Genome-Wide Association Studies and Cancer

Eric Jorgenson PhD and Iona Cheng PhD, MPH

What is a Genome-Wide Association Study?

Genetic association studies examine the effect of inherited genetic variants on disease traits. For cancer, such traits include differences in the risk of developing cancer, response to therapy, disease progression and mortality. The most common study design involves comparing the frequency of a particular genetic variant in a group of cancer patients (cases) and a group of healthy subjects (controls). What makes genome-wide association studies unique is that they attempt to comprehensively examine all genetic variants in the human genome in one study.

Technology Behind Genome-Wide Association Studies

The power of genome-wide association studies was first recognized in 1996,[1] but it has taken a decade for advances in technology to make these studies possible and affordable. The first step in this process was the completion of the human genome project. This project sequenced the entire human genome and determined that it contains just over 3.2 billion basepairs (the smallest unit of DNA) and approximately 22,000 genes. In addition to the genes themselves, there are thought to be a large number of gene regulatory elements that control the activity of genes. If each gene were a light bulb, the regulatory elements would be light switches that turn the genes on and off. Variation in the genes or the sequences that regulate them can alter the risk of human disease.

The second step that has made genome-wide association studies possible is the identification of genetic variants across the genome. A number of projects that have followed the sequencing of the human genome have identified millions of genetic variants in multiple ethnically diverse populations.[2-4] The most common type of genetic variant is the single nucleotide polymorphism or SNP, which alters a single basepair in the DNA sequence. There are estimated to be more than 10 million SNPs in the human genome.[5] The International HapMap project (http://hapmap.ncbi.nlm.nih.gov) has genotyped and validated nearly 4 million SNPs in several ethnic groups that can be incorporated into genotyping platforms in genome-wide association studies.

The third step is the dramatic technological advances and corresponding decrease in the cost of genotyping these variants as a result of advances in DNA microarray technology. High throughput genotyping platforms that are designed to genotype hundreds of thousands to over two million SNPs are currently available from two companies, Affymetrix and Illumina, with some laboratories capable of genotyping >2,000 subjects per week. As it is possible to genotype these large sets of genetic variants at a low cost of approximately $500 per subject, it is estimated that the cost per genotype has reduced 2,000-fold in the past 10 years.[6] In addition, large populations from previous epidemiologic studies provide a readily available resource of subjects for genome-wide association studies, helping to limit the substantial cost of recruiting and phenotyping a large number of patients.

Assumptions Underlying Genome-Wide Association Studies

The success of genome-wide association studies will depend on several critical assumptions. For a genome-wide association study to detect an association between genetic variant(s) and human diseases, those variants must occur at a high enough frequency and have a strong enough effect on disease. The common disease-common variant (CDCV) hypothesis states that the majority of genetic variants that affect common human diseases will be frequent (or common) in the population and have small to modest effects on disease risk.[7] The rare variant (RV) hypothesis suggests that, because disease causing variants are likely to be disadvantageous to those who carry them and therefore be selected against, most disease variants will be rare with modest to large effects on disease risk. Genome-wide association studies will perform best under the CDCV hypothesis as they have good statistical power to detect genetic variants with small to modest effects as long as they are common.[8] Genome-wide association studies have limited power, however, when the genetic variants underlying disease are rare even when the increases in disease risk and sample size of a study are large.[9]

The second assumption underlying current genome-wide association studies is that the genetic variants that are genotyped on DNA microarrays will serve as proxies for those variants that are not genotyped. This is possible because of a phenomenon called linkage disequilibrium, where genetic variants that are not genotyped on the DNA microarray are captured due to their correlation with variants that are genotyped. Typically, the strongest correlation occurs between markers that are located in close physical proximity, allowing for localization of the causal variant. As a result, only a subset of genetic variants needs to be genotyped to capture the effects of all genetic variants. This type of genome-wide association study is referred to as an indirect association study, because potentially causal variants are examined indirectly through the variants that are genotyped.[10] The recently published results from genome-wide association studies have utilized DNA microarray genotyping platforms for indirect genome-wide association studies, which have consisted of up to 2.5 million SNP markers. The next generation platforms will be able to examine up to five million SNPs, still a considerable reduction from the more than 10 million SNPs thought to exist in the human genome.[11]

