Decision Tree–Based Modeling of Androgen Pathway Genes and Prostate Cancer Risk

Jill S. Barnholtz-Sloan1,2, Xiaowei Guan1,2, Charnita Zeigler-Johnson4, Neal J. Meropol1,3, and Timothy R. Rebbeck4

Abstract

Background: Inherited variability in genes that influence androgen metabolism has been associated with risk of prostate cancer. The objective of this analysis was to evaluate interactions for prostate cancer risk by using classification and regression tree (CART) models (i.e., decision trees), and to evaluate whether these interactive effects add information about prostate cancer risk prediction beyond that of “traditional” risk factors.

Methods: We compared CART models with traditional logistic regression (LR) models for associations of factors with prostate cancer risk using 1,084 prostate cancer cases and 941 controls. All analyses were stratified by race. We used unconditional LR to complement and compare with the race-stratified CART results using the area under curve (AUC) for the receiver operating characteristic curves.

Results: The CART modeling of prostate cancer risk showed different interaction profiles by race. For European Americans, interactions among CYP3A43 genotype, history of benign prostate hypertrophy, family history of prostate cancer, and age at consent revealed a distinct hierarchy of gene–environment and gene–gene interactions, whereas for African Americans, interactions among family history of prostate cancer, individual proportion of European ancestry, number of GGC androgen receptor repeats, and CYP3A4/CYP3A5 haplotype revealed distinct interaction effects from those found in European Americans. For European Americans, the CART model had the highest AUC whereas for African Americans, the LR model with the CART discovered factors had the largest AUC.

Conclusion and Impact: These results provide new insight into underlying prostate cancer biology for European Americans and African Americans.

Introduction

Among American men, prostate cancer has the highest incidence of any noncutaneous tumor and is a leading cause of cancer-related mortality (1). In addition, the incidence of prostate cancer is over twice as high in African American men as compared with any other racial group in the United States (2). Although the causes of prostate cancer are not well understood and are likely to involve many factors, androgen metabolism genotypes have been hypothesized to be involved in prostate cancer etiology (3–6). Recently, there has been significant debate about the utility of inherited genetic information in clinical applications including risk assessment (7). These concerns relate both to the proportion of variability in clinically relevant traits that can be explained by common variants and the ability to translate this information to clinical practice. Therefore, a critical research question is whether genotype adds information to risk prediction beyond that of “traditional” risk factors, such as age, race, or family history of prostate cancer (as reviewed in refs. 8, 9).

Among the pathways that have been identified as playing an important role in prostate cancer etiology is that of androgen (testosterone) metabolism (Fig. 1). Testosterone is a major determinant of prostate growth and differentiation. Although serum levels of testosterone do not correlate well with prostate cancer risk, serum levels of dihydrotestosterone (DHT) and other testosterone metabolites do correlate with prostate cancer risk (10, 11). Androgens are related to the growth and development of prostate tumors and androgen ablation in men with hormone-sensitive prostate tumors reduces tumor size and decreases the associated disease burden (12). There are several enzymes that determine the activation or inactivation of testosterone,
which subsequently influences the signaling capability of testosterone metabolites in androgen-sensitive cells and are all involved in the androgen receptor (AR) pathway. These genes include the AR, the 5 alpha-reductase type II (SRD5A2), and the cytochromes p450 genes, CYP3A4, CYP3A5, and CYP3A43 (Fig. 1). In addition, electron spin resonance 1 (ESR1), which encodes the estrogen receptor α, might increase prostate cancer risk, although evidence is conflicting (13–15).

Among the possible explanations for the lack of replication of associations between these genes and prostate cancer is that the etiology of this disease may involve interactions among many risk genotypes. This may be particularly true of alleles that influence common metabolic pathways such as those involved in androgen metabolism. Thus, studies that focus on the univariable effect of a single gene do not detect relevant joint effects of multiple genes acting in a common pathway. On the basis of these observations, the objective of this analysis was to evaluate gene–gene and gene–environment interactions for prostate cancer risk by using classification and regression tree (CART) models and to evaluate whether these genotypic effects add information about prostate cancer risk prediction beyond that of "traditional" risk factors, such as age, race, or family history of prostate cancer (as reviewed in refs. 8, 9, 16, and 17).

