nature genetics

Multiple loci identified in a genome-wide association study of prostate cancer

Gilles Thomas[1], Kevin B Jacobs[2], Meredith Yeager[1,3], Peter Kraft[4], Sholom Wacholder[1], Nick Orr[1], Kai Yu[1], Nilanjan Chatterjee[1], Robert Welch[1,3], Amy Hutchinson[1,3], Andrew Crenshaw[1,3], Geraldine Cancel-Tassin[5], Brian J Staats[1,3], Zhaoming Wang[1,3], Jesus Gonzalez-Bosquet[1], Jun Fang[1], Xiang Deng[1,3], Sonja I Berndt[1], Eugenia E Calle[6], Heather Spencer Feigelson[6], Michael J Thun[6], Carmen Rodriguez[6], Demetrius Albanes[1], Jarmo Virtamo[7], Stephanie Weinstein[1], Fredrick R Schumacher[4], Edward Giovannucci[8], Walter C Willett[8], Olivier Cussenot[5], Antoine Valeri[5], Gerald L Andriole[9], E David Crawford[10], Margaret Tucker[1], Daniela S Gerhard[11], Joseph F Fraumeni Jr[1], Robert Hoover[1], Richard B Hayes[1], David J Hunter[1,4] & Stephen J Chanock[1,12]

We followed our initial genome-wide association study (GWAS) of 527,869 SNPs on 1,172 individuals with prostate cancer and 1,157 controls of European origin—nested in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial prospective study—by testing 26,958 SNPs in four independent studies (total of 3,941 cases and 3,964 controls). In the combined joint analysis, we confirmed three previously reported loci (two independent SNPs at 8q24 and one in *HNF1B* (formerly known as *TCF2* on 17q); $P < 10^{-10}$). In addition, loci on chromosomes 7, 10 (two loci) and 11 were highly significant (between $P < 7.31 \times 10^{-13}$ and $P < 2.14 \times 10^{-6}$). Loci on chromosome 10 include *MSMB,* which encodes β-microseminoprotein, a primary constituent of semen and a proposed prostate cancer biomarker, and *CTBP2,* a gene with antiapoptotic activity; the locus on chromosome 7 is at *JAZF1,* a transcriptional repressor that is fused by chromosome translocation to *SUZ12* in endometrial cancer. Of the nine loci that showed highly suggestive associations ($P < 2.5 \times 10^{-5}$), four best fit a recessive model and included candidate susceptibility genes: *CPNE3, IL16* and *CDH13.* Our findings point to multiple loci with moderate effects associated with susceptibility to prostate cancer that, taken together, in the future may predict high risk in select individuals.

In developed countries, prostate cancer is the most common non-cutaneous malignancy in men[1]. The only established risk factors are age, African ancestry, and a positive family history of disease[1]. Twin studies and epidemiologic observations have suggested a substantial genetic contribution to disease risk[2]. Recently, linkage, admixture mapping and genome-wide studies have identified variants with moderate effects on prostate cancer risk at multiple loci in the 8q24 region[3–7], two of which have been replicated by independent groups, and in *HNF1B* (formerly *TCF2*)[8]. These loci account for a fraction of the elevated risk for relatives of individuals with prostate cancer, suggesting that additional loci exist[9]. Recognizing the limitations of both candidate gene and linkage studies in identifying common susceptibility genes in prostate cancer, we conducted a two-stage GWAS in order to search for common variants with moderate risk. For the first stage of our cost-effective approach[10], we used 527,869 SNPs, monitoring 91% of common autosomal SNPs in HapMapII typed from data on the population of European ancestry (CEU; $r^2 = 0.8$, minor allele frequency (MAF) $>5\%$)[7,11]. For the second step, we analyzed the most promising 27,326 SNPs from the first scan in 3,941 cases and 3,964 controls. The large number of SNPs genotyped in the second stage enabled us to follow up on regions with moderate association in the initial genome-wide scan ($P < 0.068$).

We conducted the initial genome-wide scan in a nested case-control study of 1,172 prostate-specific antigen (PSA) screened cases

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20892, USA. [2]Bioinformed Consulting Services, Gaithersburg, Maryland 20877, USA. [3]Core Genotyping Facility, Advanced Technology Program, SAIC Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, USA. [4]Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [5]CeRePP Hopital Tenon, Assistance Publique-Hôpitaux de Paris, Paris 75970, France. [6]Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Georgia 30329, USA. [7]Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Helsinki FIN-0030, Finland. [8]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [9]Division of Urologic Surgery, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [10]Department of Surgery, University of Colorado at Denver and Health Sciences Center, Denver, Colorado 80014, USA. [11]Office of Cancer Genomics and [12]Pediatric Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20892, USA. Correspondence should be addressed to S.C. (chanocks@mail.nih.gov).

