Analysis of Candidate Genes for Prostate Cancer

James K. Burmester\textsuperscript{a} Brian K. Suarez\textsuperscript{b} Jennifer H. Lin\textsuperscript{c} Carol H. Jin\textsuperscript{d} Raymond D. Miller\textsuperscript{e} Kai-Qi Zhang\textsuperscript{a} Sherry A. Salzman\textsuperscript{a} Douglas J. Reding\textsuperscript{f} William J. Catalona\textsuperscript{g}

\textsuperscript{a}Department of Cancer Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisc., \textsuperscript{b}Departments of Psychiatry and Genetics, Washington University School of Medicine, St. Louis, Mo., \textsuperscript{c}Department of Medicine, Harvard Medical School, Boston Mass., \textsuperscript{d}Department of Psychiatry and \textsuperscript{e}Division of Dermatology, Washington University, St. Louis, Mo., \textsuperscript{f}Department of Oncology, Marshfield Clinic, Marshfield, Wisc., \textsuperscript{g}Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, Ill., USA

Key Words
Prostate cancer · Genetics · Familial disease · Aggressiveness · Metastasis · Single nucleotide polymorphism

Abstract
Considerable evidence demonstrates that genetic factors are important in the development and aggressiveness of prostate cancer. To identify genetic variants that predispose to prostate cancer we tested candidate SNPs from genomic regions that show linkage to prostate cancer susceptibility and/or aggressiveness, as well as genes that show a significant difference in mRNA expression level between tumor and normal tissue. Cases had histologically verified prostate cancer. Controls were at least 65 years old, never registered a PSA above 2.5 ng/ml, always had digital rectal examinations that were not suspicious for cancer, and have no known family history of prostate cancer. Thirty-nine coding SNPs and nine non-coding SNPs were tested in up to 590 cases and 556 controls resulting in over 40,000 SNP genotypes. Significant differences in allele frequencies between cases and controls were observed for ID3 (inhibitor of DNA binding), $p = 0.05$, HPN (hepsin), $p = 0.009$, BCAS1 (breast cancer amplified sequence 1), $p = 0.007$, CAV2 (caveolin 2), $p = 0.007$, EMP3 (epithelial membrane protein 3), $p < 0.0001$, and MLH1 (mutL homolog 1), $p < 0.0001$. SNPs in three of these genes (BCAS1, EMP3 and MLH1) remained significant in an age-matched subsample.

Objectives
Prostate cancer is the most common non-cutaneous cancer in men in the United States, accounting for 31\% of new cases. Prostate cancer is the second most common cause of death from cancer in men aged 60 years old and older [1]. Metastatic prostate cancer is currently incurable and is more likely to occur when the cancer has extended to, or spread beyond the capsule of the prostate gland. In about 30\% of all cases, cancer has extended to the surgical margin or beyond at the time of diagnosis [2]. The aggres-
siveness of prostate tumors, measured by Gleason score, varies substantially, with some tumors remaining indolent and others becoming life-threatening, suggesting that individual variation contributes to tumor aggressiveness.

Germline mutations are estimated to account for approximately 9% of all prostate cancers and 45% of cases in men younger than 55 years of age [3, 4]. Numerous linkage mapping studies have identified many regions throughout the genome likely to contain genes that predispose to prostate cancer [5, 6].

Our whole genome linkage analysis of affected brothers with prostate cancer was used to identify regions of the genome likely to contain genes that predispose individuals to prostate cancer and/or aggressive disease [7–9]. In our cohort, regions of the genome that were associated with prostate cancer included 2q, 12p, 15q, 16p, and 16q, with the strongest signal occurring on chromosome 16q. Several additional regions were identified when the data were stratified based on positive family history for hereditary prostate cancer, median age-at-diagnosis, and family history of breast cancer. The subgroup with a family history of prostate cancer had a signal on chromosome 1 proximal to HPC1 (OMIM#601518). The subgroup with breast cancer revealed a strong linkage signal at 1p35.1. The families with late age-at-diagnosis had a signal on chromosome 4q. Analysis of the same data set using Gleason grade as an index of tumor aggressiveness demonstrated prostate cancer aggressiveness loci at 5q31–33, 7q32, and 19q12 [10, 11].

Here we report results from the analysis of 48 SNPs from genomic regions reported to be linked to prostate cancer susceptibility or tumor aggressiveness, as well as for genes reported to differ in expression level between normal prostate and tumor. Non-synonymous coding SNPs were emphasized in this study because these variants have the potential to change protein/enzyme activity. Non-coding SNPs in hepsin (HPN) and EZH2 were also tested because these genes have recently been reported as associated with aggressive disease at the RNA and/or protein level, and these genes reside on chromosomes 7 and 19, respectively, within regions of linkage to aggressive disease. We identified significant associations with prostate cancer for 6 genes; the association remained significant for three genes when cases and controls were matched for age.