Potential Pitfalls in Genome-Wide Association Studies

Because genome-wide association studies rely on these assumptions, there are a number of factors that can decrease their effectiveness to detect genetic variants underlying human disease. For example, genome-wide association studies have much better statistical power to detect genetic variants that are common than those that are rare. For this reason, genotyping platforms have been designed to capture common variants, typically those that have a frequency of 5% or greater in the population under study, by utilizing linkage disequi-

librium. As a result, variants that occur less frequently will not be captured well (i.e., lack coverage), further decreasing the power to capture the effect of these variants on disease.[9] Many important variants may be missed as a result, including some that change the amino acid sequence of genes which often occur at frequencies below 5%.[11] In addition, as the identification of common genetic variants has relied on data from the International HapMap project, which initially focused on African, Asian, and European populations and more recently included Latinos and Native Americans, population-specific variants for less common groups such as Polynesians have yet to be characterized.

A great debate has been ongoing on the importance or self identified race/ethnicity in medical studies. The geographic origin of the population under study has important implications for genome-wide association studies. The extent of linkage disequilibrium in the human genome varies by population. Notably, populations of African descent have lower levels of linkage disequilibrium than those of European or East Asian descent.[12-13] African populations also have more genetic variants, including SNPs.[2-3] For these reasons, a genotyping platform would need to contain hundreds of thousands of additional SNPs to capture the same variants in African descent populations.

In the United States, this issue is further complicated by recent admixture in African-American populations between African and European populations. Studies that are focused on admixed groups need to be wary of the potential for population stratification, which can lead to false positive association signals when the frequency of genetic variants and the frequency of the disease under study differ across populations. For example, the incidence of prostate cancer is higher in African-American populations compared to European populations. Men with a greater proportion of African admixture are also more likely to have genetic variants that are more frequent in African populations. If African-Americans who have prostate cancer also have a greater proportion of African admixture compared to those who do not have prostate cancer, the variants that are more frequent in the ancestral African population can appear to be associated with prostate cancer in an African-American sample when they are not causal. There are a number of methods that are currently being used to address the problem of population stratification.[14-16] Conversely, admixture can also be used to map the location of disease causing variants in admixed populations by identifying a local region of increased or decreased admixture in disease subjects compared to healthy controls. A recent study used this type of admixture mapping to identify a second prostate cancer susceptibility locus on chromosome 8q24 that appears to explain some of the increased risk of prostate cancer in African-Americans.[17]

A final caveat in genome-wide association studies is the issue of allelic heterogeneity, where multiple variants in the same gene increase (or decrease) the risk of the disease under study. These variants are difficult to capture using linkage disequilibrium and so it may be necessary to take an alternative approach to study them. While we have outlined some of the possible pitfalls in genome-wide association studies, many of these issues can be handled using a careful and thoughtful approach to the design and analysis of these studies. In the future, technological improvements in genotyping even more variants at a decreasing cost are likely to make genome-wide association studies even more comprehensive and more powerful.

