Studying Genetic Variations in Cancer Prognosis (and Risk): A Primer for Clinicians

SEVTAP SAVAS, GEOFFREY LIU

Division of Applied Molecular Oncology, Department of Medical Biophysics, Ontario Cancer Institute, Toronto, Canada; Division of Medical Oncology and Hematology, Department of Medicine, Princess Margaret Hospital, Toronto, Canada

Key Words: Cancer predisposition • Treatment response and prognosis • Low-penetrance alleles • SNPs • Genome-wide association studies • Candidate gene approach

Disclosures

Sevtap Savas: None; Geoffrey Liu: None.

The content of this article has been reviewed by independent peer reviewers to ensure that it is balanced, objective, and free from commercial bias.

LEARNING OBJECTIVES

1. Evaluate SNPs as genetic markers and functional genetic variations.
2. Select candidate genes for cancer research based on knowledge of their biological function.
3. Assess candidate gene and genome-wide association studies for their potential to improve translational research.

ABSTRACT

Rare, high-penetrance genetic variations account for a small portion of genetic cancer syndromes. In contrast, most cancers develop from a combination of minor genetic influences and environmental factors. There are numerous publications on cancer susceptibility. In contrast, genetic studies in treatment response and outcome analyses are a rapidly emerging field. Approaches used in disease susceptibility can be adapted for genetic outcome studies. In this review, we summarize the current knowledge on how candidate genes and genetic variations are selected to evaluate gene–outcome, gene–prognosis, and gene–treatment response relationships as applicable to the practicing oncologist. The Oncologist 2009;14:657–666

INTRODUCTION

Familial or inherited high-penetrance genes account for only a small portion of cancer cases. In the remaining sporadic cases, genetic variations in the form of low-to-mod-erate penetrance alleles may predispose individuals to cancer in combination with environmental factors [1, 2]. Individual low-penetrance risk alleles are insufficient to cause cancer, but may influence cancer risk.

Correspondence: Sevtap Savas, Ph.D., Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, 300 Prince Philip Drive, Room H4333, St. John’s, Newfoundland, A1B 3V6 Canada. Telephone: 709-777-2233; Fax: 709-777-7497; e-mail: savas@mun.ca Received March 9, 2009; accepted for publication May 16, 2009; first published online in The Oncologist Express on July 6, 2009. ©AlphaMed Press 1083-7159/2009/$30.00/0 doi: 10.1634/theoncologist.2009-0042

The Oncologist 2009;14:657–666 www.TheOncologist.com
Genetic variation in the form of low-penetrance alleles can also explain variable treatment response and prognosis among patients. Serious side effects and differential efficacy in chemotherapy and radiotherapy are frequent observations in the clinical setting and could be determined by inherited genetic factors [3]. Genetic risk factors (affecting carcinogenesis and therefore the biology of the disease) and treatment response factors may thus have direct effects on overall prognosis, survival, and outcome [4, 5]. There is a relatively small but growing number of genetic studies of cancer treatment response and outcome, and many more on cancer susceptibility. Classically, molecular and genetic epidemiologists have studied genetic variations in cancer risk. For cancer outcome studies, a clinician’s role is essential. Here, we summarize the genetic approaches applied to disease susceptibility studies that can be adapted for outcome studies. The goal is to aid clinicians with methodological concepts in understanding how studies related to the genetic basis of interpatient variability in treatment response and prognosis are developed.

**GENETIC VARIATION: USING SINGLE NUCLEOTIDE POLYMORPHISMS AS A PROTOTYPE POLYMORPHISM**

The human genome contains a massive amount of genetic variation [6, 7], such as the insertion/deletion of one or more nucleotides (indels), the copy number variations (CNVs) that can involve DNA sequences of a few kilobases up to millions of bases, and single nucleotide polymorphisms (SNPs), which are the substitution of a single nucleotide along the DNA (Fig. 1). This review primarily focuses on SNPs to illustrate concepts, with the understanding that most concepts can be applied to most genetic variations.