Materials and Methods

Participants

Participants in this study were from 275 multiplex prostate cancer sibships and 556 unrelated controls. Twenty-five patient samples were excluded from analysis due to insufficient quantities of DNA. All of the multiplex sibships were ascertained from patients seen at Washington University School of Medicine by staff urologists or were referred by other urologists, or were participating in prostate cancer support groups, or responded to published solicitations.

Control subjects were followed for many years as part of a long-term prostate cancer screening study in which men were screened at 6- to 12-month intervals with prostate-specific antigen (PSA) blood tests and digital-rectal examination (DRE) of the prostate [12]. The control subjects were required to meet the following four criteria: (a) be at least 65 years old, (b) never have registered a PSA level above 2.5 ng/ml, (c) always had DRE findings that were not suspicious for prostate cancer, and (d) have no known family history of prostate cancer. Family history of prostate cancer was assessed by inquiring about the subjects’ brothers, fathers, grandfathers and maternal and paternal uncles. As a result of the first criterion, the mean age of control subjects was greater than the mean age of case subjects. All of the subjects in this study were of European ancestry. The protocol for this study was approved by the Human Studies Committee of Washington University and the Institutional Review Board of Marshfield Clinic Research Foundation. Written informed consent was obtained from all participants.

SNP Selection, Validation and Genotyping

Non-synonymous coding SNPs were selected from public databases in 2001. Candidate SNPs were selected from genomic regions that had yielded evidence of linkage in our family studies of multiplex prostate cancer sibships: 1p36, 3q36-37, 3p22-21, 8p21-p12, 12p13, 15q13-14, 16p13, 16q23-24, 21q22.3 [7]; from linkage regions highlighted in our QTL analysis of Gleason scores [10]: 3p22, 6p22-21, 7q31-32, 19q12-13, or were selected from genomic regions that other groups, at the time, had reported as linked to prostate cancer, i.e., 1q24-25, 1q42-43, 5q32-33 [13]; 8q22-23, 11p15-13 [14]; 17p13-11 [15]; and 20q11-13 [16]. Only candidate genes whose transcripts are expressed in prostate were considered. Our SNP search used MAP Viewer (http://www.ncbi.nlm.nih.gov/mapview) and UniGene (http://www.ncbi.nlm.nih.gov/UniGene) to identify candidate genes.

Non-synonymous coding SNPs were also identified by searching gene expression databases. Serial analysis of gene expression (SAGE) was used to identify transcripts that showed a two-fold or greater difference between normal prostate tissue and prostate tumor tissue using 3 libraries of normal prostate bulk tissue (SAGE_Chen_Normal_Pr, SAGE_PR317_normal_prostate, and SAGE_normal_prostate), and comparing these with three pooled libraries of prostate tumor tissue (SAGE_Chen_Tumor_Pr, SAGE_PR317_prostate_tumor, and SAGE_PrCA-I). We required a two-fold difference in expression level as well as the Bayesian posterior probability of the difference to be >50%. We were aware that many alleged SNPs, especially in the early databases, were in fact not polymorphic. These errors were more likely to occur for coding SNPs that were the main focus of our investigation. Accordingly, after we had compiled the list of putative coding SNPs for candidate genes, the first order of business was to validate the SNPs in a test panel of DNAs. From a total of 144 candidate SNPs that we identified from our search of expression
SNPs tested in cases and controls. The major allele is the first one listed under variant.

a  HWE = Test for Hardy Weinberg Equilibrium.

libraries 26 could be validated in Caucasian samples and three are analyzed in this report. Analysis of the remaining SNPs is currently in progress and will be available in a future report. Single nucleotide polymorphisms were validated by amplification in pooled anonymous genomic DNAs (5 genomic DNAs/pool, 60 DNAs total). Amplification primers were chosen in regions flanking SNPs to produce amplicons approximately 100 base pairs in length. Amplicons were digested with restriction enzymes recognizing the DNA sequence encompassing the SNP site. Digested amplicons were electrophoresed through a 3% Agarose 1000 (Invitrogen, Hum Hered 2004;57:172–178

Burmeister/Suarez/Lin/Jin/Miller/Zhang/
Salzman/Reding/Catalona

174
Carlsbad, Calif., USA agarose gel containing ethidium bromide. Fragments were analyzed for changes in restriction digestion patterns defining SNP sites.

DNA polymorphisms were detected as described [17], using polymerase chain reaction followed by detection with the Invader assay (Third Wave Technologies, Madison, WI). For EMP1, about 1% of samples were checked by sequencing to verify the genotype. All samples that were checked had the same genotype as that determined in the Invader assay. Approximately 25% of MLH1 genotypes were repeated with the Invader assay to verify the genotypes.

