Published in final edited form as:

Cancer Res. 2011 May 1; 71(9): 3287–3295. doi:10.1158/0008-5472.CAN-10-2646.

Large Scale Exploration of Gene-Gene Interactions in Prostate Cancer Using a Multi-stage Genome-wide Association Study

Julia Ciampa¹, Meredith Yeager², Laufey Amundadottir², Kevin Jacobs², Peter Kraft³, Charles Chung², Sholom Wacholder¹, Kai Yu¹, William Wheeler¹, Michael J. Thun⁴, W. Ryan Divers⁴, Susan Gapstur⁴, Demetrius Albanes¹, Jarmo Virtamo⁵, Stephanie Weinstein¹, Edward Giovannucci⁶, Walter C. Willett⁶, Geraldine Cancel-Tassin⁷, Olivier Cussenot⁷, Antoine Valeri⁷, David Hunter³, Robert Hoover¹, Gilles Thomas¹, Stephen Chanock², and Nilanjan Chatterjee^{1,*}

¹ Epidemiology and Biostatistics Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892, USA ² Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, NCI-Frederick, Frederick, MD 21702, USA ³ Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA ⁴ Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA 30329 ⁵ Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Helsinki, Finland ⁶ Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts ⁷ CeRePP Hopital Tenon, Assistance Publique-Hôpitaux de Paris, Paris, France

Abstract

Recent genome-wide association studies have identified independent susceptibility loci for prostate cancer (CaP) that could influence risk through interaction with other, possibly undetected, susceptibility loci. We explored evidence of interaction between pairs of 13 known susceptibility loci and single nucleotide polymorphisms (SNPs) across the genome to generate hypotheses about the functionality of CaP susceptibility regions. We used data from Cancer Genetic Markers of Susceptibility: Stage I included 523,841 SNPs in 1175 cases and 1100 controls; Stage II included 27,383 SNPs in an additional 3941 cases and 3964 controls. Power calculations assessed the magnitude of interactions our study is likely to detect. Logistic regression was used with alternative methods that exploit constraints of gene-gene independence between unlinked loci to increase power. Our empirical evaluation demonstrated that an empirical Bayes (EB) technique is powerful and robust to possible violation of the independence assumption. Our EB analysis identified several noteworthy interacting SNP pairs, although none reached genome-wide significance. We highlight a Stage II interaction between the major CaP susceptibility locus in the subregion of 8q24 that contains *POU5F1B* and an intronic SNP in the transcription factor *EPAS1*, which has potentially important functional implications for 8q24. Another noteworthy result involves interaction of a known CaP susceptibility marker near the prostate protease genes *KLK2* and *KLK3* with an intronic SNP in *PRXX2*. Overall, the interactions we have identified merit follow-up study, particularly the *EPAS1* interaction which has implications not only in CaP but also in other epithelial cancers that are associated with the 8q24 locus.

*Corresponding author: chattern@mail.nih.gov; phone, 301-402-7933; fax, 301-402-0081.

Introduction

Recent genome-wide association studies (GWAS) have identified multiple single nucleotide polymorphisms (SNPs) associated with risk of prostate cancer (CaP) (1–8). Functional variants in linkage disequilibrium (LD) with the established markers may directly contribute to CaP risk through interactions with other, yet undetected, susceptibility loci. Large datasets are required for exploration of interactions between known risk alleles and the remainder of the genome. Such large scale studies could give rise to the discovery of novel susceptibility regions and a better understanding of the biology of CaP. In this report, we present the first results from a study of gene-gene interactions in the etiology of CaP using data from Stages I and II of the Cancer Genetics Markers of Susceptibility (CGEMS) Initiative.

To identify SNPs that may interact with established susceptibility regions to affect CaP risk, we conducted a series of conditional genome scans. The susceptibility (conditioning) regions included nine gene regions and four independent regions within the chromosomal region 8q24. All have demonstrated strong associations with CaP in recent GWAS, including but not limited to CGEMS (Table 1).

