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

Research Article

Cancer Epidemiology, Biomarkers & Prevention

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

Jill S. Barnholtz-Sloan[1,2], Xiaowei Guan[1,2], Charnita Zeigler-Johnson[4], Neal J. Meropol[1,3], and Timothy R. Rebbeck[4]

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. *Cancer Epidemiol Biomarkers Prev; 20(6); 1146–55. ©2011 AACR.*

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,

Authors' Affiliations: [1]Case Comprehensive Cancer Center; [2]Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine; [3]Division of Hematology and Oncology, University Hospitals Case Medical Center and Case Western Reserve University, Cleveland, Ohio; and [4]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

Corresponding Author: Jill S. Barnholtz-Sloan, Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, 11100 Euclid Ave–Wearn 152, Cleveland, OH 44106-5065. Phone: 1-216-368-1506; Fax: 1-216-368-2606. E-mail: jsb42@case.edu

doi: 10.1158/1055-9965.EPI-10-0996

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

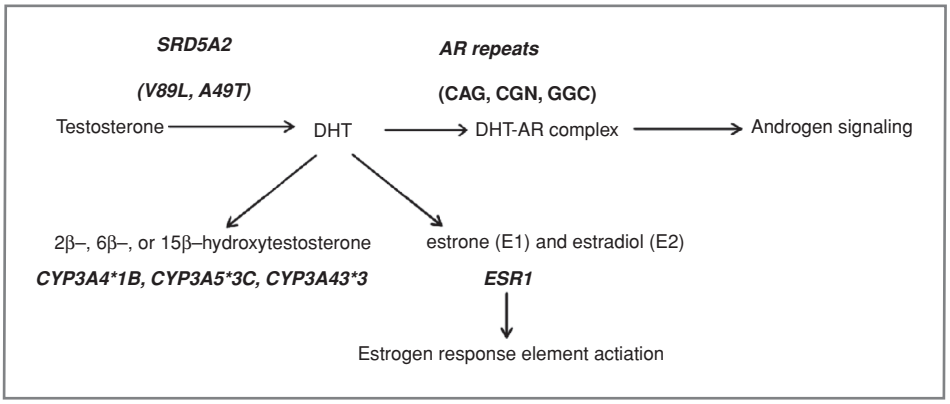

Figure 1. Overview of androgen metabolism pathway genes considered, including candidate genes and genetic variants studied.

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 multivariate 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

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

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/CYP3A45* 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* (*A49T*; rs9282858; Lewontin's D′ = 1.00). *CYP3A43**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 (ORs) 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 ROCR package in R. AUC values from the final LR models were compared with the baseline LR model in a pairwise fashion by race by using the χ^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 = 0.90; 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

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

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

	European American (*n*, column %)		African American (*n*, column %)	
	Cases	Controls	Cases	Controls
Family history of prostate cancer				
Yes	207 (22.23)	51 (9.46)	43 (28.1)	59 (14.68)
No	684 (73.47)	482 (89.42)	90 (58.82)	319 (79.35)
Missing	40 (4.3)	6 (1.11)	20 (13.07)	24 (5.97)
Personal history of benign prostate hypertrophy (history of BPH)				
Yes	189 (20.3)	52 (9.65)	29 (18.95)	39 (9.7)
No	699 (75.08)	476 (88.31)	96 (62.75)	333 (82.84)
Missing	43 (4.62)	11 (2.04)	28 (18.3)	30 (7.46)
European ancestry proportion (%)				
<25	519 (55.75)	363 (67.35)	118 (77.12)	356 (88.56)
25–75	1 (0.11)	0 (0)	29 (18.95)	38 (9.45)
>75	411 (44.15)	176 (32.65)	6 (3.92)	8 (1.99)
Missing	0 (0)	0 (0)	0 (0)	0 (0)

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.

