the use of endophenotypes in order to understand a complex disorder. *Schizophr Bull*. In press.
4. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994;265:2037–2048. [Erratum in: *Science*. 1994;266: 353.]
5. Jarvik LF, Deckard BS. The Odyssean personality. A survival advantage for carriers of genes predisposing to schizophrenia? *Neuropsychobiology*. 1977;3:179–191.
6. Koob GF, Swerdlow NR. The functional output of the mesolimbic dopamine system. *Ann N Y Acad Sci*. 1998;537:216–227.
7. Light GA, Braff DL. Mismatch negativity deficits are associated with poor functioning in schizophrenia patients. *Arch Gen Psychiatry*. 2005;62:127–136.
8. Cadenhead KS, Perry W, Shafer K, Braff DL. Cognitive functions in schizotypal personality disorder. *Schizophr Res*. 1999;37:123–132.
9. Calkins ME, Dobie DJ, Cadenhead KS, et al. The Consortium on the Genetics of Endophenotypes in Schizophrenia (COGS): model recruitment, assessment, and endophenotyping methods for a multi-site collaboration. *Schizophr Bull*. In press.
10. Callinan PA, Feinberg AP. The emerging science of epigenomics. *Hum Mol Genet*. 2006;15:R95–R101.
11. Feil R. Environmental and nutritional effects on the epigenetic regulation of genes. *Mutat Res*. 2006;600:46–57.
12. Smith KR. Gene therapy: the potential applicability of gene transfer technology to the human germline. *Int J Med Sci*. 2004;1:76–91.
13. Ott J. Revised edition. *Analysis of Human Genetic Linkage*. Baltimore, MD: Johns Hopkins University Press; 1991.
14. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*. 1993;52:506–516.
15. Spielman RS, McGinnis RE, Ewens WJ. The transmission/disequilibrium test detects cosegregation and linkage. *Am J Hum Genet*. 1994;54:559–560.
16. Gershon ES, Badner JA. Progress toward discovery of susceptibility genes for bipolar manic-depressive illness and schizophrenia. *CNS Spectr*. 2001;6:965–968977.
17. Owen MJ, Williams NM, O'Donovan MC. The molecular genetics of schizophrenia: new findings promise new insights. *Mol Psychiatry*. 2004;9:14–27.
18. Riley B. Linkage studies of schizophrenia. *Neurotox Res*. 2004;6:17–34.
19. Norton N, Williams HJ, Owen MJ. An update on the genetics of schizophrenia. *Curr Opin Psychiatry*. 2006;19:158–164.
20. Riley B, Kendler KS. Molecular genetic studies of schizophrenia. *Eur J Hum Genet*. 2006;14:669–680.
21. Kelsoe JR, Spence MA, Loetscher E, et al. A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. *Proc Natl Acad Sci U S A*. 2001;98:585–590.
22. Abdolmaleky HM, Thiagalingam S, Wilcox M. Genetics and epigenetics in major psychiatric disorders: dilemmas, achievements, applications, and future scope. *Am J Pharmacogenomics*. 2005;5:149–160.
23. Zhang L, Rao F, Wessel J, et al. Functional allelic heterogeneity and pleiotropy of a repeat polymorphism in tyrosine hydroxylase: prediction of catecholamines and response to stress in twins. *Physiol Genomics*. 2004;19:277–291.
24. Hodge SE. Do bilineal pedigrees represent a problem for linkage analysis? Basic principles and simulation results for single-gene diseases with no heterogeneity. *Genet Epidemiol*. 1992;9:191–206.
25. Gardner M, Gonzalez-Neira A, Lao O, Calafell F, Bertranpetit J, Comas D. Extreme population differences across Neuregulin 1 gene, with implications for association studies. *Mol Psychiatry*. 2006;11:66–75.
26. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol*. 2001; 60:227–237.
27. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*. 2001;60:155–166.
28. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004;3:512–517.

29. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet*. 2004;36:388–393.
30. Nievergelt CM, Schork NJ. Admixture mapping as a gene discovery approach for complex human traits and diseases. *Curr Hypertens Rep*. 2006;7:31–37.
31. DeLisi LE, Mesen A, Rodriguez C, et al. Genome-wide scan for linkage to schizophrenia in a Spanish-origin cohort from Costa Rica. *Am J Med Genet*. 2002;114:497–508.
32. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*. 2005;37:413–417.
33. De Luca V, Tharmalingam S, Muller DJ, Wong G, de Bartolomeis A, Kennedy JL. Gene-gene interaction between MAOA and COMT in suicidal behavior: analysis in schizophrenia. *Brain Res*. 2006;1097:26–30.
