CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Elsevier - Publisher Connector

Cell
Leading Edge
Essay

Genetic Heterogeneity in Human Disease

Jon McClellan[1,*] and Mary-Claire King[2,*]
[1]Department of Psychiatry
[2]Departments of Medicine and Genome Sciences
University of Washington, Seattle, WA 98195-7720, USA
*Correspondence: drjack@uw.edu (J.M.), mcking@uw.edu (M.-C.K.)
DOI 10.1016/j.cell.2010.03.032

Strong evidence suggests that rare mutations of severe effect are responsible for a substantial portion of complex human disease. Evolutionary forces generate vast genetic heterogeneity in human illness by introducing many new variants in each generation. Current sequencing technologies offer the possibility of finding rare disease-causing mutations and the genes that harbor them.

"Every unhappy family is unhappy in its own way," wrote Tolstoy in Anna Karenina. Tolstoy was reflecting on the individually unique nature of human tragedy. We suggest that this principle also captures the misfortune of human disease. That is, from the perspective of genetics, we suggest that complex human disease is in fact a large collection of individually rare, even private, conditions. Disentangling the paradoxes embedded in this idea will be the subject of this Essay.

In molecular terms, we suggest that human disease is characterized by marked genetic heterogeneity, far greater than previously appreciated. Converging evidence for a wide range of common diseases indicates that heterogeneity is important at multiple levels of causation: (1) individually rare mutations collectively play a substantial role in causing complex illnesses; (2) the same gene may harbor many (hundreds or even thousands) different rare severe mutations in unrelated affected individuals; (3) the same mutation may lead to different clinical manifestations (phenotypes) in different individuals; and (4) mutations in different genes in the same or related pathways may lead to the same disorder.

This degree of allelic, locus, and phenotypic heterogeneity has important implications for gene discovery. In particular, causality in this context can almost never be resolved by large-scale association or case-control studies. This degree of heterogeneity also has important implications for development of molecular treatments and their appropriate use by individual patients.

An Evolutionary Perspective

The genetic bases of disease in modern humans reflect the architecture and evolution of the human genome. The oldest human alleles originated in Africa, in parallel with the development of our species, millions of years before people first migrated out of Africa 50,000 to 60,000 years ago (Cavalli-Sforza et al., 1994; Cavalli-Sforza and Feldman, 2003) (Figure 1A). These ancient polymorphisms are shared by all human populations and account for approximately 90% of human variation. These polymorphisms provided the single-nucleotide polymorphisms (SNPs) of the HapMap (Tishkoff and Verrelli, 2003; The International HapMap Consortium, 2007).

Yet new alleles constantly arise, at an estimated rate of approximately 175 per diploid human genome per generation (Nachman and Crowell, 2000) (Figure 1B). Exponential population growth, fueled by the development of agriculture in the past 10,000 years and of urbanization in the past 700 years, has resulted in a vast number of new alleles. Collectively, these alleles have generated an immense degree of genetic variation. Given the size of the present day human population, every point mutation compatible with life is likely present in someone, somewhere. Many of these rare alleles are found in only one person or family. Thus the paradox: most human variation is ancient and shared, but most alleles are recent and rare.

Whole-genome sequencing efforts have revealed millions of previously unreported variants in healthy individuals, including single base pair substitutions, small insertions and deletions, and larger copy number mutations (McKernan et al., 2009). Most of these mutations likely have no functional significance and persist by chance in the absence of selective pressure. In contrast, mutations with deleterious effects, either on viability before reproduction or on fertility, are less frequently transmitted to subsequent generations. By this reasoning, it should be no surprise that severe mutations associated with early-onset disease are disproportionately of recent origin and are therefore individually very rare.

Common Disease—Many Rare Alleles

Rare large-effect mutations are now recognized as causes of many different common medical conditions. A thorough review of the rare variant literature for human disease is well beyond the scope of this commentary. Here we briefly consider a few very different complex disorders for which rare mutations have been implicated, drawing in part from our own research.

Inherited Predisposition to Breast Cancer

Rare severe alleles have been implicated in all forms of inherited susceptibility to cancer. Inherited predisposition to breast cancer is associated with germline mutations in at least 13 genes (Walsh and King, 2007) (Figure 2A). Thousands of different loss-of-function mutations have been detected in *BRCA1* and *BRCA2* (Figure 2B); all of these mutations are individually rare, and each independently