**Statistical Analysis**

Maximum likelihood allele frequency estimates for the cases and controls were obtained from the USERM13 subroutine of MENDEL [18, 19]. Allele frequencies in cases and controls were compared using a likelihood ratio test. Let $L_A$, $L_C$ and $L_T$ denote the respective maximum likelihoods for the cases, controls and the total sample (i.e., all cases and controls combined). The quantity $-2(\ln L_A + \ln L_C - \ln L_T)$ is asymptotically distributed as a $\chi^2$ with one degree of freedom for all of the SNPs reported in table 1 (20).

The general design of a case/control study attempts to match cases with controls, either individually or collectively, on all variables thought to be relevant except the variable under study. For genetic studies of the sort reported here, it is critical to draw cases and controls from the same ‘breeding population.’ For this reason, we have restricted this analysis to American men who are of European ancestry. For complex diseases with a variable age-of-onset, however, opinion is divided over the issue of strictly matching for age. Because of their interest in estimating relative-risks, for instance, epidemiologists tend to insist that age always be a matching variable. Many gene-hunting geneticists, however, prefer a form of ‘extreme sampling’ that calls for choosing the controls to be exceptionally healthy and as far through the ‘risk-period’ as possible. This design is expected to result in a control sample that is enriched for low-risk alleles, thereby allowing increased power to identify susceptibility loci when compared to the cases. The danger of this extreme sampling design for the present study is that we could misidentify a SNP allele as increasing risk for the development of prostate cancer when, in fact, the alternative allele is associated with longevity. To guard against this possibility, we retest each significant case/control difference by excluding the younger cases and the oldest controls, thereby creating a post-hoc age-matched subgroup of cases and controls (see fig. 1). The price paid for this strategy is a reduced sample size for both cases and controls with a concomitant loss of power. On the other hand, the prospect of committing a type II error is likely reduced using this extreme sampling strategy.

**Results**

We have completed genotyping for 39 coding SNPs and 9 non-coding SNPs in up to 590 cases and 556 healthy controls. More than 40,000 genotypes were determined for the 48 SNPs reported here.

Table 1 gives the name of the gene, the SNP rs-number, the nucleotide substitution, the chromosomal map position of the gene, and the major allele frequencies in controls and cases for each SNP. Additionally, table 1 reports the likelihood ratio ($\hat{\chi}^2$) for equality of allele frequencies and a test of Hardy-Weinberg equilibrium (HWE) in the control sample. The purpose of the HWE test is to alert us to the possibility that there might be genotypic errors for a particular SNP. Accordingly, when the SNP fails the HWE test, we put little faith in the case/control comparison of allele frequencies. All of the SNPs reported in table 1 are in HWE. The boldface entries of table 1 highlight those SNPs for which a significant difference in allele frequency between cases and controls is evident.

**Fig. 1.** Age adjustment of cases and controls. Because controls were required by design to be a minimum of 65, cases are significantly younger than controls. To obtain age-matched subsamples, all cases with an age-at-diagnosis less than 65 (left stippled area) were discarded. Controls, sequentially starting with the oldest (right stippled area), were discarded until the mean age of the remainder matched the mean age of the cases. For the 7 significant SNPs reported in table 1, the age-matched samples contain an average of 45.9% fewer cases and 13.2% fewer controls.
Table 2. Results of age matching cases and controls for SNPs that revealed a significant difference in allele frequencies in the total sample

<table>
<thead>
<tr>
<th></th>
<th>Controls total sample</th>
<th>Controls age-matched sample</th>
<th>Cases total sample</th>
<th>Cases age-matched sample</th>
<th>Δ</th>
<th>Likelihood ratio $\chi^2$</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3</td>
<td>0.781</td>
<td>0.780</td>
<td>0.739</td>
<td>0.738</td>
<td>0.000</td>
<td>3.122</td>
<td>0.078</td>
</tr>
<tr>
<td>HPN</td>
<td>0.913</td>
<td>0.921</td>
<td>0.948</td>
<td>0.937</td>
<td>0.019</td>
<td>0.955</td>
<td>0.328</td>
</tr>
<tr>
<td>HPN</td>
<td>0.915</td>
<td>0.919</td>
<td>0.939</td>
<td>0.933</td>
<td>0.010</td>
<td>0.815</td>
<td>0.367</td>
</tr>
<tr>
<td>BCAS1</td>
<td>0.526</td>
<td>0.524</td>
<td>0.592</td>
<td>0.590</td>
<td>0.000</td>
<td>5.108</td>
<td>0.024</td>
</tr>
<tr>
<td>CAV2</td>
<td>0.798</td>
<td>0.800</td>
<td>0.851</td>
<td>0.829</td>
<td>0.024</td>
<td>1.723</td>
<td>0.189</td>
</tr>
<tr>
<td>EMP3</td>
<td>0.987</td>
<td>0.987</td>
<td>0.955</td>
<td>0.959</td>
<td>0.004</td>
<td>10.162</td>
<td>0.001</td>
</tr>
<tr>
<td>MLH1</td>
<td>0.728</td>
<td>0.719</td>
<td>0.629</td>
<td>0.612</td>
<td>−0.008</td>
<td>16.352</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

It is evident that the drop in significance in the age-matched sample is simply the result of the reduction in sample size since the $\Delta$ is zero. The change in significance for $HPN$, and $CAV2$ is due to both the positive $\Delta$ and the reduced sample size.