34. Caspi A, Moffitt TE, Cannon M, et al. Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene × environment interaction. *Biol Psychiatry*. 2005;57:1117–1127.
35. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*. 1995;11:241–247.
36. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 2003;100:9440–9445.
37. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med*. 1990;9:811–818.
38. Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*. 2005;21:781–787.
39. Taylor J, Tibshirani R. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*. 2006;7:167–181.
40. Salem RM, Wessel J, Schork NJ. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum Genomics*. 2005;2:39–66.
41. Hennah W, Varilo T, Paunio T, Peltonen L. Haplotype analysis and identification of genes for a complex trait: examples from schizophrenia. *Ann Med*. 2004;36:322–331.
42. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res*. 2001;11:143–151.
43. Longmate JA. Complexity and power in case-control association studies. *Am J Hum Genet*. 2001;68:1229–1237.
44. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet*. 2001;9:291–300.
45. Gur RE, Calkins ME, Gur RC. The Consortium on the Genetics of Schizophrenia (COGS): neurocognitive endophenotypes. *Schizophr Bull*. In press.
46. Turetsky BI, Calkins ME, Light GA. Neurophysiological endophenotypes of schizophrenia: the viability of selected candidate measures. *Schizophr Bull*. In press.
47. Monks SA, Leonardson A, Zhu H. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*. 2004;75:1094–1105.
48. de Koning DJ, Haley CS. Genetical genomics in humans and model organisms. *Trends Genet*. 2005;21:377–381.
49. Swerdlow NR, Koob GF. Lesions of the dorsomedial nucleus of the thalamus, medial prefrontal cortex and pedunculopontine nucleus: effects on locomotor activity mediated by nucleus accumbens-ventral pallidal circuitry. *Brain Res*. 1987;412:233–243.
50. Braff DL, Geyer MA, Swerdlow NR. Human studies of prepulse inhibition of startle: normal subjects, patient groups, and pharmacological studies. *Psychopharmacology (Berl)*. 2001;156:234–258.
51. Ohnishi T, Hashimoto R, Mori T. The association between the Val158Met polymorphism of the catechol-O-methyl transferase gene and morphological abnormalities of the brain in chronic schizophrenia. *Brain*. 129:399–410.
52. Ho BC, Wassink TH, O'Leary DS, Sheffield VC, Andreasen NC. Catechol-O-methyl transferase Val158Met gene polymorphism in schizophrenia: working memory, frontal lobe MRI morphology and frontal cerebral blood flow. *Mol Psychiatry*. 2005;10:229, 287–298.
53. Raemaekers M, Ramsey NF, Vink M, van den Heuvel MP, Kahn RS. Brain activation during antisaccades in unaffected relatives of schizophrenic patients. *Biol Psychiatry*. 2006;59:530–535.
54. GottesmanII, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry*. 2003;160:636–645.
55. Jablensky A. Subtyping schizophrenia: implications for genetic research. *Mol Psychiatry*. 2006;11:815–836.
56. Keri S, Janka Z. Critical evaluation of cognitive dysfunctions as endophenotypes of schizophrenia. *Acta Psychiatr Scand*. 2004;110:83–91.
57. Hall MH, Schulze K, Rijsdijk F. Heritability and reliability of P300, P50 and duration mismatch negativity. *Behav Genet*. In press.
58. Niculescu AB, Lulow LL, Ogden CA, et al. PhenoChipping of psychotic disorders: a novel approach for deconstructing and quantitating psychiatric phenotypes. *Am J Med Genet B (Neuropsychiatr Genet)*. 2006;141:653–662.
59. Jamshidi N, Palsson BO. Systems biology of SNPs. *Mol Syst Biol*. 2006;2:38.
60. Ogden CA, Rich ME, Schork NJ, et al. Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: an expanded convergent functional genomics approach. *Mol Psychiatry*. 2004;9:1007–1029.
61. Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*. 2006;78:1011–1025.
62. Carter CJ. Schizophrenia susceptibility genes converge on interlinked pathways related to glutamatergic transmission and long-term potentiation, oxidative stress and oligodendrocyte viability. *Schizophr Res*. 2006;86:1–14.
63. Carter JW, Schulsinger F, Parnas J, Cannon T, Mednick SA. A multivariate prediction model of schizophrenia. *Schizophr Bull*. 2002;28:649–682.
64. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature*. 2004;429:475–477.
65. Schork NJ, Thiel B, St Jean P. Linkage analysis, kinship, and the short-term evolution of chromosomes. *J Exp Zoolog*. 1998;282:133–149.
66. Templeton AR. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics*. 1995;140:403–409.
67. Seltman H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. *Genet Epidemiol*. 2003;25:48–58.