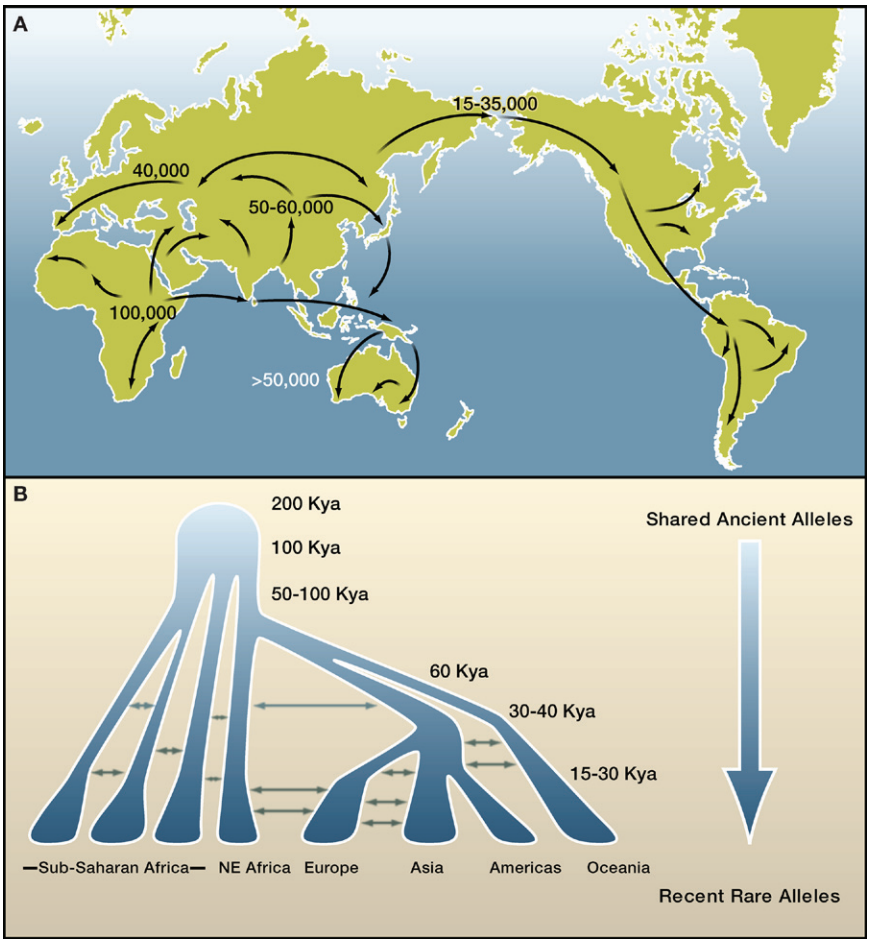

Figure 1. Human Migrations and Genetic Diversity

(A) The oldest human alleles originated in Africa well before the diasporas of modern humans 50,000–60,000 years ago. These oldest alleles are common in all populations worldwide. Approximately 90% of the variability in allele frequencies is of this sort. (Figure adapted from Cavalli-Sforza and Feldman, 2003.)

(B) Origins of common and rare alleles. KYA refers to "thousand years ago." Horizontal arrows suggest continuing cross-migration between continental populations. Development of agriculture in the past 10,000 years and of urbanization and industrialization in the past 700 years has led to rapid populations growth and therefore to the appearance of vast numbers of new alleles, each individually rare and specific to one population or even to one family. (Figure adapted from Tishkoff and Verrelli, 2003.)

confers very high risks for breast and ovarian cancers. Rare germline mutations in *p53*, *PTEN*, *CHEK2*, *PALB2*, *ATM*, *BRIP1*, *CDH1*, and *STK1* are also associated with increased risk of breast cancer, ranging from 2-fold for *CHEK2* to 10-fold for *p53*. Each of these genes operates in networks integral to DNA repair and genomic integrity. For each gene, the combination of inherited and somatic loss-of-function mutations ("2 hits") leads to errors of DNA repair and ultimately to tumor development. In addition, biallelic germline loss-of-function mutations in *BRCA2*, *BRIP1*, and *PALB2* cause various forms of Fanconi Anemia, all of which are characterized by genomic instability and early-onset cancers.

Genetic heterogeneity of inherited predisposition to breast cancer serves as a model for other complex illnesses. The disorder results from any one of thousands of different mutations in any one of multiple different genes, but all the implicated genes encode proteins in related pathways (Walsh and King, 2007). We suggest that the degree of biological complexity underlying a phenotype is an excellent predictor of locus heterogeneity. We further suggest that age at onset and severity of the illness are excellent predictors of allelic heterogeneity at each locus.

Inherited Hearing Loss

Dozens of genes harbor inherited mutations leading to nonsyndromic hearing loss, and each gene harbors multiple pathogenic mutations (Dror and Avraham, 2009). Depending on the gene and mutation, inheritance of the hearing loss may be dominant or recessive and autosomal, X-linked, or mitochondrial. All deafness-causing mutations are recent and all but one are individually rare. The one frequent mutation, 30 delG in connexin 26, is the exception that proves the rule: it is not a common shared ancestral variant but rather has occurred independently multiple times in a mutational hotspot. Genes responsible for hearing loss encode proteins involved in a wide variety of biological processes in the inner ear, including development and maintenance of cytoskeletal structures, myosin motors, gap junction transport and signaling, ion channels, and transcriptional regulators (Figure 3). Mutations in microRNA have also been identified in hearing loss. The theme is the same as for inherited predisposition to cancer: any one of many different mutations in any one of many different genes lead to related phenotypes.

Genetics of Lipid Metabolism

Individually rare variants in genes related to lipid metabolism are associated with extreme levels of high-density lipoprotein cholesterol (HDL-C) (Cohen et al., 2004) and low-density lipoprotein cholesterol (LDL-C) (Cohen et al., 2006). Although each variant only explains a small number of cases, collectively these mutations are responsible for a substantial portion of metabolic disease. The design for identifying these alleles was ingenious: candidate genes, defined as those causing rare recessive metabolic disorders, were sequenced in genomic DNA of individuals without overt disease but with extreme values of the trait of interest, i.e., levels of HDL or LDL cholesterol. This proved a powerful and efficient strategy for discovering causal variants associated with complex genetic traits (Fahmi et al., 2008).