Received 9 November 2007; accepted 26 December 2007; published online 10 February 2008; doi:10.1038/ng.91

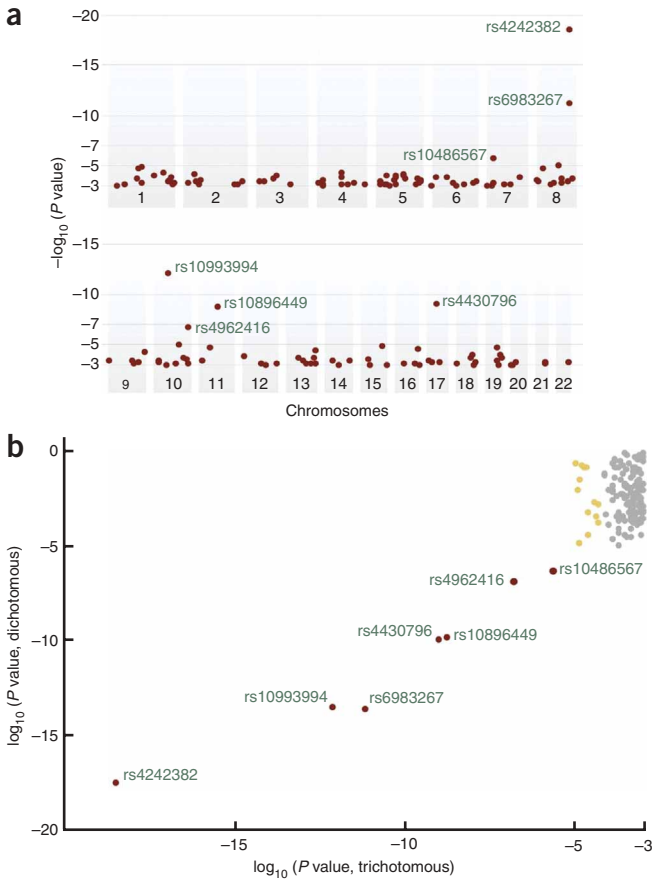

Figure 1 Association analysis of combined joint analysis in two-stage GWAS of prostate cancer. The 194 SNPs with $P < 10^{-3}$ in the trichotomous regression analysis (distinguishing nonaggressive and aggressive prostate cancer) are located in 150 distinct regions of the genome. In both panels only the SNP with the lowest such *P* value is shown. (**a**) Distribution along the chromosomes of the trichotomous *P* values for the 150 regions. (**b**) Relationship between the trichotomous *P* value and the dichotomous *P* value obtained by combining the nonaggressive and aggressive cancer cases into a single case group. See **Supplementary Table 1** for location of and LD between SNPs with $P < 10^{-3}$ in the trichotomous regression analysis.

(oversampled for aggressive cases; 484 nonaggressive prostate cancer (Gleason score <7 and disease stage <III) and 688 aggressive prostate cancer (Gleason score ≥7 and/or disease stage ≥III)) and 1,157 PSA-screened controls in men of European ancestry from the PLCO Cancer Screening Trial[12,13]. In the second stage, we successfully genotyped 26,958 SNPs in four additional replication studies totaling 4,020 cases and 4,028 controls (American Cancer Society Cancer Prevention Study II[14], 1,790/1,797; the Health Professionals Follow-up Study[15], 619/620; CeRePP French Prostate Case-Control Study[16], 671/671; and Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study[17], 940/940) (**Supplementary Methods** online). Of the 26,613 SNPs chosen on the basis of the initial genome-wide scan for attempted replication (see below), 24,748 SNPs were of sufficient quality and frequency for analysis in four follow-up studies in men of European background. Initially, we had selected 25,265 SNPs in a second stage from the whole-genome scan on the basis of a *P* value reflecting an age, center, and population stratification–adjusted polytomous/multinomial regression analysis (4 degree-of-freedom (d.f) test) that accounted for both heterozygous and homozygous variant genotypes and the two groups of cases (nonaggressive and aggressive) in PLCO. We excluded SNPs with higher *P* values that were in strong linkage disequilibrium ($r^2 > 0.8$) with a previously selected SNP. The threshold for single SNP selection was $P < 0.068$. An additional 1,348 SNPs had been selected on the basis of a 2-SNP test that improved the *P* value relative to the single-SNP statistic by at least an order of magnitude. Furthermore, 1,508 SNPs had been chosen to monitor population stratification. These SNPs detected comparable first principal components as a set of 10,693 SNPs drawn from the second stage (described above) which had been selected because they had little evidence for linkage disequilibrium over 200 kb ($r^2 < 0.0075$) (**Supplementary Methods**)[18].