Materials and Methods

Study participants and variables of interest
Incident prostate cancer cases were identified through Urologic Oncology Clinics at multiple hospitals of the University of Pennsylvania Health System (UPHS) between 1995 and 2008. Controls were men attending UPHS general medicine clinics and were ascertained concurrently with the prostate cancer cases (i.e., between 1995 and 2008; ref. 3). Our final study population consisted of 1,470 total European Americans (931 cases and 539 controls) and 555 African Americans (153 cases and 402 controls). Cases and controls were excluded if they were of "Other" race, or they had missing genotype data.

Variables of interest used in analysis were: age at consent for both cases and controls; self-reported race (European American, African American), family history of prostate cancer (yes/no; first-degree relatives only), personal history of benign prostate hypertrophy (BPH; yes/no; i.e., history of BPH), and individual maximum likelihood estimated European ancestry proportion.

Biosample collection and genotype analysis
Genomic DNA for this study was self-collected by each study participant by using sterile cheek swabs (Cyto-Pak Cytosoft Brush; Medical Packaging Corporation) and processed by using either a protocol modified from Richards and colleagues (18) as described previously (19), or by using a Qiagen 9604 robot with the QIAamp 96 DNA Buccal Swab Biorobot Kit. DNA extraction was undertaken without knowledge of case–control status, race, age, or any other variable. Extractions were undertaken in batches that included both cases and controls and individuals of all races. Previous evaluations of these extraction protocols did not reveal any evidence for differential bias in extraction success due to case status, gender, age, or other demographic characteristics (19). Genotypes were determined for putatively functional variants in a series of candidate androgen metabolism genes (Fig. 1). These genes were chosen to evaluate whether combinations of biologically plausible candidate susceptibility genotypes in a well-defined pathway might provide evidence for multivariable associations, even though the main effects of these genes have not clearly been associated with prostate cancer.

SRD5A2, CYP3A4, CYP3A5, and CYP3A43 genotypes and AR repeats were determined by using Pyrosequencing protocols accompanied by PCR-RFLP assays as previously described (3, 20). Two SNPs were genotyped in SRD5A2, rs523349, and rs9282858, with genotype call rates of 81.2% and 79.7%, respectively. One SNP was genotyped in CYP3A4, rs2740574, with a genotype call rate of 91.0%. One SNP was genotyped in CYP3A5, rs10249369, with a genotype call rate of 92.4%. One SNP was genotyped in CYP3A43, rs680055, with a genotype call rate of 75.0%. AR repeats were determined by using information from 2 SNPs, rs3138869 (GGC repeat, GGN repeat) and rs4045402 (CAG repeat), with call rates of 84.3% and 77.7%, respectively. ESR1 genotypes were...
obtained as part of a multiplex, custom candidate gene SNP panel assayed by using the Illumina GoldenGate platform. One SNP was genotyped in ESR1, rs3853250, with a genotype call rate of 98.8%.

Ancestry informative markers and ancestry estimation

A panel of 158 ancestry informative markers (AIM) was genotyped as part of a multiplex, custom candidate gene SNP panel assayed by using the Illumina GoldenGate platform. These AIMs were chosen to be maximally informative for distinguishing between African and European ancestries (21–23) and have been described elsewhere (24). Individual estimates of European ancestry were calculated from 149 AIMs by using maximum likelihood methods as reviewed in ref. 22; 9 AIMs failed genotyping. Average and median proportions of European ancestry were 0.97 and 0.99 for European Americans and 0.22 and 0.18 for African Americans, respectively.