Genetics of Severe Mental Illness

Emerging research suggests that rare mutations play important roles in neuropsychiatric disorders, including autism

and schizophrenia (Figure 4). A substantial portion of autism appears to be caused by rare point mutations, deletions, duplications, and larger chromosomal abnormalities (Bucan et al., 2009). Autism is characterized by rare structural mutations, including a disproportionately high rate of de novo large (>100 kb) deletions and duplications (Sebat et al., 2007). Large structural mutations that reoccur independently multiple times in genomic "hotspots," including on chromosomes 1q21.1, 15q11-q13, 16p11.2, and 22q11.21, may each explain small subsets of cases (Bucan et al., 2009). Copy number mutations in patients with autism may disproportionally disrupt genes involved with neuronal cell-adhesion or ubiquitin degradation, both important networks for brain development (Glessner et al., 2009). In addition, rare severe mutations in multiple genes important for brain development, including *NRXN1*, *CNTN4*, *CNTNAP2*, *NLGN4*, *DPP10*, and *SHANK3*, have been identified in patients with autism spectrum disorders (Guilmatre et al., 2009).

Similar results have been observed for schizophrenia. We found that compared to healthy controls, individuals with schizophrenia were 3-fold more likely to harbor rare structural genomic mutations that impacted genes (Walsh et al., 2008). The risk was even higher in subjects with early onset of schizophrenia. Each rare mutation disrupted a different gene or genes. However, the disrupted genes were not a random sample of the genome. Genes disrupted in patients were disproportionally involved with signaling and neurodevelopment (e.g., neuregulin and glutamate pathways), whereas genes disrupted in controls did not cluster in any particular pathways. Several independent studies have replicated and extended our findings. Compared to controls, individuals with schizophrenia carried significantly more de novo structural mutations (Xu et al., 2008) and more structural mutations at genomic hotspots, including chromosomes 1q21.1, 15q13.3, 16p13.1, and 22q11.2 (Stefansson et al., 2008; International Schizophrenia Consortium, 2008).

These results suggest that a substantial portion of autism and schizophrenia is caused by individually rare mutations—small and large—that disrupt the function

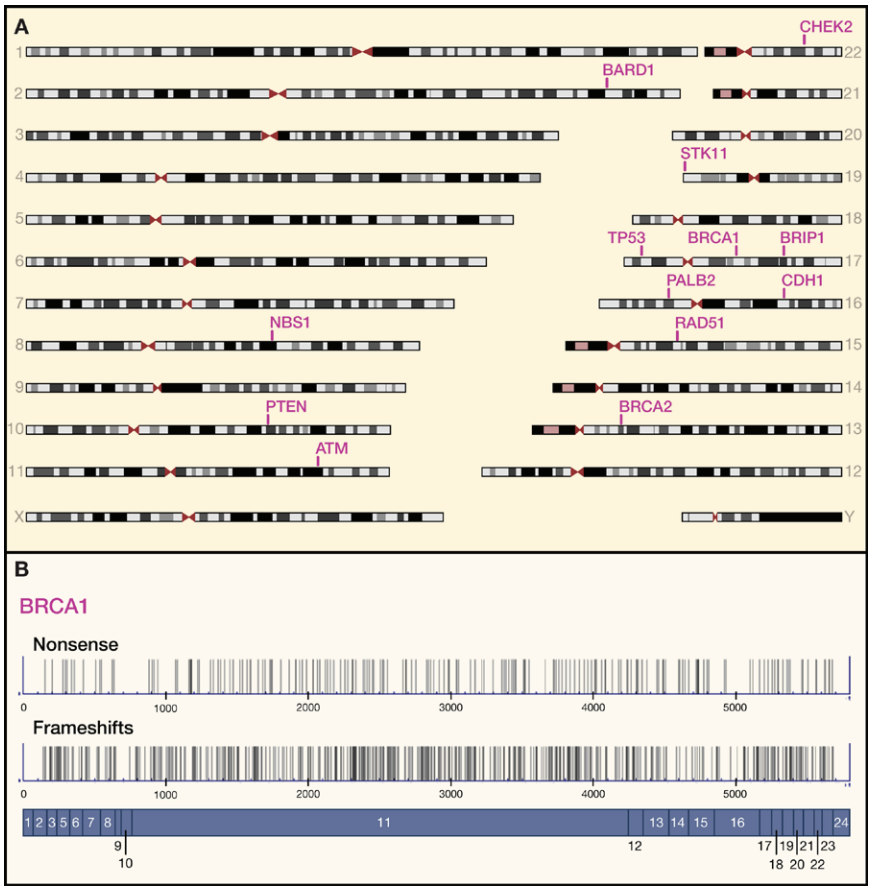

Figure 2. Genetic Heterogeneity of Inherited Predisposition to Breast Cancer

(A) Thus far, 13 genes have been identified that harbor loss-of-function mutations responsible for inherited susceptibility to human breast cancer. Each susceptible woman carries an inherited mutation in only one (or very occasionally, two) of these genes. *BRCA1* and *BRCA2* are the most frequently altered of these genes. Inherited mutations in *TP53* and *PTEN* are fortunately very rare and lead to young onset breast cancer in the context of Li-Fraumeni syndrome and Cowden disease, respectively. Most of the genes illustrated encode proteins critical to the maintenance of genomic integrity. The large number of high-incidence breast cancer families with no mutations in any of these genes suggests that additional genes remain to be found.