The region of 8q24 warrants close attention because it contains independent risk markers for prostate and additional cancers (Figure 1). Despite its strong associations with multiple cancers, the physiologic function of 8q24 remains an area of investigation. Molecular studies have focused on Region 3 of 8q24 which is associated with several epithelial cancers, including prostate and colorectal. The most significant marker in that region, rs6983267, is part of a consensus binding sequence for TCF (9), a family of transcription factors that are nuclear targets of WNT signaling. The risk allele has been shown to participate in long-range regulation of the WNT-targeted oncogene *MYC* that is telomeric to the regions of 8q24 associated with multiple cancers (10–12). To a lesser degree than *MYC*, *POU5F1B* (also called *POU5F1P1*) has drawn attention as a plausible candidate gene to explain the underlying biology of the association signals (13–16). It is the only confirmed gene within the extended 8q24 region, located specifically in Region 3. Until recently *POU5F1B* was classified as a highly homologous (15) pseudogene of *POU5F1* (also called *OCT4* or *OCT3*), which is a central gene in the regulation of stem cell pluripotency (17) (Figure 2a). Recent reports on *POU5F1B* have demonstrated that it produces a protein with similar function to *POU5F1* (14) and that it is over-expressed in CaP (13).

We applied three methods that are available for exploring gene-gene interactions in case-control studies. Traditionally, logistic regression has been the most popular method for analysis of case-control data. In recent years, a number of reports have noted that the power for exploring gene-gene interactions from case-control studies can be greatly enhanced by alternative methods that exploit the assumption of gene-gene independence between distant loci (18,19). These methods, however, can be very sensitive to violation of the underlying gene-gene independence assumption (20). Our analysis provides the first empirical assessment of the performance of these alternative methods for large scale exploration of gene-gene interactions in a GWAS setting.

Methods

CGEMS Stage I

The details of CGEMS Stage I have been published previously (6). Briefly, the subjects available for this study included 1175 cases and 1100 controls of European ancestry from the Prostate, Lung, Colon and Ovarian Screening Trial. They were genotyped using two Illumina chips (HumanHap 300 and HumanHap240) that constituted 523,841 autosomal SNPs.

CGEMS Stage II

The details of CGEMS Stage II have been published previously (5). Briefly, the subjects included an additional 3941 cases and 3964 controls of European ancestry. They represent five studies (case/control): Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study in Finland (929/921), Health Professionals Follow-up Study in America (596/611), American Cancer Society Cancer Prevention Study II Nutrition Cohort (1760/1775), and CeRePP French Prostate Case-Control Study (656/657). Subjects were genotyped on 27,383 autosomal SNPs that showed evidence of association in single-SNP Stage I analyses ($p < 0.05$).

We performed a series of conditional genome scans, using data from CGEMS Stage I or Stage II. We chose to study the 13 regions that have already been shown to be strongly associated with risk of CaP in CGEMS or other recent GWAS (Table 1). For each region, except *EEFSEC* we used the SNP from the original publication. For *EEFSEC*, we chose the most significant SNP within the gene in the combined CGEMS Stage I and II analysis.

In each scan, we “conditioned” on the most notable SNP from each of 13 well-established susceptibility regions and tested for interaction of the conditioning SNP with all the remaining “scan” SNPs across the genome. Tests for interaction for each scan SNP used a logistic regression model that included the main effect of the conditioning SNP, the main effect of a scan SNP, an interaction term that captured any non-multiplicative effect between the two markers, and adjusting covariates for study and DNA extraction. We assumed alleles within a locus affect risk of the disease in an additive fashion (on the logistic scale) and, thus, coded genotype data for each SNP as a continuous variable and the interaction as the product of the allele counts. In addition to performing standard tests for interactions, we also conducted “omnibus” (21) tests that simultaneously evaluated the significance of the main- and interaction- effect of the scan SNP.

