Genome-Wide Association Studies of Cancer: Principles and Potential Utility

Review Article | June 15, 2010 | Cancer and Genetics [2], Oncology Journal [3]

By Zsofia K. Stadler, MD [4], David J. Gallagher, MD [5], Peter Thom, MS [6], and Kenneth Offit, MD, MPH [7]

Genome-wide association studies (GWAS) have emerged as a new approach for investigating the genetic basis of complex diseases. In oncology, genome-wide studies of nearly all common malignancies have been performed and more than 100 genetic variants associated with increased risks have been identified. GWAS approaches are powerful research tools that are revealing novel pathways important in carcinogenesis and promise to further enhance our understanding of the basis of inherited cancer susceptibility. However, “personal genomic tests” based on cancer GWAS results that are currently being offered by for-profit commercial companies for cancer risk prediction have unproven clinical utility and may risk false conveyance of reassurance or alarm.

Until relatively recently, the field of medical genetics has focused on the identification and treatment of rare, single-gene disorders that are usually associated with a high-risk of a particular disease or trait. Since Watson and Crick’s groundbreaking explanation of DNA structure in 1953, our understanding of patterns of inheritance has progressed considerably. We now recognize that the genetic architecture of complex diseases such as cancer is probably better characterized by polygenic and multifactorial inheritance, wherein heritability is determined by the joint action of multiple genes and their interaction with environmental factors.

Technological advances have facilitated detailed interrogation of the human genome and moved investigations of genetic causation from linkage studies in high-risk cancer families to genome-wide association studies (GWAS). Through a hypothesis-neutral genome-based approach, GWAS compare common DNA variations in a large set of unrelated cases and controls to identify genetic variants associated with disease risk. Common genetic variations that contribute to susceptibility to more than 40 different diseases, including heart disease, diabetes, asthma, psychiatric disorders and inflammatory bowel disease, have been identified.[1]

GWAS have also been performed for most common malignancies and more than 100 genomic variants associated with risks of various cancers have been described. This unprecedented rapid amassing of new genetic risk variants associated with cancer risk has generated hope that these germ-line markers may also prove useful for cancer prevention, improve our understanding of cancer pathogenesis, and possibly direct treatment of patients. However, considerable scientific barriers must be overcome before these genomic data can meaningfully contribute to patient care. In this review we detail the role of genomic variation in cancer susceptibility and discuss the potential clinical implications of recent discoveries. We review warnings by professional societies and government advisory bodies regarding the use of genomic information to guide clinical decisions, with particular emphasis on the potential risks posed by for-profit, direct-to-consumer marketing of genomic risk panels.

The Genetics of Cancer Predisposition

Genetic linkage studies performed in the 1980s and 1990s identified a number of highly penetrant syndromes responsible for cancer predisposition within certain families. Molecular insight into breast-ovarian cancer (BRCA1/2), Lynch (mismatch repair genes), Li Fraumeni (p53) and Cowden (PTEN) syndromes, among others, has had a powerful impact on current clinical management of families affected by these syndromes. These syndromes, however, account for only a fraction of the familial risk of cancer, leaving much of the heritability of cancer unexplained. Epidemiologic evidence provides strong support for a hereditary component to cancer. Twin studies demonstrating concordance for most cancers among monozygotic vs dizygotic twins and siblings,[2] and population studies of familial clustering,[3] suggest that genetic factors predispose to even environmentally induced diseases such as lung cancer. Using a candidate gene approach, pathways thought to be important in carcinogenesis have been investigated. In breast cancer, for example, a
small number of candidate genes involved in the response to DNA damage (ATM,[4-6] CHEK2,[7-9] BRIP1,[10] and PALB2[11]) have been associated with a modest increase in breast cancer risk (approximately two-fold). It is estimated that these genes, in combination with the high penetrance but rare cancer-predisposition–syndrome genes such as BRCA, account for only about one-quarter to one-third of familial risk, with the remainder remaining unexplained.[12] One explanation for this seemingly “missing” fraction of heritable disease is the common variant common disease (CVCD) hypothesis,[13] which assumes that many different common genetic variants, each with a small effect size, collectively cause disease.

Recent developments, including the completion of the Human Genome Project, the HapMap Project, and the emergence of high-throughput genotyping, have allowed further investigation of this hypothesis, through the GWAS approach.