(B) At *BRCA1*, more than 1000 different alleles increase susceptibility to breast and ovarian cancers. The figure shows positions of frameshift and nonsense mutations along the *BRCA1* gene, which is in blue, with exons indicated. Among mutation carriers, cancer development is caused by the combination of the inherited mutation and a subsequent somatic mutation, leading to complete loss of *BRCA1* function in the affected tissue.

of genes operating in critical neurodevelopmental pathways. Several genomic hotspots have been implicated in more than one psychiatric or neurocognitive phenotype (Cook and Scherer, 2008). The converse is also true: the same mutation may be associated with different psychiatric disorders or with no illness at all. For example, in the Scottish kindred harboring a chromosomal translocation disrupting *DISC1*, 29 translocation carriers presented variously with schizophrenia, bipolar disorder, major depressive disorder, or no mental illness (Chubb et al., 2008). Similarly, there are several genes, including *NRXN1*, *PRODH*, *DOC2A*, and *CNTNAP2*, for which rare deleterious mutations have been detected in individuals with autism, mental retardation, and schizophrenia (Guilmatre et al., 2009).

Thus, mutations in genes involved with brain development may lead to a wide range of pleiotropic effects. Conversely, each neuropsychiatric outcome may be influenced by a variety of interacting factors, including dose and timing of gene expression, epistasis, RNA regulatory elements, epigenetic effects, and environmental exposures. In principle, there are thousands of candidate genes for

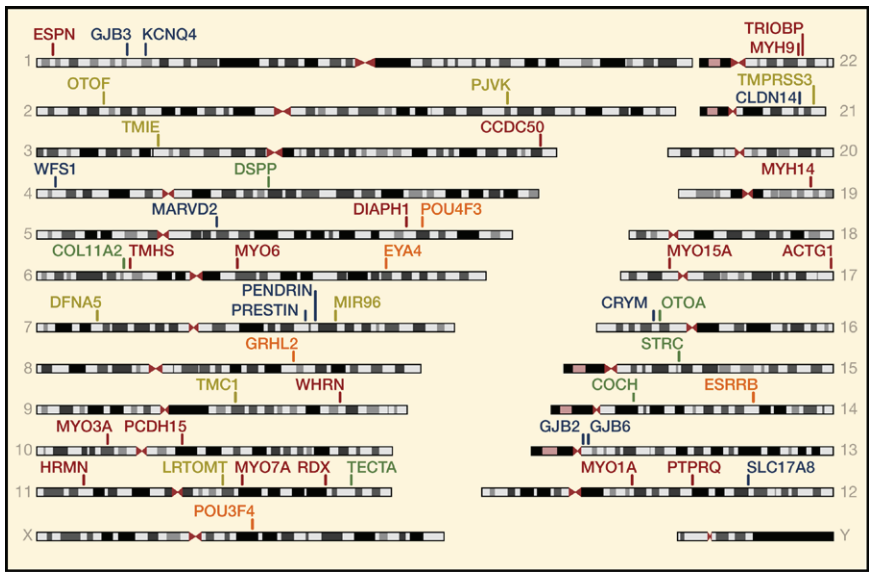

Figure 3. Genetic Heterogeneity of Inherited Hearing Loss

Thus far, 48 genes have been identified that harbor mutations responsible for inherited hearing loss. The genes encode proteins for a wide range of functions in the inner ear: hair bundle morphogenesis, including cytoskeleton, adhesion, scaffolding, and motor proteins (red); ion homeostasis, including connexins, ion channels, and tight junctions (blue); extracellular matrix proteins (green), transcription factors (orange), and proteins whose function in hearing is not yet known (black). Every gene includes multiple deafness-associated mutations. *GJB2*, which encodes a gap junction protein, is the most frequently altered of these genes, although all mutations are individually rare. Each affected individual carries one dominant or two recessive mutations in only one gene.

these illnesses, given that most human genes are expressed in brain. Many more genes responsible for neurological and psychiatric disorders will be identified as sequencing technologies improve and the cost of screening individual genomes falls. Characterization of the functional consequences of these mutations will help to elucidate both normal brain development and neurodevelopmental processes leading to disease (Geschwind, 2008).

Common Disease—Common Variants

The recognition that rare alleles are important contributors to common complex human diseases is a major paradigm shift in human genetics. In the past decade, until this shift, the search for disease-associated genes has been predicated almost entirely on the common disease-common variant model, which postulates that common illnesses stem from the additive or multiplicative effects of combinations of common variants (Risch and Merikangas, 1996). In this model, each "risk variant" is postulated to confer only a small degree of risk, with no one variant sufficient to cause the disorder. Disease onset is postulated as the result of the combined effects of many such alleles. The model dates to Francis Galton (1872) and early population genetics theory (Fisher, 1918; Wright, 1934; Falconer, 1965) and with recent technology could be tested directly.

For each common disease, "risk variants" have been identified by comparing allele frequencies at hundreds of thousands of polymorphic sites (SNPs) in thousands of cases versus thousands of controls. To date, genome-wide association studies (GWAS) have published hundreds of common variants whose allele frequencies are statistically correlated with various illnesses and traits. However, the vast majority of such variants have no established biological relevance to disease or clinical utility for prognosis or treatment. For example, a recently published 12 year follow-up study of cardiovascular disease (CVD) in more than 19,000 women found that the 101 SNPs identified by GWAS as risk variants for CVD did not predict cardiovascular outcomes (Paynter et al., 2010). More generally, it is now clear that common risk variants fail to explain the vast majority of genetic heritability for any human disease, either individually or collectively (Manolio et al., 2009).