Statistical analysis

Lewontin’s $D^*$ was calculated as a measure of linkage disequilibrium between SNPs within the same gene and between genes on the same chromosomal location by using SAS Genetics (25). SNPs were selected for haplotype analysis if Lewontin’s $D^*$ values were 0.7 or more. Haplotypes for CYP3A4/CYP3A5 were made from CYP3A5*3C (rs10249369) and CYP3A4*1B (rs2740574; Lewontin’s $D^* = 0.7571$) and for SRD5A2 were made from SRD5A2 (v89L; rs523349) and SRD5A2 (A97T; rs9282585; Lewontin’s $D^* = 1.00$). CYP3A4*3 (rs680055) and ESR1 (rs3853250) were analyzed as single SNPs only because they are not in linkage disequilibrium with any other SNPs, and thus were not eligible for haplotype analysis. AR repeats were analyzed as a continuous variable for each type of repeat (CAG, GGN, and GGC) based on previous literature (26, 27). Unconditional logistic regression (LR) analysis to test for SNP (additive “per allele” models) and AR repeat length associations with prostate cancer risk stratified by race generating odds ratios (OR) and 95% confidence intervals (95% CI), adjusted for family history of prostate cancer, age at consent, individual European ancestry, and history of BPH. Haplotypes were discovered and scored as implemented in haplo.stats in R (28), using the race-specific common haplotype as the referent. Generalized linear models to test for haplotype associations with prostate cancer risk stratified by race were used and generated ORs and 95% CIs for each haplotype, adjusted for family history of prostate cancer, age at consent, individual European ancestry, and history of BPH.

SNPs and haplotypes were then further tested for association with prostate cancer risk, along with other potential variables of interest, using 2 methods: (i) standard unconditional multivariable LR analysis and (ii) decision tree–based analysis (i.e., CART analysis; ref. 29). CART is a binary recursive partitioning tree modeling technique that allows for the hierarchical modeling of interactions between variables of interest associated with risk of prostate cancer. For node size restrictions, the algorithm required a minimum node size of 5 individuals. Variables were included at each possible split to evaluate whether they improved the node purity. Nodes were split by using the best split values, which maximizes the Gini index splitting criterion. After initial tree growing from top to bottom, trees were pruned at the cost complexity value which minimizes the mean square error for each split. Final trees were grown and validated by using 10-fold cross-validation. The left-most node on each tree, representing a control group of subjects, was used as the reference node for calculation of ORs with 95% CIs. All CART models considered all information on all genes within the AR pathway and the other variables of interest. The area under curve (AUC) and its 95% CI for the receiver operating characteristic (ROC) curves for each LR model and each CART tree were calculated and compared by using the ROC package in R. AUC values from the final LR models were compared with the baseline LR model in a pairwise fashion by race using the $\chi^2$ statistic. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for each terminal node of the final pruned CART models by race.

Results

Baseline characteristics of our study population by race/ethnicity are shown in Table 1. In European Americans, CYP3A5*3C was associated with increased risk of prostate cancer (OR = 2.11; 95% CI: 1.13–3.97; Table 2); increasing number of CAG repeats in the AR gene was associated with decreased risk of prostate cancer (OR = 0.98; 95% CI: 0.83–0.98; Table 2); and the CYP3A4/CYP3A5 AG haplotype was associated with increased risk of prostate cancer (OR = 2.86; 95% CI: 1.15–7.12; Table 3).

No significant SNP or haplotype associations were found in African Americans (Tables 2 and 3).