For statistical inference under the logistic model, we used three alternative methods: (a) Unconstrained maximum-likelihood (UML), (b) Constrained maximum-likelihood (CML) and (c) Empirical Bayes (EB). The UML method corresponds to the traditional prospective logistic regression analysis that obtains maximum-likelihood estimates of the odds ratio parameters with no constraint on the joint distribution of scan and conditioning SNPs in the underlying population. The CML method obtains the maximum-likelihood estimates of the same odds ratio coefficients assuming gene-gene independence in the underlying population between the scan and conditioning SNPs (22). While the advantage of the UML method is that its validity does not require any assumption of joint genotype distribution in the underlying population, the CML and analogous methods are known to be much more powerful when the underlying assumptions of independence are valid (19,23). For the CML analysis, we excluded scan SNPs within 500Kb of the conditioning SNP to minimize gene-gene dependence due to physical proximity. For assessment and estimation of odds ratio interactions, the CML method essentially produces inferences very similar to the popular case-only method (18,19). Unlike the case-only approach, however, the CML method yields both interaction coefficients and main effects, which are needed for performing omnibus tests as well as for contextual interpretation of interaction.

The third approach we considered involved the recently proposed EB method which exploits the gene-gene or gene-environment independence in a more data-adaptive fashion so that bias can be avoided when the assumptions of independence are violated in the underlying population (24). The method obtains parameter estimates by taking weighted averages of those from the UML and CML methods where the weights depend on two key quantities: (a) the bias of the CML method which could be estimated by the difference of the estimates it produces from those obtained from the UML method and (b) the variance of the parameter

estimates from the UML method. As the magnitude of the bias of the CML method increases, the EB method puts more weight on the UML method. Previous simulation studies have suggested that the EB method can strike a good balance between bias and efficiency in large scale studies where the assumptions of independence are likely to be satisfied for most, but not all, combinations of gene-gene and gene-environment factors under study (25). The parameter estimates and standard errors for logistic regression coefficients were obtained in the R package called Case Control. Genetics (26) that implements all three methods.

For each conditional region, we ran two separate scans, one for all 523,841 Stage I SNPs and the other for the subset of 27,383 SNPs that were available in both Stages I and II. We performed the scan for Stage II SNPs using only the Stage II data to make the analysis independent of the selection effect from Stage I. For each scan, we considered Bonferroni adjustment for the appropriate number of SNPs to declare genome-wide significance at p -value<0.05 level. To give higher priority for potential “cis” effects, we separately examined the significance of interaction for scan SNPs within 500Kb of the conditioning gene(s) by adjusting for multiple testing only within that region, using the EB method.

As we aimed to generate hypotheses about the functionality of susceptibility regions in the etiology of CaP, we conducted literature reviews and bioinformatics searches to investigate potential biological mechanisms that could underlie the interactions we observed. For SNPs in or near gene regions, we considered whether those genes were known or thought to participate in processes that relate to cancer or to the function of the relevant conditioning region. This work can help prioritize interacting SNP pairs for follow-up study. It guided our follow-up analysis on the top interaction result for 8q24 Region 3. Using Stage II data, we jointly modeled the top SNPs in *ESRRB* and *SALL4* for interaction with 8q24 Region 3 in a single logistic regression that we analyzed using UML. We focused on the transcription factors *ESRRB* and *SALL4* because, like the top-ranking *EPAS1* (also known as *HIF-2A*), they are positive regulators of *POU5F1* (27,28) (Figure 2). They differ from *EPAS1* in that they do not require hypoxic stimuli and that they are also target genes of *POU5F1* (29). We examined the top SNP from each gene based on Stage I main effects results (*ESRRB* p -value=0.001, *SALL4* p -value=0.28). Both SNPs are located in an intron of their respective gene. rs7155416 in *ESRRB* on chromosome 14q24 has a minor allele frequency (MAF) of 0.12 and rs6021460 in *SALL4* on chromosome 20q13 has a MAF of 0.44.

Results

Inspection of the quantile-quantile plots for omnibus and interaction tests for all conditional genome scans suggest that, in general, none of the methods was affected by large scale systematic bias or overdispersion. However, in some instances, the CML method showed more statistically significant associations than would be expected by chance (see, for example, Figure 3) even when we excluded all scan SNPs on the same chromosome as the conditioning SNP. In contrast, the UML and data adaptive EB methods did not show any evidence of such excess significance. A possible explanation for this phenomenon is population stratification that could cause long range dependence, violating the underlying assumption of gene-gene independence of the CML method (30). In the subsequent sections, we report the main findings of our conditional scans based on the EB method that is known to be more powerful than standard logistic regression analysis and yet, unlike CML type methods, is resistant to bias due to violation of the gene-gene independence assumption due to population stratification or otherwise.