Do common variants ever make a major contribution to disease? Of course. Sickle cell anemia and the thalassemias are caused by multiple mutations in hemoglobin genes that persist at polymorphic frequencies in malarial endemic regions worldwide (Patrinos et al., 2005). Autoimmune conditions, such as systemic lupus erythematosis, multiple sclerosis, type I diabetes, and rheumatoid arthritis, are strongly influenced by common polymorphic variation at the histocompatibility (MHC) loci (Fernando et al., 2008). Alzheimer's disease is strongly influenced by an allele of *APOE4* that occurs at polymorphic frequencies in most populations (Bird, 2005). Lactose intolerance (or lactase persistence) is caused by the effect of any one of several different alleles in noncoding enhancers of the lactase promoter; different regulatory alleles are common in different populations (Tishkoff et al., 2007). Pharmacogenomics holds immediate promise for personalized medicine, for example by genotyping *CYP2C9* and *VKORC1* to improve the safety of warfarin (Gurwitz and Motulsky, 2007).

These traits and the common alleles influencing them illustrate two important points. First, each common risk variant leads to a demonstrable functional difference in the protein it encodes or regulates. In contrast, the majority of variants detected by GWAS have no demonstrated biological significance. Second, the persistence of these common alleles reflects evolutionary forces: deleterious alleles in hemoglobins were maintained by selection for heterozygotes driven by malaria (sickle cell anemia and the thalassemias); geographic-specific variation in immune response is protective against geographic-specific infection (autoimmune conditions); the illness has no effect on fitness because it appears well after reproductive life (Alzheimer's disease); a medication introduces a new environmental challenge not previously under evolutionary pressure (drug response).

The genetic architecture of a disorder, i.e., the number and frequency of susceptibility alleles, is shaped by evolution-

ary factors, including the impact of the illness on selection (Pritchard and Cox, 2002). In order to be maintained at polymorphic frequencies worldwide, common variants with even modest influence on disease must withstand selective pressure in every generation. Not surprisingly, therefore, the common alleles with the best documented relationship to disease are associated with disorders that arise later in life, i.e., Alzheimer disease's or age-related macular degeneration. For illnesses that impact reproductive fitness, balancing positive selection is often demonstrable. Illness in these cases may arise from interaction between genetic and environmental factors, such that an otherwise adaptive mechanism or trait is deleterious in certain individuals. For example, adaptive inflammatory responses can cause autoimmune disorders when turned against the host, or efficient storage of calories can lead to type II diabetes or to obesity in food-rich cultures.

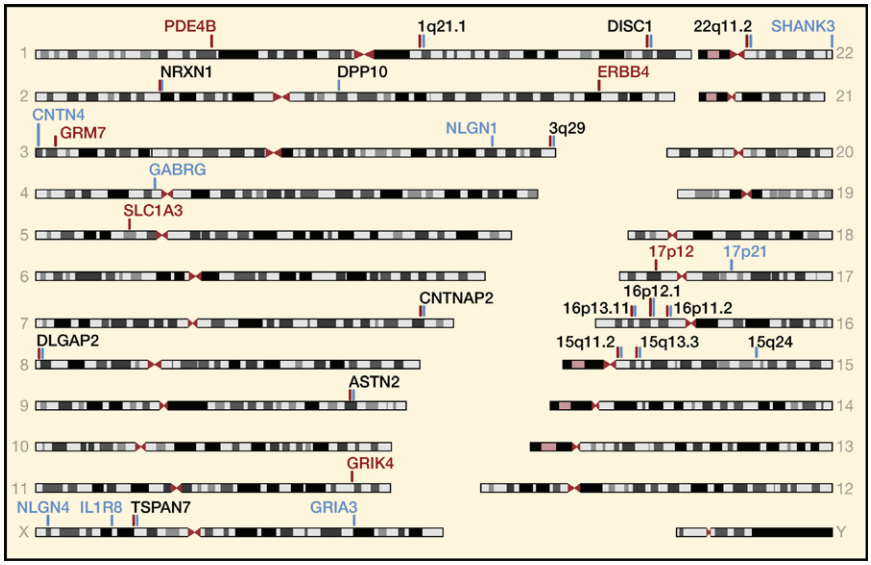

Figure 4. Genetic Heterogeneity of Severe Mental Illness
Recent genomic analyses have revealed many individually rare, or even de novo, micro-deletions, micro-duplications, and point mutations associated with schizophrenia (red), autism (blue), or both (black). The most frequently replicated genes and loci include *DISC1*, *NRXN1* (neurexin), *CNTNAP2*, and *SHANK3*, as well as genomic hotspots at 1q21.1, 15q13.3, 16p11.2, and 22q11.2.

Both common and rare alleles may lead to the same disease. For example, multiple rare mutations in any one of several genes (e.g., *APP*, *PS1*, *PS2*, and *UBQLN1)* lead to early-onset Alzheimer's disease (Bird, 2005). Rare mutations in genes involved in immune response (e.g., *DNASEI*, *TREX1*) confer a very high risk for lupus (Moser et al., 2009). Each of these conditions is characterized by allele and locus heterogeneity. However, it should not be anticipated that all complex diseases will have substantial contributions from common risk variants. This is especially true for illnesses that reduce fertility, either biologically or through social selection against marriage with individuals with severe diseases. Alleles of significant effect must be enabled by evolutionary forces to persist at polymorphic frequencies.

Genome-wide Association Studies

It has become commonplace in the genetics community to aver that genome-wide association studies have led to the identification of hundreds of SNPs as "risk variants" for common diseases, while acknowledging that most of the heritability for these traits remains to be explained (Manolio et al., 2009). We agree with this conclusion: that most of the inherited basis of common traits remains to be explained (Goldstein, 2009). We further suggest that many GWAS findings stem from factors other than a true association with disease risk. The bases of our concern are both statistical and experimental.