The 26,958 SNPs reliably genotyped in the second stage covered 7,608 distinct chromosomal regions defined by a maximal distance between two SNPs of less than 100 kb. We found that 2,853 regions contained only one SNP, and 501 regions contained 10 or more SNPs. (**Supplementary Methods** and **Supplementary Table 1** online). Of these regions, we followed up on 150 that had at least one SNP with an observed *P* value $<10^{-3}$ (**Fig. 1**). SNPs with a *P* value $>10^{-3}$ in the combined analysis of our two stages are unlikely to reach convincing genome-wide significance in a combined analysis including an additional large third stage—effectively doubling the overall number of cases and controls of the study—as is planned in the next stage of Cancer Genetic Markers of Susceptibility (CGEMS) follow-up (see **Supplementary Methods**). The largest region on 8q24 (between position 127.4 Mb and 128.9 Mb) contained 28 SNPs with *P* values $<10^{-3}$, but this region was deliberately explored with 192 SNPs drawn from the initial genome scan on the basis of strong prior evidence for associations with breast, colon and prostate cancer[7,19,20]. An additional 21 regions contained two or more significant SNPs. A remaining 115 significant SNPs were each located in a different region, a possible consequence of the filtering procedure based on linkage disequilibrium (see **Supplementary Methods**). Of note, almost one-fifth of the candidate regions identified were located on chromosomes 5 and 10, which may harbor multiple susceptibility loci for prostate cancer.

Our two-stage GWAS in prostate cancer has identified two sets of SNPs: one comprises established and newly identified loci associated with prostate cancer risk, and the second includes SNPs with $P < 10^{-3}$ that merit additional follow-up studies[18]. Most of the SNPs that achieved or approached genome-wide significance in the follow-up studies were not ranked in the top 1,000 SNPs in the initial genome-wide scan (**Table 1**). Indeed, the four most significant newly identified SNPs ranked 319 (*CTBP2*), 2,439 (chromosome 11), 24,223 (*MSMB*) and 24,407 (*JAZF1*), respectively, in the initial scan. Thus, these results support the value of genotyping a large number of SNPs in follow-up studies. Genotype counts for each SNP in each of the three phenotype groups (controls, nonaggressive cases and aggressive cases) are freely available as computed association statistics. Individual genotype data in the initial GWAS study are available for registered users through the CGEMS portal, thus providing investigators with an opportunity for additional analyses.

The results of our combined joint analysis by polytomous/multinomial logistic regression of the initial genome-wide scan and the four follow-up studies provide the necessary independent replication of an association with a common genetic variation in *HNF1B* (**Table 1** and **Supplementary Tables 2–7** online)[8]. We confirmed two independent loci in 8q24 previously associated with prostate cancer in a population

Table 1 Results from the pooled trichotomous association analysis of 2,109 nonaggressive prostate cancer cases, 2,651 aggressive prostate cancer cases and 5,133 controls

							Nonaggressive[a] vs. control		Aggressive[b] vs. control		Initial GWAS[c]	
dbSNP ID[d]	Risk allele[e] (frequency)	Region[f]	Chr.	LOC[g]	χ^{2h}	*P* value	Het. OR (95% CI)	Hom. OR (95% CI)	Het. OR (95% CI)	Hom. OR (95% CI)	Rank	*P* value
Previously reported												
rs4242382	A (0.12)	8q24	8	128586755	92.98	3.07×10^{-19}	1.41 (1.24–1.60)	1.86 (1.27–2.72)	1.66 (1.47–1.87)	2.22 (1.51–3.26)	116	1.12×10^{-4}
rs6983267	G (0.53)	8q24	8	128482487	58.31	6.58×10^{-12}	0.79 (0.70–0.90)	0.64 (0.55–0.74)	0.78 (0.69–0.87)	0.66 (0.57–0.75)	300	3.92×10^{-4}
rs4430796	A (0.54)	*HNF1B*	17	33172153	47.97	9.58×10^{-10}	0.72 (0.64–0.82)	0.66 (0.57–0.77)	0.85 (0.76–0.96)	0.72 (0.62–0.83)	384	5.21×10^{-4}
Newly reported												
rs10993994	T (0.40)	*MSMB*	10	51219502	62.85	7.31×10^{-13}	1.24 (1.10–1.39)	1.66 (1.42–1.95)	1.16 (1.04–1.29)	1.57 (1.36–1.81)	24,223	0.042
rs10896449	G (0.52)	11q13	11	68751243	46.70	1.76×10^{-09}	0.78 (0.69–0.88)	0.65 (0.56–0.76)	0.91 (0.81–1.02)	0.71 (0.62–0.82)	2,439	0.004
rs4962416	C (0.27)	*CTBP2*	10	126686862	37.12	1.70×10^{-07}	1.20 (1.07–1.34)	1.63 (1.33–1.99)	1.17 (1.05–1.30)	1.46 (1.22–1.76)	319	4.09×10^{-4}
rs10486567	G (0.77)	*JAZF1*	7	27749803	31.76	2.14×10^{-06}	0.74 (0.66–0.83)	0.71 (0.55–0.90)	0.89 (0.80–0.98)	0.84 (0.67–1.05)	24,407	0.042