The CART decision tree–based modeling of prostate cancer risk showed that age at consent and family history of prostate cancer were significant predictors in both races (Figs. 2 and 3). For European Americans, the final pruned decision tree included the CYP3A43 genotype, family history of prostate cancer, age at consent and history of BPH (Fig. 2). As compared with the reference node, men with the CYP3A43 GC or CC genotype, no family history of prostate cancer, age at consent 49 or less and a history of BPH had a more than 8-fold risk of prostate cancer (OR = 8.77; 95% CI: 7.27–10.28). Multiple other significant risk groups were discovered including those with the CYP3A43 GG genotypes only (OR = 21.37; 95% CI: 20.52–22.23), those with the CYP3A43 GC or CC genotype and family history of prostate cancer (OR = 4.95; 95% CI: 4.56–5.34) and those with the CYP3A43 GC or CC genotype, no family history of prostate cancer and age at consent between 50 and 70 (OR = 2.81; 95% CI: 2.52–3.09). The interactions among CYP3A43, the history
of BPH, family history of prostate cancer, and age at consent reveal the hierarchy of the gene–environment factors that further attenuates the risk of prostate cancer in this racial group. The terminal node which included interactions between the CYP3A43 genotype, the history of BPH, family history of prostate cancer, and age at consent had a specificity of 99% and a PPV of 87%, whereas the terminal node which included only the CYP3A43 genotype also had the same specificity, but a higher PPV of 94% (Table 4).

The final pruned African American decision tree included family history of prostate cancer, individual European ancestry proportion, number of GGC AR repeats, and the CYP3A4/CYP3A5 haplotype (Fig. 3). The 2 risk groups were characterized as having (i) family history of prostate cancer, individual European ancestry proportion less than 20.4% and number of GGC repeats less than 16 [OR = 5.46 (4.21, 6.70), and (ii) family history of prostate cancer, individual European ancestry proportion 20.4% or more and CYP3A4/CYP3A5 haplotypes GA, AG or GG (OR = 6.24; 95% CI: 5.30–7.17). The interactions among family history of prostate cancer, individual proportion of European ancestry, number of GGC AR repeats, and CYP3A4/CYP3A5 haplotype revealed gene–environment effects that further attenuate the risk of prostate cancer in African Americans. The terminal node which included interactions between the family history of prostate cancer, European ancestry, and AR GGC repeats had a specificity of 90% and a PPV of 64%, whereas the terminal node which included only family history of prostate cancer had a specificity of 99% and a PPV of 87%. (Table 4).

### Table 1. Baseline characteristics of prostate cancer cases and controls stratified by race

<table>
<thead>
<tr>
<th></th>
<th><strong>European American (n, column %)</strong></th>
<th><strong>African American (n, column %)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Cases</strong></td>
<td><strong>Controls</strong></td>
</tr>
<tr>
<td><strong>Family history of prostate cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>207 (22.23)</td>
<td>51 (9.46)</td>
</tr>
<tr>
<td>No</td>
<td>684 (73.47)</td>
<td>482 (89.42)</td>
</tr>
<tr>
<td>Missing</td>
<td>40 (4.3)</td>
<td>6 (1.11)</td>
</tr>
<tr>
<td><strong>Personal history of benign prostate hypertrophy (history of BPH)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>189 (20.3)</td>
<td>52 (9.65)</td>
</tr>
<tr>
<td>No</td>
<td>699 (75.08)</td>
<td>476 (88.31)</td>
</tr>
<tr>
<td>Missing</td>
<td>43 (4.62)</td>
<td>11 (2.04)</td>
</tr>
<tr>
<td><strong>European ancestry proportion (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;25</td>
<td>519 (55.75)</td>
<td>363 (67.35)</td>
</tr>
<tr>
<td>25–75</td>
<td>1 (0.11)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>&gt;75</td>
<td>411 (44.15)</td>
<td>176 (32.65)</td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

**NOTE:** The number of cases and controls for European American are 931 and 539 and for African American are 153 and 402, respectively. The average age for European American is 59.58 and for African American is 58.15.