Genome-Wide Association Studies and Cancer

Over the past three years, a multitude of genome-wide association studies of cancer have identified and replicated numerous genetic variants associated with cancer risk and prognosis. Results from genome-wide association studies are summarized and updated in the National Human Genome Research Institute's "Catalog of Published Genome-Wide Association Studies" (http://www.genome.gov/gwastudies/). As of July 2010, this catalog includes 68 genome-wide association study publications on a wide array of cancer sites such as bladder, breast, colorectal, esophagus, lung, ovary, prostate, skin, and testes. In particular, there have been 11 genome-wide association study reports of prostate cancer, 10 for breast cancer, 10 for lung cancer, and 7 for colorectal cancer with the remaining for other cancer sites. From these genome-wide association studies of cancer, 197 distinct cancer-associated SNPs have been identified in 124 known genes and 31 non-genic (gene desert) loci. Above all, the most notable finding has been a non-genic locus on chromosome 8q24. This locus has been found to harbor SNPs that impact the risk of several cancer sites—bladder, breast, colorectal, prostate, and chronic lymphocytic leukemia—suggesting that chromosome 8q24 is acutely involved in the carcinogenic process. With no known genes at the 8q24 locus, mechanistic studies have focused on the proximal proto-oncogene, *MYC*, in which transcriptional enhancing interactions have been observed with the 8q24 SNP (rs6983267).[18-19]

In addition to illuminating novel genomic regions, genome-wide association studies of cancer have confirmed the involvement of key etiologic pathways. For example, genome-wide association studies of lung cancer[20-21] have identified a region of chromosome 15q25.1 as a susceptibility locus, harboring genes that encode for subunits of the nicotinic acetylcholine receptors, which also influence nicotine dependence.[22] It remains to be determined whether this association at 15q25.1 is attributable to primarily lung cancer or rather smoking behavior, the strongest risk factor for lung cancer. To disentangle this complex gene-smoking interaction, additional large studies with rigorous epidemiologic design and details are essential.

Because the statistical power of the genome-wide association approach is greatest for detecting common genetic variants, the majority of the newly identified risk variants are common and most impart moderate effects on risk of disease to those who carry them. At present, these risk variants capture only a small portion of the heritability of disease and are not useful for risk prediction. However, as additional genome-wide association study loci are found with new technologies and larger studies, it is expected that risk prediction will improve and have important implications in targeting those at greatest risk of disease.

Future Perspective

With the identification of over a hundred genetic variants associated with a multitude of cancers, the era of genome-wide association studies promises to greatly enhance our understanding of the genetic causes of human cancers. Genome-wide association studies will continue to become more comprehensive as genotyping platforms are being developed that will capture greater numbers of genetic variants and provide more complete coverage of the human genome. The 1,000 Genomes project (http://www.1000genomes.org) is currently in progress of sequencing the entire genome of at least one thousand subjects from a number of different ethnic

groups, providing an in depth resource of less common variants for association testing. Soon, genome-wide genotyping platforms will be replaced by genome-wide sequencing technology, making it possible to examine the entire human genome as part of any study of human disease.

Authors' Affiliation:
- Department of Neurology, Ernest Gallo Clinic and Research Center at the University of California, San Francisco, Emeryville, CA 94608. (E.J.)
- Epidemiology Program, Cancer Research Center of Hawai'i, University of Hawai'i, Honolulu, HI 96813. (I.C.)

References

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science.* 1996;273(5281):1516-7.
2. Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet.* 2004;74(4):610-22.
3. Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de la Cruz M, et al. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc Natl Acad Sci USA.* 2003;100(10):5896-901.
4. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. *Science.* 2005;307(5712):1072-9.
5. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet.* 2006;15 Spec No 1:R57-66.
6. Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, et al. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry.* 2009;166(5):540-56.
7. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant... or not? H*um Mol Genet.* 2002;11(20):2417-23.
8. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet.* 2005;6(2):109-18.
9. Jorgenson E, Witte JS. Coverage and power in genomewide association studies. *Am J Hum Genet.* 2006;78(5):884-8.
10. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics.* 2002;18(1):19-24.
11. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nat Rev Genet.* 2006;7(11):885-91.
12. Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, et al. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet.* 2004;13(6):577-88.
13. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science.* 2001;293(5529):489-93.
14. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet.* 1999;65(1):220-8.
15. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol.* 2001;60(3):155-66.
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904-9.
17. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci USA.* 2006;103(38):14068-73.
18. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet.* 2009;41(8):882-4. PMCID: 2763485.
19. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet.* 2009;41(8):885-90.
20. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008;40(5):616-22. PMCID: 2713680.
21. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature.* 2008;452(7187):633-7.
22. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature.* 2008;452(7187):638-42.