The result of the trichotomous (nonaggressive and aggressive disease distinguished) logistic regression of the combined genotypes generated in the initial PLCO study and the four follow-up studies—adjusted for age in ten-year intervals, study/study center and four eigenvectors to control population stratification—is shown for the SNP providing the strongest signal in each region[10]. Included are the individual ranks and *P* values observed for the same SNP in the primary genome-wide scan in PLCO using a trichotomous regression analysis adjusted for the age in five years, study/center and three eigenvectors to control population stratification in an incident density sampling strategy. For five of the regions, additional SNPs were typed and reached *P* value $< 2.14 \times 10^{-6}$: for rs4242382 (rs4242384, rs1447295, rs7837688, rs11988857, rs7017300 and rs9656816), for rs6983267 (rs10505477, rs7837328 and rs7014346), for rs4430796 (rs7501939), for rs10993994 (rs11006207) and for rs4962416 (rs11245446, rs7077275 and rs4962708); see **Supplementary Tables 1** and **2**. Het., heterozygote; Hom., minor allele homozygote.

[a]Prostate cancer cases with a Gleason score <7 and disease stage <III. [b]Prostate cancer cases with a Gleason score ≥7 or disease stage ≥III. [c]CGEMS genome-wide association scan. [d]SNP identifier based on NCBI dbSNP. [e]SNP allele that confers susceptibility to prostate cancer. [f]Relative to SNP position. SNPs are included in the region of a gene if they are located within 20 kb of its transcription start site or within 10 kb from its last exon. SNPs at 8q24 are also indicated. [g]Chromosomal location based on NCBI Human Genome Build 35 coordinates. [h]4-d.f.genotype score test.

of European origin but did not find associations for three SNPs in this region related to prostate cancer risk in men of other ancestral origin[6]; namely, rs7000448, rs6983561 and rs13254738. The *P* value of a second SNP, rs4242382, in strong linkage disequilibrum (LD) with the previously reported rs1447295 was nearly an order of magnitude smaller than that of rs1447295 (refs. 3–7). Several newly identified loci met or approached the 'standard of genome-wide significance', $P < 10^{-7}$. These include four loci on chromosomes 7, 10 (two loci) and 11 that are highly significant ($P < 2.14 \times 10^{-6}$); moreover, the three loci on chromosomes 7 and 10 include candidate susceptibility genes, *CTBP2, MSMB* and *JAZF1*. Furthermore, rs10896449 lies on the long arm of chromosome 11 and is located 67 kb upstream of a gene overexpressed in myeloma, *MYEOV*[21]. Of note, we observed lower *P* values when we analyzed aggressive and nonaggressive tumors together with a 1-d.f. trend test: rs10486567 in *JAZF1* showed association with $P = 1.2 \times 10^{-7}$, and rs4962416 in *CTBP2* showed association with $P = 2.7 \times 10^{-8}$.

rs10993994, the SNP with the lowest *P* value among those in the newly identified loci, resides in the proximal promoter of *MSMB* (**Fig. 2**) and functionally alters *in vitro* gene expression[22]. *MSMB* encodes MSP, a 10.7-kDa nonglycosylated cysteine-rich protein that is a member of the immunoglobulin binding factor family synthesized by epithelial cells in the prostate and secreted into seminal plasma. MSP and its binding protein in serum, PSPBP, are potential serum markers for early detection of high-grade prostate cancer[23,24]. Although its expression has been noted in normal and neoplastic prostate tissue, *MSMB* can be silenced by EZH2 in advanced, androgen-insensitive prostate cancer[25]. Thus, variants in *MSMB* may predispose to prostate cancer through altered gene expression.

To determine whether additional, putative functional variants in *MSMB* are in strong LD with rs10993994, we conducted bidirectional sequence analysis across the gene region in 90 HapMap CEU individuals and in 40 men with aggressive phenotype and 40 controls from the PLCO prostate cancer study (**Supplementary Fig. 1** online). We identified a nonsynonymous variation in the initiation codon (ATG to ATA, altering Met to Ile) in a single CEU individual. We did not observe any other potentially functional variants. The probability that a SNP (with a MAF ~40%) in perfect linkage disequilibrium with rs10993994 might have escaped detection in the resequenced regions is remote ($P < 10^{-60}$).