SNPs are the most common form of genetic variation. There are ≈10 million SNPs estimated to be in the human genome [1, 8]. There is approximately one SNP in every 300 bp, but the density of the SNPs changes significantly over SNP-dense and SNP-desert regions in the genome. Although the majority of SNPs are common to at least three historic human populations (white, African, and Asian), there are also SNPs that are specific to different ethnicities [8]. These differences among human populations have important implications in biomedical science. For example, ethnic genetic differences should always be considered in disease risk, treatment response, and outcome studies.

SNPs as genetic markers are used in linkage and association studies, in which a specific trait can be linked or associated with a specific genomic region. These SNPs do not necessarily contribute to the phenotype (i.e., clinical disease entity) directly, but, rather, they may be in linkage disequilibrium (LD)
TagSNPs as Genetic Markers

The international human haplotype mapping consortium (HapMap) has played a significant role in human genetic variation studies by genotyping >4 million SNPs in 269 individuals from three major human populations (African, Asian, and white) [9, 10]. One of the major highlights of this project was the identification of a draft LD structure of the human genome. According to this finding, the human genome consists of block-like LD regions inherited as a unit and characterized by limited DNA recombination and low haplotype diversity. Use of LD information considerably benefits current genetic studies, because it led to the idea of using tagSNPs, which are SNPs that are in LD with other SNPs in the same LD block. For example, a genetic variation may have a tendency to be coinherited with its neighboring genetic variations. In this way, only one neighbor SNP needs to be studied to understand the entire neighborhood. This chosen neighbor is termed the tagSNP (because it tags the neighborhood). Therefore, by analyzing tagSNPs, other SNPs in close vicinity need not be included in the analysis, which significantly reduces the number of SNPs to be analyzed (Fig. 1). Thus, if one SNP is the one producing a functional gene effect, then studying one of the neighboring SNPs will capture this effect.

SNPs as Functional Genetic Variations

Functional SNPs represent a subset of genetic variations that have a direct effect on gene expression and protein function. They are usually located in and around known genes. Because specific gene regions have specific biological functions, depending on the location of the SNPs along the genes one can predict the potential biological role of a SNP (Fig. 2, Table 1).

Functional SNPs may explain a significant portion of complex disease predisposition, treatment response, and prognosis by acting as low-penetrance alleles. Therefore, identification of the functional status of SNPs is important in basic science as well as in clinical studies and is achieved by either experimental studies or in silico analyses. Although experimental results are available for many proteins and SNPs [11], and more and more high-throughput technologies are becoming available to test specific functional features of SNPs [12], current technology is not advanced enough to systematically and functionally characterize many interesting SNPs. Therefore, computational in silico (i.e., in computer simulation) methods have been widely used to analyze, select, and prioritize SNPs (the main concepts are summarized in Figure 3). One caveat of the in silico approach is that most in silico tools are characterized by certain false-positive and false-negative rates. Therefore, experimental assessment is essential to validate the putative biological features of specific genetic variations.

Table 1. Potential biological effects of different types of SNP

<table>
<thead>
<tr>
<th>SNP type</th>
<th>Genic location</th>
<th>Potential biological effect</th>
<th>Biology behind SNP effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>rSNPs</td>
<td>Promoter regions</td>
<td>Altered gene expression</td>
<td>Altered regulatory motifs, such as transcription factor binding sites</td>
</tr>
<tr>
<td>5′-UTR SNPs</td>
<td>5′-UTRs (noncoding exons)</td>
<td>Altered gene expression, altered protein translation</td>
<td>Altered regulatory motifs, such as cap structure</td>
</tr>
<tr>
<td>3′-UTR SNPs</td>
<td>3′-UTRs (noncoding exons)</td>
<td>Altered gene expression, altered mRNA stability</td>
<td>Altered regulatory motifs, such as microRNA binding sites</td>
</tr>
<tr>
<td>nsSNPs</td>
<td>Coding exons</td>
<td>Altered protein function and structure</td>
<td>Altered protein amino acid sequence</td>
</tr>
<tr>
<td>sSNPs</td>
<td>Coding exons</td>
<td>Altered mRNA stability, altered protein function</td>
<td>Altered mRNA sequence</td>
</tr>
<tr>
<td>X-SNPs</td>
<td>Coding exons</td>
<td>Altered mRNA stability and protein function</td>
<td>Altered coding region sequence, altered protein amino acid sequences</td>
</tr>
<tr>
<td>Splice-site SNPs</td>
<td>Splice acceptor (ag dinucleotide in intron immediately before exon sequences) and splice donor site (gt dinucleotide in intron immediately after exon sequences)</td>
<td>Altered mRNA splicing, mRNA stability and protein function</td>
<td>Altered consensus splice site sequences</td>
</tr>
</tbody>
</table>