Cryptic Population Stratification

A major limitation complicating genome-wide association studies is the potential for cryptic population stratification. Subtle differences in ancestry between cases and controls can produce spurious association solely due to sampling. GWAS study designs typically control for population structure in two ways: by taking into account differences among populations in average allele frequencies and by excluding individual subjects with extreme outlier genotypes (e.g., Price et al., 2006; Purcell et al., 2007). These approaches are appropriate and necessary but not sufficient. Neither adjustment addresses the problem posed by individual SNPs that are outliers with respect to variation in allele frequencies among healthy populations. A hypervariable SNP may be falsely identified as a risk variant if cases and controls are not perfectly matched. Stratifying cases and controls by self-reported ethnicity is not nearly sufficient to control for this problem (Serre et al., 2008).

A recently published genome-wide association study of autism (Wang et al., 2009) provides a particularly dramatic example of the perils of cryptic population stratification. The SNP most significantly associated with the illness was rs4307059 on chromosome 5p14.1. The ancestral T allele of rs4307059 was identified as the "risk variant," with allele frequencies in the discovery series of 0.65 among cases and 0.61 among controls (odds ratio = 1.19, $p = 3.4 \times 10^{-8}$). All cases and controls were of European ancestry. The p value is compelling. However, the frequency of the proposed "risk variant" varies from 0.21 to 0.77 across European populations (Coop et al., 2009), a range 14-fold greater than the difference in allele frequencies between cases and controls. Even a subtle difference in the ancestries of cases and controls within Europe could explain the difference in allele frequencies that was attributed to the autism phenotype. An odds ratio of 3.0, or even of 2.0 depending on population allele frequencies, would be robust to such population stratification. However, odds ratios of the magnitude generally detected by GWAS (<1.5) can frequently be explained by cryptic population stratification, regardless of the p value associated with them. The variation in allele frequencies at rs4307059 is

particularly striking worldwide: in virtually all African populations, the frequency of the "risk variant" is 1.00. A subsequent analysis did not replicate the association between rs4307059 and autism (Weiss et al., 2009) and in fact found the opposite trend: the minor allele was more frequent among affected persons.

In large trans-continental studies of human populations, there is almost never perfect ancestral matching of cases and controls. To the extent that variants with greatest variability in frequency among control populations disproportionately emerge as significant findings in genome-wide association studies, population stratification may explain apparent disease associations. Replication studies, based on still larger samples, are generally suggested as the resolution of this problem. However, depending on sampling strategies for cases and controls in each geographic locale, larger samples may replicate, rather than resolve, imperfect matching.

Investigators devote a great deal of effort to the problem of population stratification. Subjects who are deemed outliers based on substructure analysis are generally removed from GWAS. However, hypervariable polymorphisms remain vulnerable to stratification even after this adjustment. Strategies to address this problem include using family designs to compare genotypes of cases to their healthy relatives and removing hypervariable SNPs from analyses.

Function: The Definition of "in" a Gene

A major limitation of genome-wide association studies is the lack of any functional link between the vast majority of risk variants and the disorders they putatively influence. It has become common practice to describe risk variants derived from GWAS as "in" a gene, suggesting that the gene harboring the variant influences the disorder. But "in" in this context has a purely physical meaning: that the risk variant lies somewhere in a genomic locus that also includes a gene. In the human genome, approximately 35% of base pairs lie in introns, and therefore approximately the same proportion of SNPs lie "in" genes. In this context, "in" is a tautology, not a proof of biological relevance. Very few published risk variants lie in coding regions, in UTRs, in promoters, or even in predicted intronic or intergenic regulatory regions. Far fewer have been shown to alter the function of any of these sequences.

How did genome-wide association studies come to be populated by risk variants with no known function? Standard genotyping platforms are designed to screen common SNPs, which were selected to be the most variable among individuals. As described above, evolutionary forces have led to most common variation being neutral. Given that common variants are surveyed, it should not be surprising if most reported associations are neutral.

Search for Meaning in GWAS: Linkage Disequilibrium

Two approaches have been used to demonstrate the biological importance of risk variants detected in GWAS. One hypothesis is that a risk variant is not itself a critical functional variant, but is in linkage disequilibrium, in a subset of cases, with a rarer mutation of clear functional effect. The principle is that linkage disequilibrium (or LD) of risk variants with rare mutations of functional effect leads to statistical associations in genome-wide association studies. The hypothesis is reasonable if genetic heterogeneity of the disease is very low in the series of cases under study. That is, a significant association in a GWAS may reflect a functional mutation by LD *if* the (unknown) functional mutation is responsible for a substantial proportion of the illness in the cases surveyed.

For example, a GWAS of sickle cell anemia among African Americans yielded a meaningful association for an SNP near the beta hemoglobin gene *HBB* (Dickson et al., 2010) because most cases of sickle cell anemia among African Americans are caused by the same mutation in *HBB*. Selection over the past several thousand years in favor of heterozygosity for the *HBB-S* mutation has maintained the allele at common frequency in West Africa and hence an elevated prevalence among Americans with West African ancestry. Similarly, a GWAS of inherited hearing loss among European American children yielded a meaningful association for an SNP near the gap junction protein *GJB2* (Dickson et al., 2010) because the same mutation in *GJB2* is responsible for most cases of inherited hearing loss among European American children.