Our two-stage GWAS identified six SNPs in a region on chromosome 10 that harbors two candidate genes, *CTBP2* and *ZRANB1* (**Supplementary Fig. 2** online). The strongest signal was observed for rs4962416 in the fifth intron of *CTBP2* ($P = 1.70 \times 10^{-7}$), which encodes a member of the C-terminal binding protein (CTBP) family

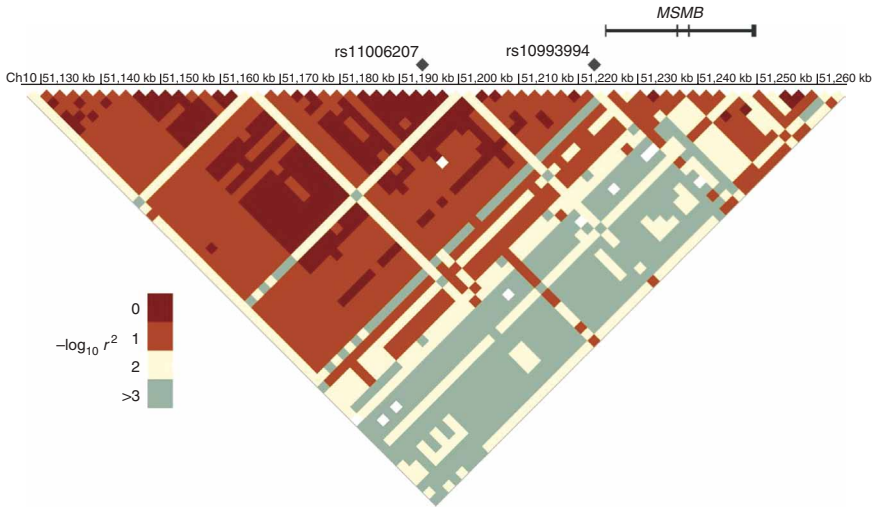

Figure 2 Location of the association signal and LD across the *MSMB* region. The pattern of LD across the *MSMB* gene on chromosome 10 is depicted using data from the CEU population of HapMapII (MAF > 5%; $r^2 > 0.8$). The color code indicates the estimated LD on the $-\log_{10}(r^2)$ scale. rs10993994 and rs11006207, the two SNPs in *MSMB* with significant *P* values, are indicated above. LD between the SNPs, $r^2 = 0.71$.

known to be transcriptional corepressors activated under metabolic stress. *CTBP2* is highly expressed in prostate tissue, and its expression has been associated with decreased *PTEN* expression and activation of the phosphatidylinositol 3-kinase pathway[26].

We confirmed rs10486567 in the second intron of the JAZF zinc finger 1 gene, *JAZF1*, located on chromosome 7 ($P = 2.14 \times 10^{-6}$). *JAZF1* encodes a three C2-H2-type zinc finger protein that is a transcriptional repressor of *NR2C2*, a nuclear orphan receptor that is highly expressed in prostate tissue and that reportedly interacts with the androgen receptor. *JAZF1* is a component of a fusion gene with *SUZ12* (also known as *JJAZ1*) found in endometrial stromal tumors[27].

This strategy—analyzing two phenotypes of prostate cancer, aggressive and nonaggressive—combined with genotype-based statistics rather than a trend statistic yielded strong associations that might have been missed using standard approaches. We found 13 loci that have at least one SNP with $P < 5 \times 10^{-5}$ (**Supplementary Table 2**). Of these SNPs, three are located within or in proximity of possible candidate genes—*IL16*, *CDH13* and *CPNE3*—whose associations with disease risk are best described using a recessive model for the at-risk allele yet do not meet this threshold under a trend model. Two SNPs are notable because they seem to be associated primarily with aggressive disease. In *IL16*, the lowest *P* value is observed for rs4072111, a nonsynonymous SNP that shifts serine to proline (Grantham score of 74) in the twelfth exon of this lymphocyte chemoattractant factor gene. The amino-acid change resides in the C terminus, shown to modulate T-cell activation, whereas the N terminus regulates cell cycle. Variation in gene expression of *IL16* has been associated with alleles including rs7179134, rs7180245 and rs11637363, but none of these SNPs showed association with prostate cancer in our initial GWAS ($P > 0.66$)[28]. Through analysis of the LD pattern of *IL16*, we found that the at-risk proline variant is carried by a haplotype associated with high expression. A SNP located in intron 1 of *CDH13* is also of note because of the observed reduction in transcript levels of *CDH13* in human prostate cancer cell lines and tissue samples: *in vitro*, increased expression of *CDH13* in DU145 cells inhibits tumorigenesis, whereas its silencing in BPH1 cells facilitates tumorigenesis[29]. The SNP with the lowest *P* value ($P = 1.26 \times 10^{-5}$) near *CPNE3*, a member of the calcium-dependent membrane-binding proteins known to aggregate phosphatidylserine membranes in a calcium-dependent manner, is located 7 kb downstream of its transcript. This SNP, rs4961199, is associated with altered expression of *CPNE3* in lymphoblastoid cell lines (http://www.sph.umich.edu/csg/liang/asthma/).