In the Stage I analysis, one SNP, rs2002865, reached genome-wide significance for interaction after Bonferroni correction (p -value=9.14e-10) in the conditional scan of 8q24

Region 4. A second SNP, rs4960563, in strong LD with rs2002865 ($r^2=0.68$), was also highly significant for interaction with the same 8q24 conditioning region (p-value=3.79e-6). Both interactions, however, failed to replicate (p-value>0.05), when the SNP was followed-up with further genotyping in a subset of the Stage II sample. Within the extended *JAZF1* conditioning region, one SNP met “region-wide” significance for potential cis-interaction: rs4857841 on chromosome 7p15. It also failed to replicate in follow-up.

In the Stage II conditional scans, no SNP reached genome-wide significance for interaction. A list of top-ranking SNPs (p-value < 1.0e-4) from each conditional scan is shown in Table 2. The most notable finding considering biologic plausibility was an interaction between rs6983267 in 8q24 Region 3 which contains *POU5F1B* and rs4953347 in the first intron of *EPAS1* which upregulates *POU5F1* (p-value=9.69e-5, multiplicative odds ratio=1.13). In a follow-up analysis we detected a significant interaction between 8q24 Region 3 and both *ESRRB* and *SALL4* (Table 3), which have functional similarities to *EPAS1* (Figure 2). Also of note is the top-ranking SNP for interaction with the *KLK2-KLK3* conditioning region: rs1558874 intronic to *PRRX2* (p-value=4.80e-5, multiplicative odds ratio=1.33). Those SNPs demonstrated evidence of an interaction in the same direction and of similar magnitude in the independent CGEMS Stage I data (p-value=0.047, multiplicative odds ratio=1.27).

We highlight two region-wide significant results from our investigation of potential “cis” interactions. The first involves rs4314620 in the extended 8q24 region. It showed significant interaction in the conditional scans for Region 2 (p-value=0.003), Region 3 (p-value=0.0004) and Region 4 (p-value=0.02) (Figure 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (p-value=2.48e-5). The second result involves rs17714461 for the *KLK2-KLK3* region (p-value=7.14e-4). That SNP resides in chromosome 19q13, located ~15Kb from *KLK4*. It is ~60Kb from the conditioning SNP with which it does not demonstrate LD ($r^2<0.001$).

Discussion

Our study presents one of the first large scale explorations of gene-gene interactions in the setting of a multi-stage GWAS. Our analysis identifies a list (Table 2) of pair-wise SNP interactions that, through follow-up study, may elucidate the functional relevance of CaP susceptibility SNPs. Our analysis also provides insights into future methodological challenges that large scale studies will face in establishing conclusive interaction with either the primary or a secondary trait. It presents an empirical evaluation of modern methods for interaction analyses using case-control data.

The results we report for the conditional scans were obtained using a recently proposed EB method. We focused on that method because previous simulations demonstrated it is more powerful than standard logistic regression (25) and our empirical evaluation suggests it is more robust than case-only type methods. Our observation that the CML method can suffer bias even when scan and conditioning SNPs are on different chromosomes is particularly cautionary. The robustness of the EB method is expected to be similarly beneficial in studies of gene-environment interactions for which it can be difficult to assess an independence assumption.

Perhaps the most noteworthy result of our study is the top SNP for interaction with 8q24 Region 3 in Stage II: rs4953347 which is intronic to *EPAS1*. That gene belongs to a family of hypoxia-inducible factors that promote key carcinogenic processes such as angiogenesis and metastasis (31). Under hypoxic conditions, which are common in malignant tumors,

EPAS1 directly binds and activates *POU5F1* (32,33). By activating *POU5F1*, EPAS1 has been shown to promote tumorigenesis (34). Both *EPAS1* and *POU5F1B* are over-expressed in CaP, but *POU5F1* is not expressed in either healthy or malignant prostate tissue (13,31). Given these data, we propose that *POU5F1B* mediates the observed *EPAS1*-8q24 Region 3 interaction. Our hypothesis involves an assumption that EPAS1 participates in the regulation of *POU5F1B*, which is currently poorly understood.