**Experimental Design of Genome-Wide Association Studies**

Genomic variation in GWAS reported to date is largely represented by single base pair changes known as single nucleotide polymorphisms (SNPs). Although structural genomic variations such as deletions, duplications, inversions, and copy-number variations also exist and can be measured, assessment of these more complex genomic changes are just beginning to be incorporated into large genome-wide studies. GWAS rely on the phenomenon of linkage disequilibrium,[14] wherein SNPs are not inherited individually but instead are in linkage disequilibrium blocks, with many nearby SNPs being highly correlated. This enables the selection of one SNP (the tag SNP) that represents up to 50,000 surrounding base pairs. Through the HapMap project[15] we have learned that genotyping sets of 500,000 to 1,000,000 tag SNPs can cover approximately 80% of all common SNPs in the genome.[16]

![Design of a multistep cancer GWAS](image)

These striking technological advances have facilitated GWAS studies using multistep designs (Figure 1). As large GWAS are often cost-prohibitive, a tiered, multistage approach is commonly used. In a simple two-stage design, all tag SNPs are tested during the first stage on a small subset of cases and controls, usually between 25% to 50% of total participants. During the second stage, significant SNPs, comprising approximately 10% of the SNPs tested, are genotyped on all remaining samples. The same sample of SNPs is then tested on the initial subset, allowing for a combined analysis.[17] For study designs with three or more stages, the significant SNPs are included for replication testing in different case-control sample sets.

A typical GWAS, containing 1,000 to 2,000 cases with often a larger number of controls, is powered to detect differences in common SNPs with an effect size of 1.5–2.0-fold. Importantly, some smaller GWAS reporting effect sizes of 1.1–1.3 must be interpreted with caution, as these studies may have been underpowered to detect such small differences. In fact, the detection of modest genetic effects with odds ratios of 1.3 or less and minor allele frequencies under 10% may require more than 10,000 cases and 10,000 controls for adequate statistical power.[18] Most recent significant findings have been for alleles with relative risks of 1.1 to 1.3, and their detection has required large, global, collaborative group efforts.

GWAS approaches require careful selection and accurate characterization of case and control participants. In breast cancers, for example, gene-expression profiling has defined different subtypes of breast tumors (e.g., luminal, basal subtype) with variations in response to treatment and clinical outcomes. As we learn more about the heterogeneity of malignancy it may be necessary to perform individual GWAS for different subtypes of disease, to accurately detect underlying genetic heterogeneity. As an example, the 10q26 locus that maps to FGFR2 and has been implicated in a number of breast cancer GWAS, and is most strongly associated with the diagnosis of estrogen receptor (ER)-positive disease.[19]
GWAS for women with triple-negative breast cancer are ongoing and may demonstrate different genetic associations. A well-matched and carefully characterized control cohort is another key component of GWAS design that is a challenging but essential prerequisite for a precise study. For example, cases and controls must be well-matched to avoid population stratification, which can bias results due to differences in race or ethnicity of those with and without the disease of interest.[20-23]

**Genome-Wide Association Study Interpretation**

Analysis of GWAS data produces an odds ratio and $P$ value for each SNP, but great care is required to avoid false-positive or false-negative reporting. A quality control ‘data cleaning’ is first performed to detect problems such as poor genotyping results or unexpected relatedness among participants. Allele frequencies among controls, for example, are compared with an accepted standard (Hardy-Weinberg equilibrium), as deviations from this equilibrium often indicate a methodologic flaw in the study. Population heterogeneity can introduce a bias in studies of participants with different ancestries, although a post-hoc principal components analysis can be performed to account for this.[24] Further stringent statistical analysis must be applied to account for the multiple testing performed in GWAS and is usually achieved using a Bonferroni correction, in which the threshold $P$ value (usually $5 \times 10^{-2}$) is divided by the number of tests performed (~500,000 depending on array used) resulting in a $P$ value in the region of $< 1 \times 10^{-7}$.

Only associations between a SNP and an outcome of interest that fall below this threshold should be considered statistically significant, but despite this rigid approach false-positive results frequently occur, as evidenced by failure to replicate findings in subsequent studies using the same SNPs and similar populations. Despite optimal design and careful analysis, interpretation of GWAS results has been complicated by a publication bias whereby positive associations are preferably published.