For simplicity, only the most common eukaryotic splice acceptor and donor sites are shown. Abbreviations: nsSNP, nonsynonymous SNP; rSNP, regulatory SNP; SNP, single nucleotide polymorphism; sSNP, synonymous SNP; UTR, untranslated region; X-SNP, termination codon introducing SNP.
Nonsynonymous SNPs as Functional Genetic Variations

Nonsynonymous SNPs (nsSNPs) are located within coding exons and lead to amino acid substitutions (Fig. 2, Table 1), and therefore are likely to change protein function. Perhaps the best evaluation of the functional consequences of an nsSNP can be done using the three-dimensional (3D) structure of the protein. Although there are several programs and tools available to analyze 3D structures of proteins, they are limited in their application because of the unavailability of the full 3D structure for most proteins. Therefore, other indirect approaches have been developed, such as evolutionary conservation analysis, to predict the biological importance of nsSNPs. Evolutionary conservation analysis

Figure 2. Hypothetical examples of biological consequences of different types of single nucleotide polymorphisms (SNPs). (A): Nonsynonymous SNPs (nsSNPs) can affect protein structure and function. (B): Synonymous SNPs (sSNPs) can affect mRNA stability. (C): Splice-site SNPs can affect mRNA splicing. (D): Regulatory SNPs (rSNPs) can affect gene expression. (E): Termination codon introducing SNPs (X-SNPs) can affect mRNA stability and protein function.
is based on the notion that if an amino acid is conserved among the members of the protein family, then it is important for the structure and function of the protein. Several computational tools have been developed to analyze evolutionary conservation status. Among these, SIFT [13, 14] and PolyPhen [15] are the most widely used [16–20].

**Synonymous SNPs as Functional Genetic Variations**

Synonymous SNPs (sSNPs) are located in exons yet do not substitute encoded amino acids, and thus are typically considered functionally benign (Fig. 2, Table 1). However, there are rare examples of such SNPs with functional consequences. For example, in the multidrug resistance 1 gene (MDR1), an sSNP characterized by a C to T substitution at mRNA position 3435 (3435C>T) causes a change in the mRNA secondary structure, leading to decreased mRNA stability [21]. The same sSNP was later shown to affect the substrate affinity of the MDR1 protein, probably through altering the timing of cotranslational protein folding [22]. Currently, only experimental approaches are available to detect this kind of functional genetic variation.

**Splicing Modifying SNPs as Functional Genetic Variations**

Substitutions in the consensus 5’ and 3’ splice sites at exon–intron junctions (Fig. 2, Table 1) can affect the splicing process. For example, a splice-site polymorphism in the LILRA2 gene generates a splice isoform that is associated with systemic lupus erythematosus and microscopic polyangiitis [23]. Similarly, an sSNP in the MCAD gene was shown to inactivate an exonic splicing enhancer (ESE) element that is responsible for recognition of exons during splicing, causing loss of the functional protein product [24]. SNPs located in ESEs have been previously hypothesized as candidate functional genetic variations [25]. Fortunately, it is possible to predict those SNPs that alter ESEs using the data from a previously published study [26].