We observed a notable *P* value (1.01×10^{-5}) for rs12771728 on chromosome 10, chosen on the basis of a 2-SNP analysis, which suggested an augmented signal relative to the primary SNP, rs6586085, chosen in the 1-SNP model ($P = 0.00488$) in the genome-wide scan (**Supplementary Methods**). rs12771728 and rs6586085 reside near *MINPP1*, a gene encoding a phosphatase with similarity to *PTEN*. *MINPP1* has been proposed to be a low-penetrant susceptibility allele in malignant follicular thyroid tumors[30]. However, in the joint analysis, the *P* value for rs6586085 alone was 0.3, and the 2-SNP analysis with rs12771728 did not improve upon this *P* value. Further investigation is required, as this association has not been formally replicated[18].

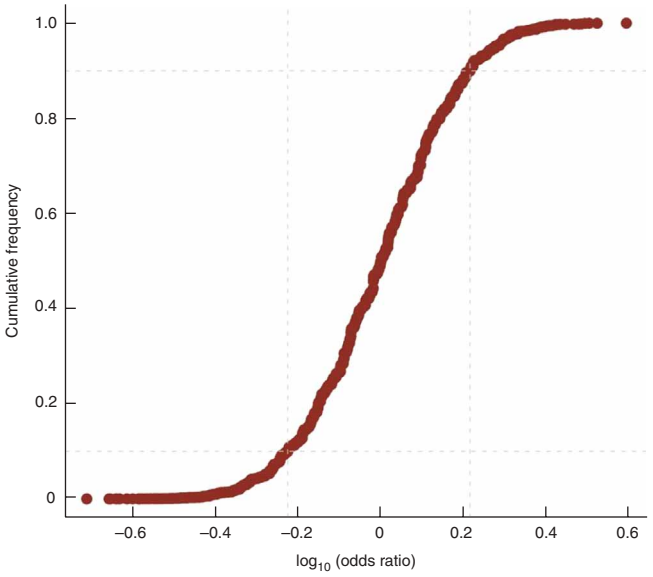

Figure 3 Distribution of odds ratios for prostate cancer risk in the European population. The allele frequencies and single locus genotype odd ratios shown in **Table 1** were used to infer the distribution of odds ratios for prostate cancer risk in the European population assuming a multiplicative multilocus odds ratio model across the seven independent risk markers. An odds ratio of 1 ($\log_{10}$ OR = 0) corresponds to the median risk in the population. With a total of 2,187 different seven-locus genotypes combinations, odds ratios vary over a roughly fourfold range (OR = 0.5–2.0; $\log_{10}$ OR = –0.3 – 0.3), with extreme categories showing greater but more imprecise risk differentials. The validity of the multiplicative model cannot be assessed, and estimates of single-locus odd ratios from the initial scans that identified the loci are likely to be exaggerated.

Of the two independent 8q24 markers (rs4242382 and rs6983267) and five SNP markers outside of 8q24 with $P \leq 2.14 \times 10^{-6}$ in the polytomous analysis, all remained strongly associated ($P \leq 3.24 \times 10^{-7}$) with risk of prostate cancer after mutual adjustment for the other SNP markers (**Supplementary Table 8** online). As there is little evidence for multiple functional variants in a single region, we examined possible interactions between regions with a single SNP per region. There was no compelling evidence that the interactions of the seven independent risk markers departed from a multiplicative model on the odds ratio scale (the minimum P value among the ${}_7C_2 = 21$ tests for adding pairwise interaction terms to the joint model was $P = 0.01$). The odds ratio comparing the men at low risk (corresponding to 10th percentile for risk) to those at high risk (corresponding to 90th percentile for risk) was 2.70 (**Fig. 3**, dashed lines). Although genotypes at these seven loci may identify men at substantially increased or decreased risk of prostate cancer, most men are expected to fall into intermediate risk categories.