### Table 2. SNP and androgen repeat analysis for androgen pathway genes by race (all models adjusted by family history of prostate cancer, age at consent, individual European ancestry, and history of BPH)

<table>
<thead>
<tr>
<th>Gene</th>
<th>rsNumber</th>
<th>Alleles</th>
<th>Referent allele frequency (%)</th>
<th>Additive (per allele) OR (95% CI)</th>
<th>Alleles</th>
<th>Referent allele frequency (%)</th>
<th>Additive (per allele) OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP3A5</td>
<td>rs10249369</td>
<td>A/G</td>
<td>92.72</td>
<td>2.11 (1.13–3.97)</td>
<td>G/A</td>
<td>61.33</td>
<td>0.59 (0.28–1.22)</td>
</tr>
<tr>
<td>CYP3A4</td>
<td>rs2740574</td>
<td>G/A</td>
<td>95.19</td>
<td>1.17 (0.54–2.54)</td>
<td>A/G</td>
<td>59.90</td>
<td>1.20 (0.56–2.56)</td>
</tr>
<tr>
<td>CYP3A43</td>
<td>rs680055</td>
<td>G/C</td>
<td>86.04</td>
<td>1.31 (0.78–2.19)</td>
<td>G/C</td>
<td>62.66</td>
<td>0.75 (0.35–1.57)</td>
</tr>
<tr>
<td>SRD5A2</td>
<td>rs523349</td>
<td>C/G</td>
<td>70.54</td>
<td>1.45 (0.98–2.14)</td>
<td>C/G</td>
<td>72.54</td>
<td>0.85 (0.41–1.77)</td>
</tr>
<tr>
<td>SRD5A2</td>
<td>rs3282858</td>
<td>A/G</td>
<td>97.36</td>
<td>1.57 (0.56–4.31)</td>
<td>A/G</td>
<td>99.04</td>
<td>1.71 (0.08–34.82)</td>
</tr>
<tr>
<td>ESR1</td>
<td>rs3853250</td>
<td>C/A</td>
<td>54.17</td>
<td>1.37 (0.91–2.05)</td>
<td>A/C</td>
<td>54.22</td>
<td>1.08 (0.57–2.04)</td>
</tr>
<tr>
<td>AR</td>
<td>CAG (rs4045402)</td>
<td>A/G</td>
<td>92.72</td>
<td>2.11 (1.13–3.97)</td>
<td>G/A</td>
<td>61.33</td>
<td>0.59 (0.28–1.22)</td>
</tr>
<tr>
<td>AR</td>
<td>GGN (rs3138869)</td>
<td>A/G</td>
<td>95.19</td>
<td>1.17 (0.54–2.54)</td>
<td>A/G</td>
<td>59.90</td>
<td>1.20 (0.56–2.56)</td>
</tr>
<tr>
<td>AR</td>
<td>GGC (rs3138869)</td>
<td>A/G</td>
<td>97.36</td>
<td>1.57 (0.56–4.31)</td>
<td>A/G</td>
<td>99.04</td>
<td>1.71 (0.08–34.82)</td>
</tr>
</tbody>
</table>

www.aacrjournals.org Cancer Epidemiol Biomarkers Prev; 20(6) June 2011 1149

Published OnlineFirst April 14, 2011; DOI: 10.1158/1055-9965.EPI-10-0996

Downloaded from cebp.aacrjournals.org on October 14, 2017. © 2011 American Association for Cancer Research.
prostate cancer had a specificity of 85% and much lower PPV of 42% but a reasonable NPV of 76% (Table 4).

Race-stratified ROC curves were used to compare 5 models within each racial group: (i) a CART model with only the adjustment factors of age, European ancestry, family history of cancer, and history of BPH (CARTBASE; ii), a final pruned CART model (CART); (iii) a LR model including only the adjustment factors significant at the 0.05 level, which were family history of prostate cancer and history of BPH for European Americans and family

Table 3. Haplotype analysis for CYP3A4/CYP3A5 and SRD5A2 by race (all models adjusted by family history of prostate cancer, age at consent, individual European ancestry, and history of BPH)a