**Conclusions**

We have observed 7 SNPs in 6 different genes that show a significant association with prostate cancer. Three of these remain significant after the cases and controls are matched for mean age. Except for $HPN$ and $CAV2$, the drop in significance appears to be due to the decrease in sample size that accompanied the age matching procedure. Indeed, for these 7 SNPs, the mean $\Delta$ between the total and age-matched sample is 0.007.

$ID3$ encodes a helix-loop-helix protein that controls cell differentiation by interfering with DNA binding of transcription factors [21]. The polymorphism changes Ala at amino acid 105 to Thr. $HPN$ codes for a cell surface protease that has been associated with aggressive prostate cancer at the mRNA and protein level [22]. The polymorphism in $HPN$ is located within the intron. $BCAS1$ codes for a protein of unknown function that is over-expressed in breast cancer [23]. The variant amino acid is Lys 24 Gln. $CAV2$ produces a membrane bound protein that associates with caveolin 1 and caveolin 3 forming specialized membrane invaginations [24]. Although the role of $CAV2$ in prostate cancer has not been studied, caveolin 1 is associated with increased cell survival and contributes to metastasis in androgen-insensitive prostate cancer [25]. The amino acid that is variable in $CAV2$ is Glu 130 Gln.

The function of the protein coded by the $EMP3$ gene is largely unknown, but may be involved in regulating cell proliferation, differentiation and death [26]. The polymorphism in $EMP3$ changes amino acid 125 from Val to Ile. $MLH1$ codes for a DNA repair enzyme associated with microsatellite instability in colon cancer [27]. The observed polymorphism changes Val to Ile at amino acid 219. Any mismatch DNA repair enzyme is a natural candidate for a cancer susceptibility gene.

Our selection process for identifying candidate SNPs included both candidates from regions of linkage in whole genome scans as well as genes differentially expressed between normal prostate and tumor. All of the significant genes that we identified were located in regions of linkage. We did not identify any significant associations between genes chosen for differential expression pattern, although we have only tested a limited number of these genes to date. $HPN$ fits both selection criteria since it is located in a region of aggressive disease and its expression level differs between normal prostate and tumor.

The observation that the coding SNP in $CAV2$ in the total sample is associated with prostate cancer is interesting since previous studies have demonstrated an association between $CAV1$ and prostate cancer. $CAV1$ is associated with androgen-independent disease and with stimulation of prostate cell growth [25]. It has been demonstrated that $CAV1$ interacts with androgen receptor as a positive modulator of transactivation [28]. Our results support an interaction between $CAV1$ and $CAV2$ as part of an important regulatory pathway.

The function of the $EMP3$ protein is largely unknown but may be involved in regulating cell proliferation, differentiation and death since over expression of $EMP3$ results in cell blebbing, and activation of the caspase
apoptosis pathway [26]. It is a 163-amino acid protein that contains 4 trans-membrane domains and is a member of the protein family consisting of PMP-22, EMP1 and EMP2. PMP-22 was first identified due to its growth arrest properties.

The association between MLH1 and prostate cancer is important since it suggests a mechanism of genome instability in patients who inherit the variant allele. MLH1 is a DNA repair enzyme that functions with MSH2, MSH6, PMS1 and PMS2 to create a functional DNA repair system responsible for proofreading during DNA synthesis [29]. Germline mutations of MLH1 and MSH2 are frequent in hereditary non-polyposis colon cancer and are frequently characterized due to microsatellite instability, a global measure of genome mutation. Significantly, microsatellite instability in prostate cancer has been reported, but the actual rate of microsatellite instability has varied considerably between different studies [30]. Complementary studies have demonstrated decreased staining of PMS2, MSH2, and MLH1 in prostate tumors. Our results suggest it may be possible to identify patients who have variants of MLH1 at the genomic level and correlate this variant with increased mutation rate in tumors detectable as microsatellite instability.

Our linkage peak on chromosome 7q35 as well as the linkage peak on chromosome 19q was broad, covering approximately 30 and 10 cM respectively. Our finding of a significant association for HPN and EMP3 on chromosome 19 raises the possibility that there are multiple genes involved in prostate cancer aggressiveness within this region.

Acknowledgements

Supported by an award from the Urological Research Foundation.

References