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 *AR* 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

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)

		European American (931 cases, 539 controls)		African American (153 cases, 402 controls)			
Gene	rsNumber	Alelles (test/ referent)	Referent allele frequency (%)	Additive (per allele) OR (95% CI)	Alelles (test/ referent)	Referent allele frequency (%)	Additive (per allele) OR (95% CI)
CYP3A5	rs10249369	A/G	92.72	**2.11 (1.13–3.97)**	G/A	61.33	0.59 (0.28–1.22)
CYP3A4	rs2740574	G/A	95.19	1.17 (0.54–2.54)	A/G	59.90	1.20 (0.56–2.56)
CYP3A43	rs680055	G/C	86.04	1.31 (0.78–2.19)	G/C	62.66	0.75 (0.35–1.57)
SRD5A2	rs523349	C/G	70.54	1.45 (0.98–2.14)	C/G	72.54	0.85 (0.41–1.77)
SRD5A2	rs9282858	A/G	97.36	1.57 (0.56–4.31)	A/G	99.04	1.71 (0.08–34.82)
ESR1	rs3853250	C/A	54.17	1.37 (0.91–2.05)	A/C	54.22	1.08 (0.57–2.04)
AR	CAG (rs4045402)			**0.90 (0.83–0.98)**			0.91 (0.80–1.04)
AR	GGN (rs3138869)			0.59 (0.31–1.13)			Not estimable
AR	GGC (rs3138869)			0.85 (0.71–1.02)			1.07 (0.89–1.29)

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

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]

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

[a]Haplotypes 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).

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 ($CART_{BASE}$; 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

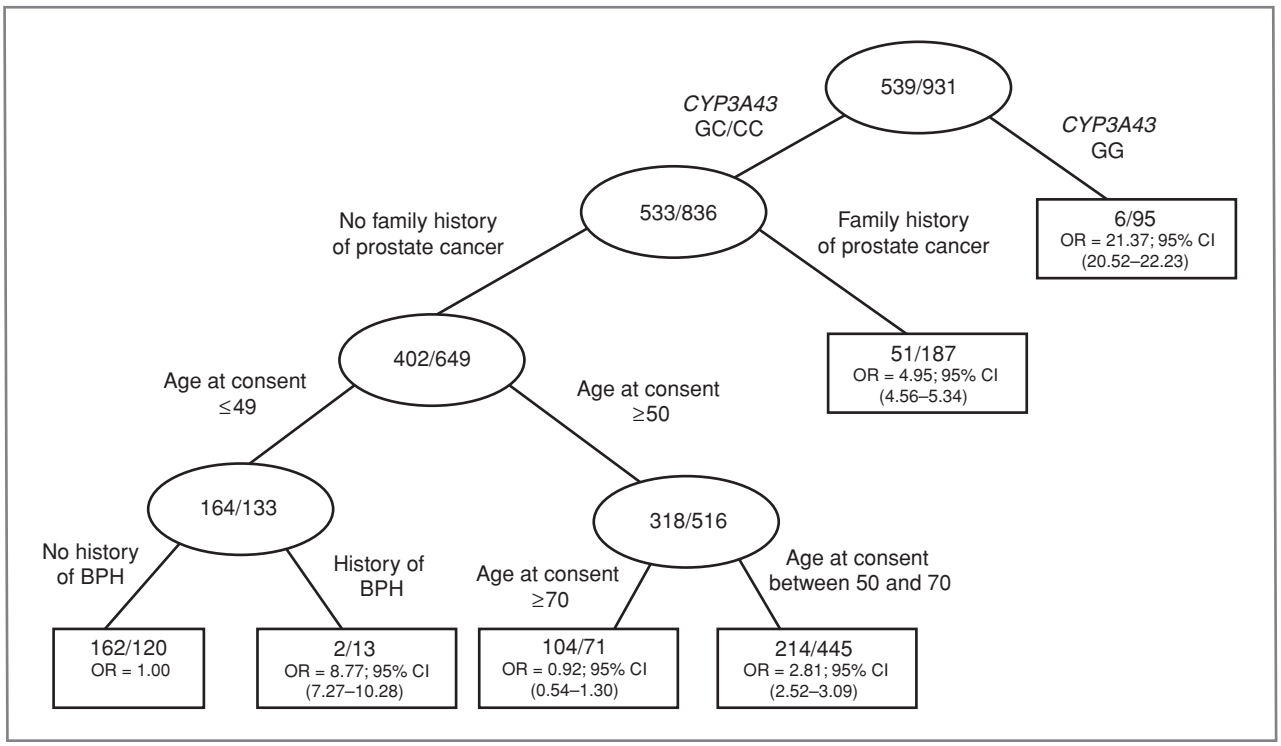

Figure 2. Pruned classification tree for androgen pathway in European Americans. The numbers within each node indicate the number of controls/the number of prostate cancer cases.