Individual population attributable risks (PARs) for prostate cancer for each of the seven independent loci ranged from 8% to 20% (**Fig. 3** and **Supplementary Table 9** online). Although the total PAR from multiple loci is always less than the sum of the individual PARs, considered together, the seven loci contribute to a substantial though yet-to-be-defined fraction of prostate cancer incidence in populations of European ancestry. Our data provide a foundation for determining the cumulative risk of moderate- to low-penetrance alleles for prostate cancer and suggest that further investigation should identify additional loci. In this regard, the alleles identified to date do not fully capture the complex contribution of genetic factors to prostate cancer incidence. For at least one of these genes, *MSMB*, the strongest signal associated with prostate cancer (rs10993994) is known to be functionally active; thus, its corresponding polymorphism is a plausible causal variant. For the other loci, a search for candidate causal variants should be conducted[18,22]. The robust evidence for multiple associations presented in this study identifies genes that may elucidate the etiologic pathways contributing to the development of prostate cancer. The multiple loci reported here as well as additional loci that are likely to be identified in follow-up of this and other studies will sharpen estimates of the increased risk of prostate cancer associated with these genetic loci along an almost continuous scale.

METHODS

Analytical plan. The primary follow-up analysis explored the association between single SNPs and prostate cancer susceptibility. Briefly, 25,265 SNPs were selected on the basis of single-SNP association tests in the previously reported GWAS in PLCO (1,172 PSA-screened cases oversampled for aggressive disease and 1,157 PSC-screened controls in men of European background). An additional 1,348 SNPs were selected on the basis of two-SNP association tests, and 897 SNPs were chosen to explore high-profile candidate genes, fine-mapping of 8q24, and other non-CGEMS hypotheses. Another 1,508 SNPs were chosen to assess population structure. To maximize the number of regions that could be explored, we applied a filter based on linkage disequilibrium such that we excluded SNPs for which there was a more significant result observed with a separate SNP with an $r^2 \geq 0.8$.

For the follow-up replication studies, all one- and two-SNPs analyses were conducted using unconditional polytomous logistic regression, adjusted for age (in ten-year categories), study, and center for the two studies for which this information was available. Four continuous covariates were included to account for population stratification based on principal components analysis of genotype correlations. Analysis was carried out in three ways: for each study separately, for the pooled replication studies (all except PLCO), and for all studies combined. Genotype effects were modeled individually, and a score test with two d.f. for each case phenotype was computed (separate effects for aggressive and nonaggressive cases result in a 4-d.f. test).

Replication samples. In the second stage, we genotyped 26,958 SNPs in four follow-up studies of men of European background totaling 4,020 cases and 4,028 controls drawn from the American Cancer Society Cancer Prevention Study II[14], the Health Professionals Follow-up Study[15], CeRePP French Prostate Case-Control Study[16], and the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study[17]. These studies were approved by the appropriate institutional review boards.

Replication genotyping. We genotyped 8,693 samples (including study duplicates) passing sample handling quality-control metrics in the Core Genotyping Facility of the National Cancer Institute using a custom-designed iSelect Infinium assay (Illumina) with content described above. Using quality-control measures, we removed samples with call rates under 90% and SNPs with call rates under 95%. Fitness for Hardy-Weinberg proportion was assessed for each SNP in control subjects by study but not used to exclude SNP assays. Genotyping was attempted twice in 141 (1.6%) of samples.

A small fraction of subjects who were successfully genotyped were excluded from analysis because they: (i) were unanticipated interstudy and intrastudy duplicates; (ii) showed unanticipated African and/or Asian admixture with <85% European ancestry; (iii) belonged to sparse groups (namely, two PLCO subjects who were the only participants from one study center, and five nonaggressive cases from the portion of the CPS-II study that provided only buccal DNA); and/or (iv) had incomplete covariate data.

In the iSelect assay, a total of 525 discordant genotypes were detected out of 9,050,673 genotype comparisons (332 duplicate DNA pairs) yielding a total discordance rate of 6×10^{-5}; the lowest rate of concordance for individual SNP assays was 98.3% for the study with buccal samples and 98.9% for the study with blood samples. Infinium genotype cluster plots for notable SNPs are included in **Supplementary Methods**. For the seven loci with the lowest associated P values (three known and four newly identified), we validated genotype calls determined by the iSelect Infinium assay by TaqMan (ABI) assay in two studies, HPFS and CPS-II (buccal DNA component)[18].

Informatics. We used GLU (Genotyping Library and Utilities), a new suite of tools that is being released as an open-source application, to manage, archive and analyze the genome-wide association data. We used the STRUCTURE and EIGENSTRAT programs to assess population heterogeneity (see URLs section below).