In both of these examples, in the population studied, the condition is not characterized by marked genetic heterogeneity. Had sickle cell anemia been investigated among affected individuals worldwide, the number of responsible mutations would be far greater and hence no one allele at any SNP would be consistently associated with the disease. Similarly, had inherited hearing loss been investigated in a region where it is more common (e.g., in the Middle East), many different genes and an even larger number of causal mutations would be involved. Thus no one allele at any SNP would be consistently associated with the condition. In geographically isolated populations, localized disease homogeneity has led to successful applications of the GWAS approach (Kristiansson et al., 2008). The approach is far less likely to be successful in studies in heterogeneous populations of common diseases that involve multiple disease genes, each harboring multiple rare mutations, in LD with different alleles of neighboring SNPs.

Search for Meaning in GWAS: Distant Regulators

The second approach to demonstrating the biological importance of risk variants detected in GWAS is to assess whether a risk variant regulates a gene at considerable genomic distance from the locale of the SNP. It is well established that mutations in noncoding sequence far removed from coding sequence can alter gene expression and lead to a severe phenotype. For example, disruption of an enhancer site more than 1 MB from the sonic hedgehog gene *Shh* leads to polydactyly in mouse (Lettice et al., 2002). Multiple different mutations in the coding sequence of *Shh* in humans can cause the comparable polydactyly phenotype as well as other far more severe patterning aberrations (Quinlan et al., 2008). Among risk variants detected by GWAS, an example that may prove parallel is SNP rs6983267, which has been identified in several studies as associated with colon cancer. rs6983267 lies in a gene desert, 335 kb from the oncogene *MYC*. Despite the considerable genomic distance, alternate alleles of rs6983267

regulate expression of MYC by 1.5-fold (Tuupanen et al., 2009; Pomerantz et al., 2009). However, there is no difference in *MYC* expression in colon tumors based on the rs6983267 genotype.

The general failure to confirm common risk variants is not due to a failure to carry out GWAS properly. The problem is underlying biology, not the operationalization of study design. The common disease—common variant model has been the primary focus of human genomics over the last decade. Numerous international collaborative efforts representing hundreds of important human diseases and traits have been carried out with large well-characterized cohorts of cases and controls. If common alleles influenced common diseases, many would have been found by now. The issue is not how to develop still larger studies, or how to parse the data still further, but rather whether the common disease—common variant hypothesis has now been tested and found not to apply to most complex human diseases.

A Time to Sequence—With an Appreciation to Maynard Olson (Olson, 1995)

Genetic factors contributing to human disease are subject to the same evolutionary forces that dictate the architecture of the human genome. The overall magnitude of human genetic variation, the high rate of de novo mutation, the range of mutational mechanisms that disrupt gene function, and the complexity of biological processes underlying pathophysiology all predict a substantial role for rare severe mutations in complex human disease. Furthermore, these factors explain why efforts to identify meaningful common risk variants are vexed by irreproducible and biologically ambiguous results.

Genome-wide screening for mutations remains the most effective and unbiased way to discover genes involved in complex illnesses. Heretofore, the identification of rare severe disease-causing variants was limited by the resolution of mutation detection strategies. The widespread availability of next-generation sequencing technology renders this limitation essentially moot. Designs based on genome-wide identification of *all* exonic variants, *all* variants in a defined genomic region, or even *all* variants in a whole genome are replacing genome-wide association approaches. However, although the power of sequencing is enormous, genetic heterogeneity remains a daunting challenge. With next-generation sequencing technology, the issue is not finding potentially deleterious mutations but rather determining which of many potential deleterious mutations in an individual play a role in disease.

Two powerful strategies for identifying critical mutations are (1) tracing coinheritance of potential disease alleles with the illness in severely affected families, and (2) identifying different rare functional mutations in the same gene in unrelated affected individuals. Experience with well-characterized disease genes indicates that the first mutations discovered are generally those with severe effects, and that subsequent characterization of the gene's mutational spectrum reveals additional point mutations and small frameshifts with severe effect, as well as hypomorphic alleles (e.g., Cohen et al., 2004, 2006), regulatory mutations, and genomically cryptic mutation sites (e.g., Walsh and King, 2007; Dror and Avraham, 2009).

New sequencing technologies provide conceptual and practical advantages over current approaches (Olson, 1995). Given genetic heterogeneity of complex traits, large numbers of affected individuals and families are necessary to identify different disease-causing variants. With complete exome or full genome sequencing of each case, the concept of replication shifts. As described in this Essay, a powerful method of replication in genetics is the identification of different biologically relevant mutations in the same gene. This important concept now expands to include the identification of functional mutations in genes acting in related biological pathways. The mutational spectrum of each gene can be established using existing large cohorts. Any one gene may be responsible for only a small fraction of cases. The occurrence of multiple functional mutations among unrelated cases provides both biological evidence and epidemiological support for the causal role of the gene or pathway.

Next-generation sequencing provides its own challenges. Whole-genome sequencing strategies detect hundreds of thousands of rare variants per individual (McKernan et al., 2009). Biological relevance must be established before a mutation can be causally linked to a disorder. The critical question is not whether cases as a group have more rare events than controls; but rather which mutation(s) disrupting a gene is responsible for the illness in the affected person harboring the variant. Variable penetrance, epistasis, epigenetic changes, and gene-environment interactions will complicate these efforts. It will be fun to sort out.