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

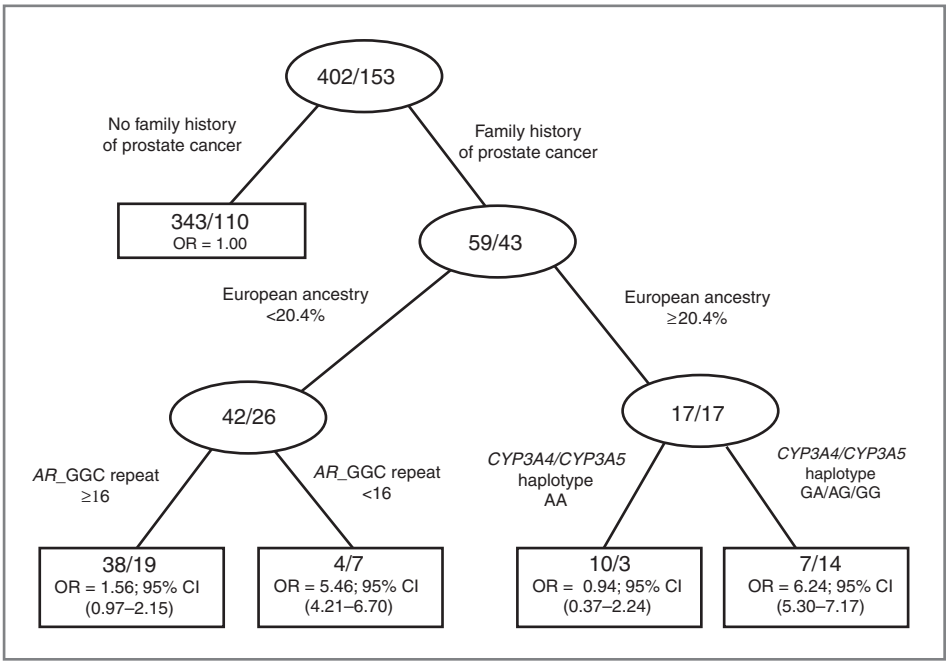

Figure 3. Pruned classification tree for androgen pathway in African Americans. The numbers within each node indicate the number of controls/the number of prostate cancer cases.

history of cancer, history of BPH, and European ancestry proportion for African Americans (LR_{BASE}); (iv), a LR model including all main effects and interaction terms as found by CART (LR_{CART}); and (v), a fully saturated backward selection LR model with all interactions (LR_{FULL}; 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 LR_{FULL} and the LR_{CART} models performed as well as the LR_{BASE} model (χ^2 P value for LR_{FULL} vs. $LR_{BASE} = 0.11$; LR_{CART} vs. $LR_{BASE} = 0.09$). For African Americans, the LR_{CART} 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 LR_{FULL} and the LR_{CART} models performed better than the LR_{BASE} model (χ^2 P value for LR_{FULL} vs. $LR_{BASE} = 0.85$; LR_{CART} vs. $LR_{BASE} = 0.51$). 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

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

NOTE: All the values are expressed as percentages.