**Regulatory SNPs as Functional Genetic Variations**

Interindividual variability in gene expression is attributable to both environmental and genetic factors [27, 28]. On the genetic side are the SNPs that can affect promoter sequences, regulatory SNPs (rSNPs) (Fig. 2, Table 1), which contain important regulatory motifs such as transcriptional start sites and transcriptional factor binding sites. Such genetic variations result in altered gene expression patterns [29–31], and their association with variable drug response has been reported [32], indicating the relevance of rSNPs in cancer research.

**Untranslated Region SNPs as Functional Genetic Variations**

Post-transcriptional regulation of gene expression involves regulatory regions located in untranslated regions (UTRs) in the 5’ and 3’ ends of the transcripts (Table 1), and its abnormalities are frequently observed in cancer [33]. 5’-UTRs contain regulatory regions, such as the cap-structure and internal ribosomal entry sites, that are involved in the initiation of translation. Therefore, any perturbation in these regions by SNPs can affect the translation process.

---

**Figure 3.** Concepts for selecting candidate single nucleotide polymorphisms (SNPs). (A): Sequence homology. (B): Evolutionary conservation. (C): Known functional motifs.
addition, 5′-UTR regulatory elements have also been implicated in gene expression and associated with disease outcome [34] and susceptibility [35].

3′-UTRs of genes contain regulatory sequences such as adenine–uridine rich elements and microRNA-binding sites that affect mRNA as well as protein stability [36–38]. Therefore, polymorphisms that affect these sequences are likely to be associated with cancer risk, treatment response, or prognosis [38]. For example, a SNP at the 3′-UTR of the DHFR gene modifies DHFR transcript levels and is associated with resistance to methotrexate [39]. A previous study predicted potential microRNA-binding sites along cancer-related genes using a variety of in silico programs and identified candidate SNPs that can affect microRNA binding [40], and a study has reported two such SNPs associated with colorectal cancer risk [41].

### Table 2. Candidate gene selection approaches applied in disease susceptibility studies

<table>
<thead>
<tr>
<th>Type of candidate gene</th>
<th>Candidate gene</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positional candidates</td>
<td>Genes located in regions somatically deleted/amplified in cancer</td>
<td>Genes determined by high-resolution techniques, such as comparative genomic hybridization</td>
</tr>
<tr>
<td></td>
<td>Genes associated or linked with cancer</td>
<td>Genes determined by genetic linkage and association studies</td>
</tr>
<tr>
<td>Functional candidates</td>
<td>Genes whose biological functions are relevant to phenotype</td>
<td>Carcinogen metabolism genes, DNA repair genes, cell cycle genes, immune system genes, signal transduction genes</td>
</tr>
<tr>
<td></td>
<td>Genes with inherited mutations in cancer syndromes</td>
<td>BRCA1, BRCA2, p53, ATM</td>
</tr>
<tr>
<td></td>
<td>Proteins interacting with cancer genes</td>
<td>BRIP1 and BRDA1 proteins interact with BRCA1 protein and are associated with breast cancer susceptibility</td>
</tr>
<tr>
<td></td>
<td>Driver genes somatically mutated in cancer</td>
<td>Some protein kinases</td>
</tr>
<tr>
<td>Cancer-site specific candidate genes</td>
<td>Genes biologically related to epidemiological risk factors</td>
<td>Genes related to estrogen metabolism in breast cancer</td>
</tr>
<tr>
<td></td>
<td>Genes expressed in the affected tissue and genes upregulated and downregulated in tumors</td>
<td>Genes identified by gene expression studies, such as microarray experiments</td>
</tr>
</tbody>
</table>

### Selection of Candidate Genes in Cancer Research

Researchers have applied a variety of approaches in their studies to identify disease risk alleles, which now can be adapted for studying cancer outcome and treatment response.