Laulima: cooperation, partnership

UPCOMING CME EVENTS

Interested in having your upcoming CME Conference listed? Please contact Nathalie George at (808) 536-7702 x103 for information.

Date	Specialty	Sponsor	Location	Meeting Topic	Contact
November 2010					
11/1-11/5	AN	California Society of Anesthesiologists	Mauna Lani Resort & Spa, Kailua-Kona, Hawai'i	2010 CSA Fall Hawaiian Seminar	Web: www.csahq.org
11/20	Multi	Hepatitis Support Network of Hawai'i and Hawai'i Consortium for Continuing Medical Education	Queen's Conference Center	Viral Hepatitis in Hawai'i - 2010	Tel: (808) 538-2881 Web: www.hepatitis.IDLinks/symposium2010
January 2011					
1/24-1/28	AN	California Society of Anesthesiologists	Mauna Lani Resort & Spa, Kailua-Kona, Hawai'i	2011 CSA Winter Hawaiian Seminar	Web: www.csahq.org
February 2011					
2/13-2/18	R	University of California San Francisco School of Medicine	Fairmont Orchid, Kohala Coast, Hawai'i	Neuro and Musculoskeletal Imaging	Web: www.cme.ucsf.edu/cme
2/16-2/20	EM	University of California San Francisco School of Medicine	Marriott Ihilani Resort & Spa, O'ahu	High Risk Hawai'i 2011	Web: www.retinameeting.com
2/19-2/20	OTO	University of California San Francisco School of Medicine	Moana Surfrider Hotel, Waikiki, O'ahu	American College of Surgeons Thyroid and Parathyroid Ultrasound Skills-Oriented Course	Web: www.osnhawaiianeye.com
2/19-2/22	OTO	University of California San Francisco School of Medicine	Moana Surfrider Hotel, Waikiki, O'ahu	Pacific Rim Otolaryngology Head and Neck Surgery Update	Web: www.csahq.org
2/20-2/25	IM	University of California San Francisco School of Medicine	Fairmont kea Lani, Maui	Infectious Diseases in Clinical Practice: Update on Inpatient and Outpatient Infectious Diseases	Web: www.csahq.org
March 2011					
3/13/-3/18	Multi	Mayo Clinic	Mauna Lani Bay Hotel, Kohala Coast, Hawai'i	14th Mayo Clinic Endocrine Course	Web: www.mayo.edu/cme
3/20-3/23	GS	University of California San Francisco School of Medicine	Wailea Beach Marriott, Maui	Postgraduate Course in General Surgery	Web: www.cme.ucsf.edu/cme
April 2011					
4/3-4/8	IM	University of California San Francisco School of Medicine	Wailea Beach Marriott, Maui	Primary Care Medicine: Update 2011	Web: www.cme.ucsf.edu/cme
May 2011					
5/14-5/19	P	American Psychiatric Association	Hawai'i Convention Center, Honolulu	164th Annual Meeting	Tel: (703) 907-7300 Web: www.psych.org
October 2011					
10/24-10/28	AN	California Society of Anesthesiologists	Grand Hyatt, Poipu Beach, Kaua'i	2011 CSA Fall Hawaiian Seminar	Web: www.csahq.org
January 2012					
1/23-1/27	AN	California Society of Anesthesiologists	Hyatt Regency Maui, Ka'anapali Beach, Maui	2012 CSA Winter Hawaiian Seminar	Web: www.csahq.org
February 2012					
2/13-2/18	IM	University of California San Francisco School of Medicine	Grand Hyatt Kaua'i	Infectious Diseases in Clinical Practice: Update on Inpatient and Outpatient Infectious Diseases	Web: www.cme.ucsf.edu/cme
April 2012					
4/2-4/7	IM	University of California San Francisco School of Medicine	Wailea Beach Marriott, Maui	Primary Care Medicine: Update 2012	Web: www.cme.ucsf.edu/cme