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

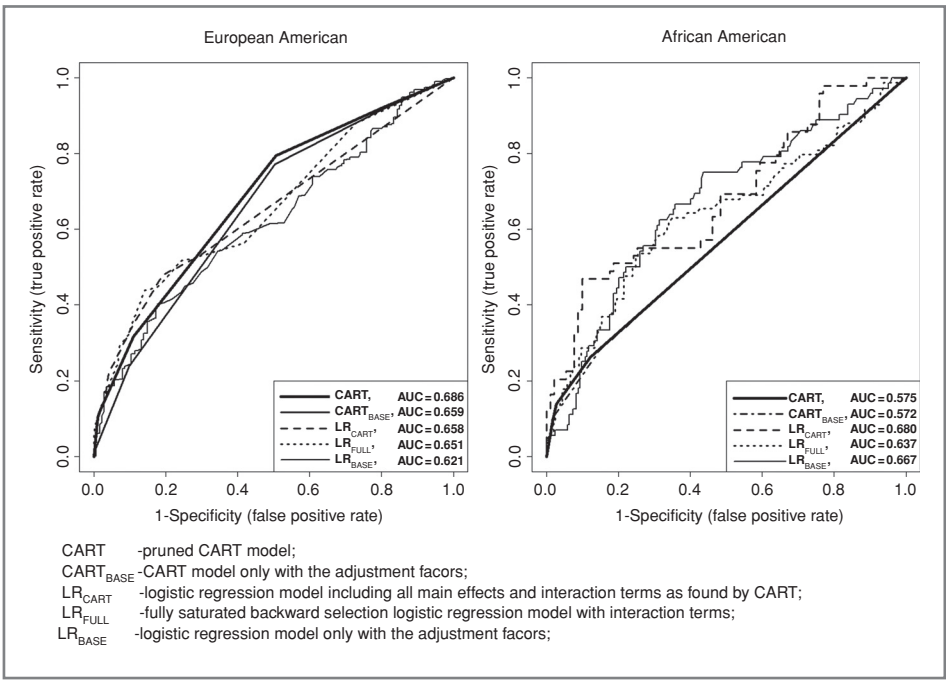

Figure 4. ROC analysis for CART versus LR stratified by race.

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 genotypes 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

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

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.