**Lessons From Genome-Wide Association Studies in Cancer**

---

**TABLE 1**

Glossary of Terms

---

**TABLE 2**
Genome-Wide Association Studies of Cancer: Principles and Potential Utility

Published on Cancer Network (http://www.cancernetwork.com)

Genome-Wide Association Studies in Solid Malignancies

TABLE 3

Genome-Wide Association Studies in Hematologic Malignancies

Genomic variation has not yet contributed positively to patient management in cancer care; nonetheless, the 50-plus GWAS that have been performed for more than 15 malignancies have generated many interesting associations, stimulating widespread interest in the clinical utility of this tool. Breast cancer (nine GWAS),[25-33] prostate cancer (12 GWAS),[34-45] colorectal cancer (seven GWAS),[46-52] and lung cancer (seven GWAS),[53-57] have been extensively investigated, and similar studies have been performed in pancreatic,[58,59] gastric,[60] esophageal,[61] bladder,[62,63] testicular,[64,65] ovarian,[66] skin,[67-71] thyroid,[72] neuroblastoma,[73,74] brain,[75,76] and hematological cancers,[77-81](Tables 2 and 3). A detailed discussion of each study is beyond the scope of this manuscript, but some notable lessons have been learned from individual findings.

The 8q24 locus, first reported in Icelandic patients, is of particular interest.[42] It is located in a region containing no known genes (referred to as a ‘gene desert’) but has been associated with prostate, bladder, breast, and colorectal cancer, and the association has been successfully replicated in independent studies. It is possible that genomic variation at this locus marks a carcinogenic pathway common to many cancers, and recent evidence suggests a role in regulation of the myc oncogene.[82-84] A second locus at 5p15.33 has also been associated with multiple cancers. A multicancer study of > 30,000 cases and > 45,000 controls identified an association between the 5p15.33 locus and risk of lung, bladder, prostate, cervical, and basal cell cancer, with a trend towards a protective effect for cutaneous melanoma.[85]

Epidemiological studies have identified a higher incidence of aggressive prostate cancer in African-American men, and there has been much debate as to whether this represents a socioeconomic difference in how African-American men are screened and treated for prostate cancer or a distinct natural history. Recent GWAS data suggest that seven independent risk variants identified in the 8q24 region are more commonly associated with prostate cancer in African-American men, possibly suggesting a genetic link for this epidemiological observation.[86] The first GWAS in pancreatic cancer provided another example of genomic data supporting prior epidemiological observations.[87] Pancreatic cancer risk had long been associated with ABO blood type, and a SNP located in the first intron of the ABO blood group gene on chromosome 9q24 significantly increased the risk of pancreatic cancer by 1.2-fold.[59]

Cancer GWAS have also identified associations in genes encoding proteins known to be important in organogenesis and organ function. Prostate cancer GWAS have identified a number of significant associations between prostate cancer diagnosis and SNPs in genes that encode key products of the prostate; serum levels of micro-seminoprotein and a number of kallikreins have been identified as potential prostate cancer biomarkers, and GWAS have now offered associated genomic markers for further study.[34,35] In testicular germ cell tumors, in which genetic susceptibility is supported by an 8- to 10-fold and a 4- to 6-fold increased risk of disease in brothers and fathers of patients, respectively, the two GWAS both identified risk variants at 12p22,[64,65] a region that contains KITLG, which is necessary for germ cell development.[88] The risk variant at 12p22 has one of the highest associated risks of any GWAS, conferring more than a 2.5-fold increase in cancer risk. The only thyroid cancer GWAS performed to date identified two significant predisposing SNPs, one at 9q22.33 near FOXE1, which has been implicated in thyroid organogenesis, and the other at 14q13.3 near NKX2-1, a gene important in thyroid gland differentiation.[72]

GWAS in hematological malignancies have identified a number of predisposing SNPs for chronic lymphocytic leukemia, non-Hodgkin lymphoma, and childhood leukemia[77-80]; however, in a study of myeloproliferative neoplasms (MPN) we noted an interesting finding. A somatic point mutation in JAK2 is commonly found in MPN, and the GWAS identified a germline SNP in this same gene that resulted in a three-fold-increased risk of developing MPN. This finding linked this germline and somatic variation, a finding similar to the I1307K APC variant and colorectal cancer.[81] Melanoma GWAS have provided further results of biological interest. CDKN2A is a well-recognized high-penetrance melanoma susceptibility gene, and a recent GWAS associated an adjacent locus with increased risk of melanoma, confirming this as an important region in melanoma
This study reported a risk allele at 16q24, a region involving cell cycle regulation, and this SNP was previously reported in a GWAS of hair color and skin pigmentation. Significant associations with variants affecting hair, eye, and skin color were identified in another melanoma GWAS, confirming the ability of this approach to reflect underlying biology.