<table>
<thead>
<tr>
<th>Race</th>
<th>Gene</th>
<th>Haplotype</th>
<th>Haplotype frequency (%)</th>
<th>OR (95% CI)</th>
<th>P (T-statistic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>European American</td>
<td>CYP3A4/CYP3A5</td>
<td>GA</td>
<td>86.52</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AG</td>
<td>3.50</td>
<td>2.86 (1.15–7.12)</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>2.74</td>
<td>1.33 (0.59–3.05)</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td>0.65</td>
<td>0.49 (0.14–1.73)</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>SRD5A2</td>
<td>GG</td>
<td>56.24</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CG</td>
<td>24.60</td>
<td>1.36 (0.98–1.89)</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>2.76</td>
<td>1.72 (0.65–4.56)</td>
<td>0.273</td>
</tr>
<tr>
<td>African American</td>
<td>CYP3A4/CYP3A5</td>
<td>AG</td>
<td>30.40</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>8.73</td>
<td>0.89 (0.40–1.99)</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>14.64</td>
<td>1.14 (0.62–2.10)</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td>9.27</td>
<td>0.64 (0.30–1.37)</td>
<td>0.251</td>
</tr>
<tr>
<td></td>
<td>SRD5A2</td>
<td>GG</td>
<td>38.59</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CG</td>
<td>15.40</td>
<td>1.17 (0.69–1.98)</td>
<td>0.563</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>0.36</td>
<td>1.88 (0.09–37.68)</td>
<td>0.681</td>
</tr>
</tbody>
</table>

aHaplotypes made from CYP3A5*3C and CYP3A4*1B (Lewontin's D′ = 0.7571); Haplotypes made from SRD5A2(V89L) and SRD5A2 A49T (Lewontin's D′ = 1.00).

Figure 2. Pruned classification tree for androgen pathway in European Americans. The numbers within each node indicate the number of controls/the number of prostate cancer cases.
history of cancer, history of BPH, and European ancestry proportion for African Americans (LRBASE); (iv), a LR model including all main effects and interaction terms as found by CART (LRCART); and (v), a fully saturated backward selection LR model with all interactions (LRFULL; Fig. 4). For European Americans, the pruned CART tree had the highest AUC (0.686). A comparison between the LR models for European Americans showed that both the LRFULL and the LRCART models performed as well as the LRBASE model ($\chi^2$ P value for LRFULL vs. LRBASE = 0.11; LRCART vs. LRBASE = 0.09). For African Americans, the LRCART model that included the final variables found by CART had the highest AUC (0.680). A comparison between the LR models for African Americans showed that both the LRFULL and the LRCART models performed better than the LRBASE model ($\chi^2$ P value for LRFULL vs. LRBASE = 0.85; LRCART vs. LRBASE = 0.51). These results show that CART was able to find novel interactions within these data that would not have been found by using traditional LR approaches.

### Table 4. Sensitivity, specificity, PPV, and NPV of all terminal nodes for the race-specific final pruned CART models

<table>
<thead>
<tr>
<th>Race/ethnicity</th>
<th>CART path</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Americans</td>
<td>CYP3A43 genotype</td>
<td>10</td>
<td>99</td>
<td>94</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>CYP3A43 genotype* family history of prostate cancer</td>
<td>22</td>
<td>90</td>
<td>79</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>CYP3A43 genotype* family history of prostate cancer* age at consent</td>
<td>86</td>
<td>33</td>
<td>68</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>CYP3A43 genotype* family history of prostate cancer* age at consent* history of BPH</td>
<td>10</td>
<td>99</td>
<td>87</td>
<td>57</td>
</tr>
<tr>
<td>African Americans</td>
<td>Family history of prostate cancer</td>
<td>28</td>
<td>85</td>
<td>42</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Family history of prostate cancer* European ancestry* AR GGC repeats</td>
<td>27</td>
<td>90</td>
<td>64</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Family history of prostate cancer* European ancestry* CYP3A4/CYP3A5 haplotype</td>
<td>82</td>
<td>59</td>
<td>67</td>
<td>77</td>
</tr>
</tbody>
</table>

NOTE: All the values are expressed as percentages.
Discussion

Our results provide evidence that genetic factors influence prostate cancer risk in a context- and race-specific manner. We used a series of models to evaluate the interaction of genetic and nongenetic factors and prostate cancer risk. Although the most parsimonious models identified in each race differed, there were commonalities in the final models with all models identifying age and family history of prostate cancer as important predictors. The best-fitting models in both races also identified genetic information as providing added predictive ability beyond that of "traditional" risk factors, such as age, race or family history of prostate cancer (as reviewed in refs. 8, 9). The specific genotypes and/or haplotypes differed between races. This difference is not unexpected because genotype/haplotype frequencies differ substantially by race. In addition, it is reasonable to hypothesize that genetic effects may differ depending on the context in which they occur.

