Successes of Genome-wide Association Studies

In a recent Essay in *Cell*, McClellan and King argue that genomic resequencing rather than genome-wide association studies (GWAS) will be necessary to understand the genetic basis of common disease (McClellan and King, 2010). Like the authors, we too are excited about the potential for emerging sequencing technologies to facilitate discoveries that explain the missing heritability of common diseases. However, we disagree with the implication that GWAS have not been successful to date. Instead, we propose that insofar as the goal of these studies is to understand the etiology of heritable diseases, GWAS have provided numerous tantalizing clues for us biologists to decipher. Rather than disprove the common disease/common variant hypothesis, we find that results from GWAS support the contention that common polymorphisms do directly contribute to disease risk, validating the linkage disequilibrium-based GWAS approach for helping to identify variants underlying disease. Although we do not dismiss the likelihood that rare variants also contribute to common diseases, we expect that whole-genome sequencing approaches will show that the full spectrum of alleles, from rare to common, play important roles in disease etiology. Here, we argue that the existence of common disease-causing polymorphisms is not inconsistent with population genetic theory and that actual results from GWAS suggest that the reported associations represent real biology rather than false positives.

The contention that deleterious alleles that cause human diseases are common in the population may seem paradoxical, but several mechanisms can explain how such pathogenic alleles can overcome negative selective pressure. First, accumulating evidence demonstrates that there is balancing selection in which a certain allele confers susceptibility to one disease while simultaneously conferring protection from another. The best known example is heterozygosity for sickle cell anemia, which affords protection against malaria. GWAS have identified other such instances, for example, the *TCF2* (or *HNF1B*) gene where alternate alleles are risk factors for type 2 diabetes and prostate cancer (Gudmundsson et al., 2007). More generally, several loci where alleles have opposite effects on the risk of developing type 1 diabetes and Crohn's disease have been reported (Wang et al., 2010), and it is likely that more examples of balancing selection are yet to be discovered.

The argument that common pathogenic variants must have withstood selective pressure throughout human history is predicated on the assumption that modern humans developed in the same environment that we exist in today. However, due to the rapid acceleration of human development in the recent evolutionary timeframe, numerous environmental changes have occurred that may impact the risk of complex common diseases. For instance, variants that were beneficial in the past may well have turned against their carriers as human lifestyles changed. The "thrifty gene hypothesis" suggests that variants that predispose to type 2 diabetes and obesity may have conferred a selective advantage in times of famine (Neel, 1962). However, in developed countries, where food tends to be in overabundance, type 2 diabetes and obesity have become common diseases. Furthermore, existing neutral variation may manifest positive or negative effects as new environmental modifiers come into play. A set of single-nucleotide polymorphisms (SNPs) associated with lung cancer and located at a locus encoding the nicotinic acetylcholine receptor appears to have a stronger effect on lung cancer risk in smokers born long ago relative to those born more recently (Landi et al., 2009); this effect has been attributed to changes in the composition of cigarettes over time. This demonstrates that recent environmental changes can alter the disease-influencing effect of a common variant. These examples are likely to be only the tip of the iceberg of phenotypic effects modulated by gene-environment interactions.

The fact that most SNPs identified by GWAS do not lie in coding regions or other known regulatory elements is expected from the study design and is not evidence of false positives. The assumption underlying the design of GWAS and choice of genotyped SNPs is that the true functional allele will be nearby and correlated with the initial SNP through linkage disequilibrium. When one considers linkage disequilibrium, there is an observed excess of GWAS hits that influence promoter regions or change the protein-coding sequence of a gene and a relative paucity of hits in intergenic regions (Hindorff et al., 2009). Moreover, many disease-associated SNPs identified by GWAS are located in genes or pathways previously known or suspected to play a role in disease etiology. Recent GWAS for Alzheimer's disease, Crohn's disease, type 1 diabetes, and type 2 diabetes have rediscovered SNP associations previously reported from candidate gene studies. The demonstration that GWAS can identify common disease susceptibility variants provides a positive control (Hindorff et al., 2009). More generally, the functional pathways of GWAS-identified genes often make sense; for instance, numerous inflammatory genes have been implicated by GWAS in inflammatory bowel disease (Hindorff et al., 2009). We have found a similar concordance between the literature and GWAS hits in our own work. In a GWAS for age-related macular degeneration, an SNP in complement factor H strongly associates with disease risk; this is consistent with previous suggestions that the complement pathway plays a role in disease etiology (Klein et al., 2005). Similarly, using GWAS, we identified a germline variant in the intron of the *JAK2* gene that is associated with myeloproliferative neoplasms; *JAK2* is known to harbor activating oncogenic somatic mutations in this disease (Kilpivaara et al., 2009). One would not expect such correlations between GWAS findings, genic regions, and known disease biology if these findings were randomly distributed false positives due to population stratification or other causes. As these associations are likely to be real, the most logical and parsimonious explanation is that, in general, GWAS successfully identify disease-associated variants and that variants found through GWAS tag regions important for the biology of these diseases.