**Translating Genome-Wide Association Studies Into Clinical Practice**

These examples and others in nonmalignant diseases provide evidence that GWAS can identify genomic variations associated with disease risk and known biology. Nonetheless, a systematic review of 260 meta-analyses of 160 SNP associations reported that there is insufficient scientific evidence to conclude that genomic profiles are useful in measuring genetic risk for common diseases or in developing personalized recommendations for disease prevention.

Importantly, neither the clinical validity (accuracy of a test in predicting the clinical outcome) nor the clinical utility (risks and benefits resulting from the use of a test) of this approach have been established. As an example of the challenge, consider the 1.26-fold increased risk of breast cancer associated with the risk allele in the FGFR2 gene. This represents a 26% relative increase in risk over a lifetime, comparable to that conferred by delaying age of first pregnancy to more than 35 years, but without knowing an individual's absolute risk this information is somewhat meaningless. Furthermore, it is unclear how to best combine risk alleles into a 'multi-SNP' prediction tool, and how to incorporate this approach into existing clinical models.

Even if the common variants identified lay within known genes, gene-gene interaction (referred to as epistasis) is poorly understood, and it is unknown if the relative risks associated with each SNP should be combined in an additive fashion or multiplicatively. Experimental multi-SNP models have been compared with predictive tools that use clinical variables alone. For example, in a modeling study, the addition of genotypes from seven breast cancer–associated SNPs to the National Cancer Institute’s Breast Cancer Risk Assessment Tool was of no clinical significance, with an increase in the area under the receiver operating characteristic curve from 0.607 to 0.632.

More recent studies have supported only a modest increased predictive accuracy of genomic risk panels. Susceptibility SNPs have also been proposed as potential biomarkers to predict outcome after diagnosis, as they can be identified at time of blood draw, do not change over time (ie, are constitutional), and are dichotomous (ie, do not require adjustment or normalization). Breast-ovarian cancer syndrome offers an example of how germline variation can predict clinical course, with BRCA-associated ovarian cancer demonstrating an improved survival vs sporadic disease, with the reverse relationship having been suggested for BRCA-associated prostate cancer.

Genome-based prediction models in prostate cancer, in which nomograms incorporating clinical variables have had a less than satisfactory impact on patient management, have thus far offered only modest advantages. A number of studies have failed to associate susceptibility SNPs with meaningful clinical endpoints in prostate cancer or with overall survival in breast cancer. However, carefully designed GWAS of specific clinical endpoints would be required before dismissing genomic variation as a potential prognostic biomarker.

Finally, recent reports from clinical trials investigating poly (ADP-ribose) polymerase 1 (PARP-1) inhibitors in BRCA-associated malignancy provide the first example of successful targeting of germline genetic variation. Genomic variation, either rare variants associated with specific disease or loci associated with different malignancies, such as 8q24, has the capacity to identify novel, potentially targetable pathways important in carcinogenesis.

The appeal of basing clinical decisions on the results of a blood or even saliva test is obvious. The current complex issues in the management of prostate cancer offer an example of how germline variation may one day contribute meaningfully to patient care. Despite recent results from two large prostate cancer screening studies, it is unclear if screening unselected patients is of benefit, and it is difficult to identify populations at highest risk. In the absence of markers of increased risk, it is also not clear whether or not 5α-reductase inhibitors should be used for routine disease prevention.