In this study, we observed that androgen metabolism genes play a role in prostate cancer etiology. Androgens play a central role in promoting the growth and differentiation of prostate tumors. Although there is no clear relationship between circulating testosterone levels and prostate cancer (11, 30–32), the factors associated with prostate cancer are likely associated with endogenous hormonal environment of an individual. We utilized a series of models to investigate the hierarchy of gene–environment and gene–gene interactions associated with prostate cancer risk stratified by race utilizing information on genes in the androgen pathway, including decision trees. For European Americans, interactions associated with risk exist for those with particular CYP3A43 genotypes, family history of prostate cancer, varying age at consent, and history of BPH (Fig. 2). For African Americans, interactions associated with risk exist for those with family history of prostate cancer, individual European ancestry proportion, number of GGC AR repeats, and CYP3A4/CYP3A5 haplotypes (Fig. 3). The CART decision tree–based model in European Americans yielded higher AUC than the standard multivariable backward selection LR models. This result shows the utility of the CART decision tree–based modeling approach to discover gene–gene and gene–environment interactions in datasets that may have limited sample sizes to detect such interactions by using standard LR models.

We identified associations of androgen metabolism genes and prostate cancer risk in both races. Our results are consistent with knowledge of gene and allele function in CYP3A4 and CYP3A43 (32–39). CYP3A4 and CYP3A5 previously have been associated with prostate cancer occurrence and severity (3, 34, 36). A number of reports have suggested that 1 or more of these genes may be associated with prostate cancer etiology or severity (3–6). However, these loci are in linkage disequilibrium with one another (3), and it remains unclear which, if any, of the variants at these loci may be causally associated with altered hormone metabolism or prostate cancer risk. Here we show that in standard haplotype analyses that the CYP3A4/CYP2A5 AG haplotype is significantly associated with increased risk of prostate cancer in European Americans but not in African Americans. Through CART modeling of prostate cancer risk we show that in European Americans, this haplotype is not associated with risk; instead, the CYP3A43 GC or CC genotypes are
associated with prostate cancer risk, in combination with family history of prostate cancer, age at consent, and/or history of BPH. However, in African Americans, the CYP3A4/CYP3A5 GA, AG, or GG haplotypes are associated with risk but only in combination with a higher proportion of European ancestry and positive family history of prostate cancer.

The AR, located on the X chromosome, plays a major role in the development and normal function of the prostate gland. Several regions of repetitive DNA sequences exist in this region, including CAG trinucleotide repeats encoding polyglutamine residues and GGN repeats encoding polyglycine residues. These repeat variants have been associated with androgen independence in syndromes associated with extremely long AR repeat sequences (such as Kennedy disease; refs. 40, 41). Several studies have shown an inverse association between the number of CAG and GGN repeats and risk of prostate cancer, advanced cancer, or risk of associated mortality (42–49), although some studies suggest that a positive association exists between prostate cancer and long GGN repeats in combination with short CAG repeats (42, 44, 50). We found through standard LR analysis that increasing number of CAG repeats was associated with a decreased risk of prostate cancer in European Americans. However, in the CART modeling of risk, AR repeats was not associated with risk in European Americans. In African Americans, having less than 16 GCC repeats was associated with increased risk but only in combination with lower European ancestry and positive family history of prostate cancer.