#### Positional Candidate Genes

There are two main types of positional candidate genes relevant to cancer research (Table 2). The first group is located in somatically amplified or deleted genomic regions in tumors. These genes represent potential tumor suppressors (deleted) or oncogenes (amplified), and their genetic variations are worth looking at for association with cancer risk, treatment response, and prognosis [45]. The second type of positional candidate gene is located in genomic regions that are identified by linkage and association studies. Emerging genome-wide association studies (GWASs, see below) have started to identify genomic regions containing such signals, for example, in breast [46, 47], lung [48], prostate [49], and colorectal [50] cancers. Further analyses of these regions may identify genes and genetic variations causing susceptibility to these cancers [51]. Similar approaches have been adapted to evaluate cancer outcomes.

#### Functional Candidate Genes

The functional candidate gene approach is based on a priori knowledge of the biological function of genes (Table 2). For example, carcinogen metabolism genes [52] are in-
volved in preventing somatic mutations and in drug metabolism [53]. DNA repair genes have been heavily investigated in cancer genetic studies probably because of the “mutator phenotype” attributable to these genes [54] as well as their involvement in chemotherapy and radiotherapy responses [55]. Similarly, genes acting in the cell cycle, apoptosis, the immune system, and signal transduction processes also need to be considered as functional candidates because of their frequent deregulation in carcinogenesis and, in some cases, their established roles in treatment response and overall outcome in cancer [56].

In addition to these genes, previously identified high-penetrance genes such as ATM, BRCA1, BRCA2, and, p53 can prove to be good candidates because of the fact that these genes have already been shown to have important roles in carcinogenesis. Therefore, their polymorphic variants can be good candidates as low-penetrance alleles. For example, the Pro72Arg polymorphism of p53 (substitution of a proline amino acid with an arginine at amino acid position 72 on the p53 protein) is one of the most extensively studied common genetic polymorphisms [57, 58]. Additionally, genes whose protein products physically interact with the proteins already implicated in cancer form an exciting group of functional candidate genes because, in cells, proteins usually act in binary and/or multisubunit complexes mediated by protein–protein interactions, and, as anticipated, abnormalities in such protein complexes are found in a variety of diseases, including cancer [59]. Supporting this hypothesis, inactivating mutations of BRIP1 (which interacts with the BRCA1 protein) were suggested as low-penetrance breast cancer susceptibility alleles [60]. Similarly, in another BRCA1-interacting protein, BARD1, a variant was also suggested to confer a slightly greater risk for breast cancer in Nordic women [61].

Genes contributing to carcinogenesis by undergoing somatic mutations can also be evaluated for low-penetrance alleles. During carcinogenesis, both pathologic somatic mutations (driver mutations that contribute to carcinogenesis) and nonpathologic somatic mutations (passenger mutations that do not contribute to carcinogenesis) occur. Because driver mutations directly contribute to carcinogenesis, polymorphic variants of such somatically mutated genes form a group of exciting candidates [62]. Recent efforts in distinguishing driver and passenger mutations [63] have led to the identification of a variety of driver genes, such as protein kinases [64, 65]. Progress in this field will significantly improve the risk, treatment response, and prognosis studies that use the candidate gene approach.

GWASs offer unique advantages and challenges versus candidate gene approaches (Fig. 4). GWASs are characterized by genotyping of a dense set of genetic markers (usually SNPs) covering the vast majority of the genome, followed by sophisticated statistical analyses [68, 69]. There are currently a variety of high-throughput genotyping platforms for GWASs characterized by different genome coverage as well as genetic marker content and density [69].

GWASs are unbiased in nature and are therefore likely to discover previously unsuspected, novel genes. Despite this critical advantage, there are serious pitfalls related to data analysis in GWASs. The main problem is the statistical analysis and meaningful interpretation of the data. For example, in complex multifactorial diseases such as cancer, gene–gene and gene–environment interactions need to be considered in data analysis, yet powerful statistical approaches to effectively address this issue still await development. More importantly, as the number of markers analyzed increases, the probability of a false-positive association increases. This is, in part, overcome by correction for multiple testing using the Bonferroni method or the less
conservative false-discovery rate method of Benjamini and Hochberg [70, 71]. Yet further improvement in the current approaches and development of alternative approaches to overcome these limitations are urged.