After diagnosis of prostate cancer, many men may not require treatment, but it is unclear how to best identify this cohort. For men with advanced, hormone-refractory disease, treatment options are limited and novel therapeutic approaches are urgently needed. It is feasible that germline variation could be used to identify populations appropriate for screening, prevention, and certain treatments. Such genetic variants may also lead to discovery of pathways important in prostate cancer pathogenesis, and the results from the prostate cancer GWAS completed to date begin to set the
The Current Promise and Peril of Clinical Translation

The potential of genomic approaches to have a significant impact on our understanding of disease biology is unquestioned, but the clinical relevance of genomic variation for most common disorders remains to be established. Recognizing the scientific merits as well as the limitations of GWAS, the American Society of Clinical Oncology has recently updated its policy statement on genetic (and genomic) testing for cancer susceptibility. Among other recommendations, the panel stresses that genetic tests with uncertain clinical utility, such as genomic risk assessments, should be administered in the context of clinical trials and should be conducted in the setting of pre- and post-test counseling by experienced healthcare professionals. The panel, as well as a recent report from the Secretary’s Advisory Committee on Genetics, Health, and Society supports further research to demonstrate the clinical validity and reproducibility of genomic profiles, the establishment of a national registry for genomic tests for disease risk, and the incorporation of genomic-risk variants into prospective clinical trials. Despite a need for caution, the commercialization of GWAS results has occurred in a largely unrestricted manner. Governance of this technology lies beyond the traditional scope of the US Food and Drug Administration, and as a result for-profit companies have been able to market SNP panels directly to the public. For conditions such as hair color or even ancestry determination there is possibly little harm in this approach. However, the reporting of clinically meaningless disease risk figures is worrisome. Increased risk scores with unproven clinical relevance may result in anxiety, and reduced risk scores based on preliminary data may result in incorrect reassurance regarding individual cancer risk. Individuals may consequently ignore medical advice, such as age-appropriate screening recommendations, because of a perceived reduction in risk. The challenge of translating technological advances into patient care is not new to the field of medicine, and a well-established patient-centered model exists to optimally educate, empower, and put into effect preventive strategies. Genetic counseling is an important component of the clinical genetics model that should occur prior to genetic testing. It details the special social, economic, and psychological ramifications of genetic and genomic information, covering the implications of testing for individuals and for their relatives. This entire process may be omitted when genetic or genomic tests are marketed directly to consumers. Post-hoc counseling for genomic tests, however, is a different matter entirely; because the results are generally not interpretable from a clinical perspective, it is often impossible to alleviate anxiety or give advice on best management.

Conclusion

The field of preventive oncology, incorporating cancer genetics, has progressed considerably in the past decade. Management paradigms have been established for distinct syndromes that account for a significant subset of cancer predisposition. This progress has had meaningful impact on the natural history of disease affecting families worldwide. Recent technological advances that vastly improve the capacity to study genomic variation across populations have identified many associations between SNPs and a variety of malignancies. Analysis of genetic variants reviewed here, ongoing studies of structural variation, and future analysis of individual sequence differences have enormous potential to yield breakthroughs in our understanding of the genetic mechanisms of cancer susceptibility. In addition to SNPs and structural variants, microRNAs, long-range promoters, and epigenetic factors that regulate gene expression will also interact to play a part in genetic susceptibility to disease. Pending a better understanding of the complex interactions of these factors, as well as longitudinal studies of human populations that have been genetically characterized, the routine use of research-derived genomic risk panels in a clinical setting is premature. The initial harvest of information resulting from the human genome project has been substantial. However, a considerable translational effort remains before the findings of this first generation of genomic association studies can be incorporated into preventive medicine and oncologic practice.

Acknowledgments: Z.S. is supported by the Robert and Kate Niehaus Clinical Cancer Genetics Initiative at Memorial Sloan-Kettering Cancer Center. K.O. was senior author of the American Society of Clinical Oncology (ASCO) policy statement update: Genetic and genomic testing for cancer susceptibility [101]; the views expressed in this article are not those of ASCO.

Financial Disclosure: The authors have no significant financial interest or other relationship with...
Genome-Wide Association Studies of Cancer: Principles and Potential Utility
Published on Cancer Network (http://www.cancernetwork.com)

References:


93. Gail MH: Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk


Source URL:
http://www.cancernetwork.com/cancer-and-genetics/genome-wide-association-studies-cancer-principles-and-potential-utility

Links:
[1] http://www.cancernetwork.com/review-article
[3] http://www.cancernetwork.com/oncology-journal
[4] http://www.cancernetwork.com/authors/zsofia-k-stadler-md
[5] http://www.cancernetwork.com/authors/david-j-gallagher-md
[6] http://www.cancernetwork.com/authors/peter-thom-ms
[7] http://www.cancernetwork.com/authors/kenneth-offit-md-mph