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

References

1. Edwards BK, Ward E, Kohler BA, Eheman C, Zauber AG, Anderson RN, et al. Annual report to the nation on the status of cancer, 1975–2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer 2010;116:544–73.
2. American Cancer Society. Cancer facts and figures 2010. Atlanta, GA: American Cancer Society, Inc; 2010.
3. Zeigler-Johnson C, Friebel T, Walker AH, Wang Y, Spangler E, Panossian S, et al. CYP3A4, CYP3A5, and CYP3A43 genotypes and haplotypes in the etiology and severity of prostate cancer. Cancer Res 2004;64:8461–7.
4. Rebbeck TR, Jaffe JM, Walker AH, Wein AJ, Malkowicz SB. Modification of clinical presentation of prostate tumors by a novel genetic variant in CYP3A4. J Natl Cancer Inst 1998;90:1225–9.
5. Plummer SJ, Conti DV, Paris PL, Curran AP, Casey G, Witte JS. CYP3A4 and CYP3A5 genotypes, haplotypes, and risk of prostate cancer. Cancer Epidemiol Biomarkers Prev 2003;12:928–32.
6. Paris PL, Kupelian PA, Hall JM, Williams TL, Levin H, Klein EA, et al. Association between a CYP3A4 genetic variant and clinical presentation in African-American prostate cancer patients. Cancer Epidemiol Biomarkers Prev 1999;8:901–5.
7. McClellan J, King MC. Genetic heterogeneity in human disease. Cell 2010;141:210–7.
8. Nelen V. Epidemiology of prostate cancer. Recent Results Cancer Res 2007;175:1–8.
9. Williams H, Powell IJ. Epidemiology, pathology, and genetics of prostate cancer among African Americans compared with other ethnicities. Methods Mol Biol 2009;472:439–53.
10. Vatten LJ, Ursin G, Ross RK, Stanczyk FZ, Lobo RA, Harvei S, et al. Androgens in serum and the risk of prostate cancer: a nested case-control study from the Janus serum bank in Norway. Cancer Epidemiol Biomarkers Prev 1997;6:967–9.
11. Gann PH, Hennekens CH, Ma J, Longcope C, Stampfer MJ. Prospective study of sex hormone levels and risk of prostate cancer. J Natl Cancer Inst 1996;88:1118–26.
12. Luke MC, Coffey DS. The male sex assessory tissues. Structure, androgen action and physiology. In: Knobil E, Neill JD, editors. The physiology of reproduction. 2nd ed. New York: Raven Press Ltd.; 1994. p. 1435–87.
13. Tanaka Y, Sasaki M, Kaneuchi M, Shiina H, Igawa M, Dahiya R. Polymorphisms of estrogen receptor alpha in prostate cancer. Mol Carcinog 2003;37:202–8.
14. McIntyre MH, Kantoff PW, Stampfer MJ, Mucci LA, Parslow D, Li H, et al. Prostate cancer risk and ESR1 TA, ESR2 CA repeat polymorphisms. Cancer Epidemiol Biomarkers Prev 2007;16:2233–6.
15. Chae YK, Huang HY, Strickland P, Hoffman SC, Helzlsouer K. Genetic polymorphisms of estrogen receptors alpha and beta and the risk of developing prostate cancer. PLoS One 2009;4:e6523.
16. Powell IJ. Prostate cancer in the African American: is this a different disease? Semin Urol Oncol 1998;16:221–6.
17. Powell IJ. Epidemiology and pathophysiology of prostate cancer in African-American men. J Urol 2007;177:444–9.
18. Richards B, Skoletsky J, Shuber AP, Balfour R, Stern RC, Dorkin HL, et al. Multiplex PCR amplification from the CFTR gene using DNA prepared from buccal brushes/swabs. Hum Mol Genet 1993;2:159–63.
19. Walker AH, Najarian D, White DL, Jaffe JF, Kanetsky PA, Rebbeck TR. Collection of genomic DNA by buccal swabs for polymerase chain reaction-based biomarker assays. Environ Health Perspect 1999;107:517–20.
20. Zeigler-Johnson CM, Walker AH, Mancke B, Spangler E, Jalloh M, McBride S, et al. Ethnic differences in the frequency of prostate cancer susceptibility alleles at SRD5A2 and CYP3A4. Hum Hered 2002;54:13–21.
21. Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF. A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet 2006;79:640–9.
22. Barnholtz-Sloan JS, McEvoy B, Shriver MD, Rebbeck TR. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. Cancer Epidemiol Biomarkers Prev 2008;17:471–7.
23. Barnholtz-Sloan JS, Chakraborty R, Sellers TA, Schwartz AG. Examining population stratification via individual ancestry estimates versus self-reported race. Cancer Epidemiol Biomarkers Prev 2005;14:1545–51.
24. Barnholtz-Sloan JS, Shetty PB, Guan X, Nyante SJ, Luo J, Brennan DJ, et al. FGFR2 and other loci identified in genome-wide association studies are associated with breast cancer in African-American and younger women. Carcinogenesis 2010;31:1417–23.
25. SAS. Statistical analysis software, version 9.2. North Carolina:Cary; 2007.
26. Giovannucci E, Stampfer MJ, Krithivas K, Brown M, Dahl D, Brufsky A, et al. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. Proc Natl Acad Sci U S A 1997;94:3320–3.
27. Rebbeck TR, Kantoff PW, Krithivas K, Neuhausen S, Blackwood MA, Godwin AK, et al. Modification of BRCA1-associated breast cancer risk by the polymorphic androgen-receptor CAG repeat. Am J Hum Genet 1999;64:1371–7.
28. Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol 2004;27:348–64.
29. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton, FL: Chapman & Hall/CRC Press; 1984.
30. Pienta KJ, Goodson JA, Esper PS. Epidemiology of prostate cancer: molecular and environmental clues. Urology 1996;48:676–83.
31. Meikle AW, Smith JA Jr. Epidemiology of prostate cancer. Urol Clin North Am 1990;17:709–18.
32. Hashimoto H, Toide K, Kitamura R, Fujita M, Tagawa S, Itoh S, et al. Gene structure of CYP3A4, an adult-specific form of cytochrome P450 in human livers, and its transcriptional control. Eur J Biochem 1993;218:585–95.
33. Westlind A, Malmebo S, Johansson I, Otter C, Andersson TB, Ingelman-Sundberg M, et al. Cloning and tissue distribution of a novel human cytochrome p450 of the CYP3A subfamily, CYP3A43. Biochem Biophys Res Commun 2001;281:1349–55.
34. Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, et al. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. Nat Genet 2001;27:383–91.
35. Koch I, Weil R, Wolbold R, Brockmoller J, Hustert E, Burk O, et al. Interindividual variability and tissue-specificity in the expression of cytochrome P450 3A mRNA. Drug Metab Dispos 2002;30:1108–14.
36. Gellner K, Eiselt R, Hustert E, Arnold H, Koch I, Haberl M, et al. Genomic organization of the human CYP3A locus: identification of a new, inducible CYP3A gene. Pharmacogenetics 2001;11:111–21.
37. Domanski TL, Finta C, Halpert JR, Zaphiropoulos PG. cDNA cloning and initial characterization of CYP3A43, a novel human cytochrome P450. Mol Pharmacol 2001;59:386–92.
38. Soderstrom T, Wadelius M, Andersson SO, Johansson JE, Johansson S, Granath F, et al. 5 alpha-reductase 2 polymorphisms as risk factors in prostate cancer. Pharmacogenetics 2002;12:307–12.
39. Shibata A, Garcia MI, Cheng I, Stamey TA, McNeal JE, Brooks JD, et al. Polymorphisms in the androgen receptor and type II 5 alpha-reductase genes and prostate cancer prognosis. Prostate 2002;52:269–78.
40. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 1991;352:77–9.
41. Doyu M, Sobue G, Mukai E, Kachi T, Yasuda T, Mitsuma T, et al. Severity of X-linked recessive bulbospinal neuronopathy correlates with size of the tandem CAG repeat in androgen receptor gene. Ann Neurol 1992;32:707–10.
42. Zeegers MP, Kiemeney LA, Nieder AM, Ostrer H. How strong is the association between CAG and GGN repeat length polymorphisms in the androgen receptor gene and prostate cancer risk? Cancer Epidemiol Biomarkers Prev 2004;13:1765–71.