Additional Challenges in Prognostic GWASs
When it comes to prognosis and outcome studies, the need for an adequate sample size to reach conclusions backed up by reasonable statistical power becomes even more important. In both risk and prognosis–outcome studies, there is a clear need to have patient cohorts as homogeneous as possible in terms of their ethnic background and disease characteristics. Multicenter collaborations and the establishment of specific cancer registries have made it possible to investigate large numbers of cancer patients in risk GWASs; however, for prognosis and outcome studies, the situation is a bit harder. In prognosis–outcome GWASs, the same patients used in risk analysis are partitioned into groups with different features based on not only their ethnic backgrounds and disease characteristics but also the treatment regimen undertaken as well as the prognosis and overall outcome achieved following this regimen. This partitioning results in a further reduction in the sample size, which brings in a serious statistical power issue to the study. Clinical trial biospecimens may be useful in this regard, but the need for multiple replication datasets may be problematic if different clinical trials are not designed to answer a specific question and involve different patient populations or treatments. Therefore, currently, prognostic studies suffer from the lack of an adequate number of patients satisfying the desired sample size requirement to reach a reasonable statistical power in GWASs.

Emerging Directions in Genetic Research
In addition to SNPs, CNVs are becoming more and more used in genetic research. CNVs are deletions or duplication of 1 kb to millions of bases of DNA sequence and are found in different copy numbers among individuals. Even though CNVs are not as abundant as SNPs, because they cover long genomic segments, often involve genes, and thus quantitatively affect gene expression, CNVs are likely to explain a substantial portion of the phenotypic variation among individuals [72]. Recently, there have been several exciting reports in the literature on the use of CNVs in both disease susceptibility and population genetics studies [73, 74]. Such studies involving CNVs can now be extended to genetic prognosis studies to investigate whether these structural genomic variants are associated with variable outcomes and prognoses in cancer patients.

Conclusion
Genetic susceptibility, variable treatment response, and prognosis are likely to be determined, in part, by inherited genetic variations acting as low-penetrance alleles. Thus far, the majority of genetic studies have focused on the identification of low-penetrance disease susceptibility alleles by applying a variety of approaches, such as candidate gene pathway studies and GWASs. Although such genetic studies on treatment response and prognosis are currently scarce, they can nevertheless benefit from previously developed and applied approaches to identify genetic factors that can explain interpatient variability in treatment response and prognosis in cancer. Genetic studies in prognosis and outcome have their own challenges, underlined by the critical need for having similar patient cohorts (i.e., clinical trials) to perform training and testing of results, which can be overcome by collaborations among investigators and improvements in methodological issues. Use of additional genetic variations, such as CNVs, also has the potential to aid identification of the genetic basis of variable prognoses and outcomes in cancer patients. Thus, the role of the clinician in collaboration with scientists to advance this area of research and its application clinically will be much greater in the coming years.

Acknowledgments
We thank many colleagues for valuable discussions and insights. G. Liu is supported by the Alan B. Brown Chair in Molecular Genomics.

Sevtap Savas is now affiliated with the Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John’s, Newfoundland, Canada.

Author Contributions
Conception/Design: Sevtap Savas, Geoffrey Liu
Financial support: Geoffrey Liu
Provision of study materials: Sevtap Savas, Geoffrey Liu
Collection/assembly of data: Sevtap Savas, Geoffrey Liu
Data analysis: Sevtap Savas, Geoffrey Liu
Manuscript writing: Sevtap Savas, Geoffrey Liu
Final approval of manuscript: Sevtap Savas, Geoffrey Liu

References
16 Savas S, Ahmad MF, Shariff M et al. Candidate nsSNPs that can affect the functions and interactions of cell cycle proteins. Proteins 2005;58:697–705.
19 Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics 2004;83:970–979.
46 Gold B, Kirchhoff T, Stefanov S et al. Genome-wide association study pro-