The ultimate goal of gene discovery in complex disease is to identify and characterize biological pathways and processes critical to the disorder. Key pathways may be disrupted via many different causes—genetic, epigenetic, and environmental. Even if the illness in every affected individual arises from a different specific cause, each will nonetheless share disruption of related key biological processes. Defining the ways in which biological networks for common disease are impacted by mutation will contribute substantially to the understanding of their pathology and provide important targets for intervention.

REFERENCES

Bird, T.D. (2005). N. Engl. J. Med. *352*, 862–864.

Bucan, M., Abrahams, B.S., Wang, K., Glessner, J.T., Herman, E.I., Sonnenblick, L.I., Alvarez Retuerto, A.I., Imielinski, M., Hadley, D., Bradfield, J.P., et al. (2009). PLoS Genet. *5*, e1000536.

Cavalli-Sforza, L.L., and Feldman, M.W. (2003). Nat. Genet. *33*, 266–275.

Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). History and Geography of Human Genes (Princeton, NJ: Princeton University Press).

Chubb, J.E., Bradshaw, N.J., Soares, D.C., Porteous, D.J., and Millar, J.K. (2008). Mol. Psychiatry *13*, 36–64.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Science *305*, 869–872.

Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Proc. Natl. Acad. Sci. USA *103*, 1810–1815.

Cook, E.H., Jr., and Scherer, S.W. (2008). Nature *455*, 919–923.

Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W., and Pritchard, J.K. (2009). PLoS Genet. *5*, e1000500.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). PLoS Biol. *8*, e1000294.

Dror, A.A., and Avraham, K.B. (2009). Annu. Rev. Genet. *43*, 411–437.

Fahmi, S., Yang, C., Esmail, S., Hobbs, H.H., and Cohen, J.C. (2008). Hum. Mol. Genet. *17*, 2101–2107.

Falconer, D.S. (1965). Ann. Hum. Genet. *29*, 51–76.

Fernando, M.M., Stevens, C.R., Walsh, E.C., De-Jager, P.L., Goyette, P., Plenge, R.M., Vyse, T.J., and Rioux, J.D. (2008). PLoS Genet. *4*, e1000024.

Fisher, R.A. (1918). Trans. R. Soc. Edinb. *52*, 399–433.

Galton, F. (1872). Proc. R. Soc. Lond. *20*, 394–402.

Geschwind, D.H. (2008). Cell *135*, 391–395.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., et al. (2009). Nature *459*, 569–573.

Goldstein, D.B. (2009). N. Engl. J. Med. *360*, 1696–1698.

Guilmatre, A., Dubourg, C., Mosca, A.L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., Layet, V., Rosier, A., Briault, S., Bonnet-Brilhault, F., et al. (2009). Arch. Gen. Psychiatry *66*, 947–956.

Gurwitz, D., and Motulsky, A.G. (2007). Pharmacogenomics *8*, 1479–1484.

International Schizophrenia Consortium. (2008). Nature *455*, 237–241.

Kristiansson, K., Naukkarinen, J., and Peltonen, L. (2008). Genome Biol. *9*, 109.

Lettice, L.A., Horikoshi, T., Heaney, S.J., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., et al. (2002). Proc. Natl. Acad. Sci. USA *99*, 7548–7553.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Nature *461*, 747–753.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., et al. (2009). Genome Res. *19*, 1527–1541.

Moser, K.L., Kelly, J.A., Lessard, C.J., and Harley, J.B. (2009). Genes Immun. *10*, 373–379.

Nachman, M.W., and Crowell, S.L. (2000). Genetics *156*, 297–304.

Olson, M.V. (1995). Science *270*, 394–396.

Patrinos, G.P., Kollia, P., and Papadakis, M.N. (2005). Hum. Mutat. *26*, 399–412.

Paynter, N.P., Chasman, D.I., Paré, G., Buring, J.E., Cook, N.R., Miletich, J.P., and Ridker, P.M. (2010). JAMA *303*, 631–637.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M., et al. (2009). Nat. Genet. *41*, 882–884.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Nat. Genet. *38*, 904–909.

Pritchard, J.K., and Cox, N.J. (2002). Hum. Mol. Genet. *11*, 2417–2423.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). Am. J. Hum. Genet. *81*, 559–575.

Quinlan, R.J., Tobin, J.L., and Beales, P.L. (2008). Curr. Top. Dev. Biol. *84*, 249–310.

Risch, N., and Merikangas, K. (1996). Science *273*, 1516–1517.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Science *316*, 445–449.

Serre, D., Montpetit, A., Paré, G., Engert, J.C., Yusuf, S., Keavney, B., Hudson, T.J., and Anand, S. (2008). PLoS ONE *3*, e1382.

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., et al. (2008). Nature *455*, 232–236.

The International HapMap Consortium. (2007). Nature *449*, 851–861.

Tishkoff, S.A., and Verrelli, B.C. (2003). Annu. Rev. Genomics Hum. Genet. *4*, 293–340.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Nat. Genet. *39*, 31–40.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymaki, I., et al. (2009). Nat. Genet. *41*, 885–890.

Walsh, T., and King, M.-C. (2007). Cancer Cell *11*, 103–105.

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., et al. (2008). Science *320*, 539–543.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M., et al. (2009). Nature *459*, 528–533.

Weiss, L.A., Arking, D.E., Gene Discovery Project of Johns Hopkins and the Autism Consortium, Daly, M.J., and Chakravarti, A. (2009). Nature *461*, 802–888.

Wright, S. (1934). Genetics *19*, 537–551.

Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A., and Karayiorgou, M. (2008). Nat. Genet. *40*, 880–885.