The CART decision tree–based method has both advantages over standard LR models for assessing associations with cancer risk and its own limitations. First, CART is an iterative, nonparametric procedure that identifies mutually exclusive risk subgroups that share common factors associated with risk of disease and is not constrained by distributional assumptions that may be violated in data applications. Therefore, information from CART models can be used to develop individualized interventions and/or treatments, whereas information from regression models applies to the average member of a population only. Second, the CART method discovers new relationships between variables and associations with risk that may not be identified by the traditional epidemiologic analysis techniques as is shown in Figures 2 and 3. In general, standard multiplicative or additive interaction modeling by using LR models requires relatively large sample sizes (51), where it seems from this study that CART has reasonable precision to find complex interactions by race given that the width of the CIs for the calculated ORs in Figures 2 and 3 are tight. However, CART models may not be able to discover particular important interactions because of limitations imposed by the stopping rules, the competitive importance of the variables and/or the pruning procedure. Third, CART allows the user to put all potential predictors into the model and prioritize variables by assigning an actual hierarchical structure to them associated with risk, whereas with traditional regression models typically the user needs to have some a priori knowledge about which interactions may be important and/or which main effects may be statistically significant before fitting higher order interactions. Fourth, CART allows the user to provide continuous variables of interest and the algorithm will generate optimal cutpoints for these variables as they relate to the best classification of cases and controls. The age, number of AR repeats, and European ancestry proportion cutpoints in Figures 2 and 3 were discovered through the CART analysis. However, other cutpoints may also be scientifically justified beyond the ones shown. Other limitations of the general approach used here include a relatively small sample size of African Americans, and the attendant problem of limited statistical power. Nonetheless, we have been able to identify relevant associations using the CART approach, which suggests that the model had adequate power to detect at least some associations.

Ultimately, the ability to stratify prostate cancer risk has several potential clinical applications. If modifiable risk factors are present (e.g., diet, exercise, and smoking), interventions can be directed accordingly. In the case of heritable risk, screening intensity might be appropriately guided by accurate risk assessment, as there is controversy about the routine screening for prostate cancer by using the prostate-specific antigen (PSA) blood test (52–54). Our findings suggest that germline polymorphisms in a panel of androgen metabolism pathway genes might have potential as a tool for selection of patients for PSA screening. Our findings about the genetic differences in the prostate cancer risk profile between European Americans and African Americans may provide a biological basis for tailored screening approaches in different populations (9). Future prospective studies and decision modeling will be required to advance the development of these clinical tools.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

The Editor-in-Chief of Cancer Epidemiology, Biomarkers & Prevention is an author of this article. In keeping with the AACR’s Editorial Policy, the paper was peer reviewed and a member of the AACR’s Publications Committee rendered the decision concerning acceptability.

Acknowledgment

The authors thank Van Ahn Tran for assistance with the ancestry estimation and Yingli Wolinsky for technical assistance with manuscript preparation.

Grant Support

This work was supported in part by Case Comprehensive Cancer Center Core Grants P30-CA043703 (J.S. Barnholtz-Sloan and N.J. Meropol), and NCI grants R01-CA08574 (T. Rebbeck), and P01-CA105641 (T. Rebbeck).

Received September 17, 2010; revised February 18, 2011; accepted April 1, 2011; published OnlineFirst April 14, 2011.
References


Androgen Genes and Prostate Cancer Risk


Decision Tree–Based Modeling of Androgen Pathway Genes and Prostate Cancer Risk

Jill S. Barnholtz-Sloan, Xiaowei Guan, Charnita Zeigler-Johnson, et al.

Cancer Epidemiol Biomarkers Prev 2011;20:1146-1155. Published OnlineFirst April 14, 2011.

Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-10-0996

Cited articles
This article cites 50 articles, 14 of which you can access for free at:
http://cebp.aacrjournals.org/content/20/6/1146.full#ref-list-1

Citing articles
This article has been cited by 1 HighWire-hosted articles. Access the articles at:
http://cebp.aacrjournals.org/content/20/6/1146.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.