In light of this, it is likely that GWAS hits found in intergenic regions far from known genes are true associations whose biology is not yet understood, rather than false positives. The human genome is incompletely annotated. Regions where GWAS associations have been found, but no known genes are located, could easily harbor unidentified new genes or regulatory elements. For instance, the authors point to the colon and prostate cancer risk SNP rs693267 located 335 kb upstream from the *MYC* gene on chromosome 8q24 (McClellan and King, 2010). This locus has been shown to physically interact with *MYC* and is associated with enhanced Wnt signaling. Therefore, although the biology of this locus is not fully understood, it suggests a paradigm where intergenic disease-associated SNPs alter enhancer elements, either directly or through linkage disequilibrium, and therefore cause differential regulation of disease-related genes.

This observation leads to a broader point: A lack of biological understanding of how these disease-associated variants are pathogenic does not mean that there is no biology to discover. Although our understanding of the mechanisms by which disease risk loci contribute to pathogenesis currently lags behind the pace at which new loci are discovered, promising stories continue to emerge. To continue the previous example, although no definitive correlation between the rs6983267 genotype and *MYC* expression has been demonstrated, *MYC* is known to be tightly regulated and the right developmental time point may need to be examined to see such a correlation. Although further work is necessary to uncover the elusive mechanism by which the SNP confers risk, we propose that the existing evidence supports rather than refutes this SNP as a true cancer risk allele. Another example is a non-protein-coding region of chromosome 9q21 in which SNPs have been robustly associated with arterial disease. A recent paper reported that targeted deletion of an orthologous region in mouse interferes with *cis*-regulation of nearby genes (*Cdkn2a/Cdkn2b*) and may influence vascular cell proliferation (Visel et al., 2010). As a third example, an intronic type 2 diabetes risk SNP (rs7903146) was recently found to overlap with a region of islet cell-selective chromatin, and the two alleles of rs7903146 correlate with the open/closed chromatin state of the region (Gaulton et al., 2010). Thus, understanding the mechanisms by which GWAS loci contribute to disease will require considerable effort and time. We take this not as a sign that the common disease-common variant model has failed but rather that a challenge exists for the scientific community—a challenge that must be addressed with both traditional experimental genetics and innovative new approaches.

Robert J. Klein,[1,*] Xing Xu,[1] Semanti Mukherjee,[1] Jason Willis,[1] and James Hayes[1]
[1]Program in Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA
*Correspondence: kleinr@mskcc.org
DOI 10.1016/j.cell.2010.07.026

REFERENCES

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). Nat. Genet. *42*, 255–259.

Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J.T., Thorleifsson, G., Manolescu, A., Rafnar, T., Gudbjartsson, D., Agnarsson, B.A., Baker, A., et al. (2007). Nat. Genet. *39*, 977–983.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Kilpivaara, O., Mukherjee, S., Schram, A.M., Wadleigh, M., Mullally, A., Ebert, B.L., Bass, A., Marubayashi, S., Heguy, A., Garcia-Manero, G., et al. (2009). Nat. Genet. *41*, 455–459.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., Sangiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Science *308*, 385–389.

Landi, M.T., Chatterjee, N., Yu, K., Goldin, L.R., Goldstein, A.M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., et al. (2009). Am. J. Hum. Genet. *85*, 679–691.

McClellan, J., and King, M.C. (2010). Cell *141*, 210–217.

Neel, J.V. (1962). Am. J. Hum. Genet. *14*, 353–362.

Visel, A., Zhu, Y., May, D., Afzal, V., Gong, E., Attanasio, C., Blow, M.J., Cohen, J.C., Rubin, E.M., and Pennacchio, L.A. (2010). Nature *464*, 409–412.

Wang, K., Baldassano, R., Zhang, H., Qu, H.Q., Imielinski, M., Kugathasan, S., Annese, V., Dubinsky, M., Rotter, J.I., Russell, R.K., et al. (2010). Hum. Mol. Genet. *19*, 2059–2067.

Strategies for Genetic Studies of Complex Diseases

In a recent Essay published in *Cell*, McClellan and King discussed genetic heterogeneity and the potential role of rare genetic variants in complex human diseases (McClellan and King, 2010). These important issues, in particular the application of high-throughput sequencing techniques to discover disease genes, are highly relevant to genetics researchers. However, the authors allocated a substantial proportion of their efforts to being critical of the utility of genome-wide association studies (GWAS). These particular sections of the Essay may lead to misinterpretation of published studies by us and others. For the broad readership of *Cell* and for the scientific community in general, we highlight our concerns in this Correspondence.

The authors refer to the fact that most single-nucleotide polymorphisms (SNPs) detected in GWAS reside in intergenic regions and consequently challenge the utility and reliability of GWAS with the question: “How did genome-wide association studies come to be populated by