URLs. GLU, http://cgf.nci.nih.gov/development/tooldev.html; STRUCTURE, http://pritch.bsd.uchicago.edu/structure.html; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; Tagzilla, http://tagzilla.nci.nih.gov; CGEMS portal, http://cgems.cancer.gov/data.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The Prostate Lung Colorectal Ovarian Cancer Screening Trial (PLCO) was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and by contracts from the Division of Cancer Prevention, National Cancer Institute, US National Institutes of Health, Department of Health and Human Services. The Health Professionals Follow-up Study (HPFS) study is supported by the US National Institutes of Health grant CA55075 and U01CA098233. The American Cancer Society (ACS) study is supported by U01 CA098710. The Alpha Tocopherol Beta-Carotene Cancer Prevention study (ATBC) Study is supported by the Intramural Research Program of the National Cancer Institute, NIH, and by US Public Health Service contracts N01-CN-45165, N01-RC-45035, and N01-RC-37004 from the National Cancer Institute, Department of Health and Human Services. F.R.S. is supported by a NRSA training-grant (T32 CA 09001). Centre de Recherche pour les Pathologies Prostatiques (CeRePP) thanks J.P.B. and Generali for their charitable donations which have contributed to this project. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Crawford, E.D. Epidemiology of prostate cancer. *Urology* **62**, 3–12 (2003).
2. Steinberg, G.D., Carter, B.S., Beaty, T.H., Childs, B. & Walsh, P.C. Family history and the risk of prostate cancer. *Prostate* **17**, 337–347 (1990).
3. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
4. Freedman, M.L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. USA* **103**, 14068–14073 (2006).
5. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
6. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
7. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
8. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.* **39**, 977–983 (2007).
9. Schaid, D.J. The complex genetic epidemiology of prostate cancer. *Hum. Mol. Genet.* **13 Spec No 1**, R103–21 (2004).
10. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
11. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
12. Prorok, P.C. *et al.* Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control. Clin. Trials* **21**, 273S–309S (2000).
13. Gohagan, J.K., Prorok, P.C., Hayes, R.B. & Kramer, B.S. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control. Clin. Trials* **21**, 251S–272S (2000).
14. Calle, E.E. *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490–2501 (2002).
15. Chen, Y.C. *et al.* Sequence variants of Toll-like receptor 4 and susceptibility to prostate cancer. *Cancer Res.* **65**, 11771–11778 (2005).
16. Valeri, A. *et al.* Segregation analysis of prostate cancer in France: evidence for autosomal dominant inheritance and residual brother-brother dependence. *Ann. Hum. Genet.* **67**, 125–137 (2003).
17. The ATBC Cancer Prevention Study Group. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Ann. Epidemiol.* **4**, 1–10 (1994).
18. Chanock, S.J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655–660 (2007).
19. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
20. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
21. Janssen, J.W. *et al.* Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t(11;14)(q13;q32). *Blood* **95**, 2691–2698 (2000).
22. Buckland, P.R. *et al.* Strong bias in the location of functional promoter polymorphisms. *Hum. Mutat.* **26**, 214–223 (2005).
23. Nam, R.K. *et al.* A novel serum marker, total prostate secretory protein of 94 aminoacids, improves prostate cancer detection and helps identify high grade cancers at diagnosis. *J. Urol.* **175**, 1291–1297 (2006).
24. Reeves, J.R., Dulude, H., Panchal, C., Daigneault, L. & Ramnani, D.M. Prognostic value of prostate secretory protein of 94 amino acids and its binding protein after radical prostatectomy. *Clin. Cancer Res.* **12**, 6018–6022 (2006).
25. Beke, L., Nuytten, M., Van Eynde, A., Beullens, M. & Bollen, M. The gene encoding the prostatic tumor suppressor PSP94 is a target for repression by the Polycomb group protein EZH2. *Oncogene* **26**, 4590–4595 (2007).
26. Paliwal, S. *et al.* The alternative reading frame tumor suppressor antagonizes hypoxia-induced cancer cell migration via interaction with the COOH-terminal binding protein corepressor. *Cancer Res.* **67**, 9322–9329 (2007).
27. Koontz, J.I. *et al.* Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. *Proc. Natl. Acad. Sci. USA* **98**, 6348–6353 (2001).
28. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).
29. Wang, X.D. *et al.* Expression profiling of the mouse prostate after castration and hormone replacement: implication of H-cadherin in prostate tumorigenesis. *Differentiation* **75**, 219–234 (2007).
30. Gimm, O. *et al.* Somatic mutation and germline variants of MINPP1, a phosphatase gene located in proximity to PTEN on 10q23.3, in follicular thyroid carcinomas. *J. Clin. Endocrinol. Metab.* **86**, 1801–1805 (2001).