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

43. Ingles SA, Ross RK, Yu MC, Irvine RA, La Pera G, Haile RW, et al. Association of prostate cancer risk with genetic polymorphisms in vitamin D receptor and androgen receptor. J Natl Cancer Inst 1997;89:166–70.
44. Hsing AW, Gao YT, Wu G, Wang X, Deng J, Chen YL, et al. Polymorphic CAG and GGN repeat lengths in the androgen receptor gene and prostate cancer risk: a population-based case-control study in China. Cancer Res 2000;60:5111–6.
45. Hakimi JM, Schoenberg MP, Rondinelli RH, Piantadosi S, Barrack ER. Androgen receptor variants with short glutamine or glycine repeats may identify unique subpopulations of men with prostate cancer. Clin Cancer Res 1997;3:1599–608.
46. Giwercman YL, Abrahamsson PA, Giwercman A, Gadaleanu V, Ahlgren G. The 5 alpha-reductase type II A49T and V89L high-activity allelic variants are more common in men with prostate cancer compared with the general population. Eur Urol 2005;48: 679–85.
47. Giovannucci E, Platz EA, Stampfer MJ, Chan A, Krithivas K, Kawachi I, et al. The CAG repeat within the androgen receptor gene and benign prostatic hyperplasia. Urology 1999;53:121–5.
48. Cude KJ, Montgomery JS, Price DK, Dixon SC, Kincaid RL, Kovacs KF, et al. The role of an androgen receptor polymorphism in the clinical outcome of patients with metastatic prostate cancer. Urol Int 2002;68:16–23.
49. Binnie MC, Alexander FE, Heald C, Habib FK. Polymorphic forms of prostate specific antigen and their interaction with androgen receptor trinucleotide repeats in prostate cancer. Prostate 2005;63:309–15.
50. Vijayalakshmi K, Thangaraj K, Rajender S, Vettriselvi V, Venkatesan P, Shroff S, et al. GGN repeat length and GGN/CAG haplotype variations in the androgen receptor gene and prostate cancer risk in south Indian men. J Hum Genet 2006;51:998–1005.
51. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 2002;155:478–84.
52. Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008;149:185–91.
53. Barry MJ. Screening for prostate cancer–the controversy that refuses to die. N Engl J Med 2009;360:1351–4.
54. Barry MJ. Review: evidence from 2 low quality screening studies does not show a reduction in death from prostate cancer. Evid Based Med 2007;12:40.

Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

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.