Kastler et. al suggested the over-expression of *POU5F1B* in CaP may mimic the ectopic expression of *POU5F1* (13), which has been shown to promote epithelial tumors (35). That hypothesis aligns well with reports that CaP progression involves the reactivation of embryonic pathways (36) because *POU5F1* is central to the regulation of stem cell pluripotency (17) and its encoded transcription factor is functionally similar to the protein of *POU5F1B* (37). Specifically, multipotential progenitor cells are thought to be seeds of tumorigenesis in CaP (37) and ectopic *POU5F1* expression is thought to promote epithelial tumorigenesis by inhibiting the differentiation of progenitor cells (35). Given these data we cautiously hypothesize that the CaP association of 8q24 Region 3 involves a type of pluripotency network centered on *POU5F1B* rather than on *POU5F1*. Our follow-up analysis of 8q24 Region 3 with *ESRRB* and *SALL4* offers preliminary support of the hypothesis because those genes, which function as both regulators and targets of *POU5F1*, demonstrate a significant interaction with 8q24 Region 3.

The preceding results should be interpreted cautiously. Future effort is needed to replicate the finding of statistical interaction for *EPAS1* and 8q24 Region 3 in independent studies. To obtain conclusive evidence, the sample size for those studies needs to be large due to modest magnitude of the anticipated interaction. Even if our findings replicate, it cannot provide direct evidence for the proposed model underlying the interaction. Additional functional studies would be needed. We believe these future studies should consider not only CaP but rather all epithelial cancers associated with 8q24 Region 3 because, in subsets of those cancers, *EPAS1* is over-expressed (31), mRNA transcripts of *POU5F1B* have been detected (15), and embryonic pathways are implicated (38). Notably, all those features characterize colon cancer.

The Stage II analysis produced other notable results, including two for the conditioning region *KLK2-KLK3*. The region-wide significant result for rs17714461 is noteworthy because its nearby gene, *KLK4*, has been shown to stimulate cellular proliferation in CaP in conjunction with *KLK2* (39), and additional reports have linked *KLK4* to various aspects of CaP progression, including mesenchymal transition, invasion and metastasis (40–42). The top SNP for interaction in the *KLK2-KLK3* conditional scan was rs1558875 an intronic SNP to *PRRX2* that is also associated with cellular proliferation (43). This interaction replicated in our relatively small, independent CGEMS Stage I analysis. These preliminary results suggest that the *KLK2-KLK3* susceptibility region contributes to CaP risk through interaction with genes involved in cellular proliferation. Functional follow-up studies and additional replication efforts are warranted.

A second notable finding from our investigation of potential “cis” interactions involved rs4314620 for the extended 8q24 region. Its pair-wise interactions with three independent known risk alleles for CaP within the 8q24 region suggest different 8q24 susceptibility loci may be related by some common underlying biologic mechanism. It is hard to speculate what that mechanism may be because it is a gene-poor region, but one possible explanation for the observed interactions is long-range gene regulation. rs4314620 resides in a sub-region of 8q24 that is associated with bladder cancer (Figure 1) (44) and contains two regulatory regions for the oncogene *MYC* that flanks 8q24 Region 4 (45).

In our analysis of CGEMS Stage I data, one SNP-pair exceeded genome-wide significance for interaction in the conditional scan for 8q24 Region 4. The result was unlikely to be due to genotyping error as a second SNP in strong LD with the original signal also showed strong significance. Yet, the interaction failed to replicate when we genotyped the SNPs in an additional 2439 cases and 2241 controls in CGEMS Stage II. This example illustrates the challenge of employing rank p-values for prioritization of interaction as well as of establishing the threshold needed for a conclusive finding. We note that Stage I of CGEMS, which included 1175 cases and 1100 controls, was underpowered for study of interactions (Figure 4). In the future, one way of reducing such false positives would be to consider Bayesian methods (46,47) that can incorporate both power and biological plausibility into measures of statistical significance. Another strategy to gain power, particularly for initial GWAS stages, is meta-analysis of multiple studies, enabling one to increase sample size while retaining the full array of SNPs.

Due to the scarcity of highly significant findings, we carefully examined the power of CGEMS Stages I and II to detect interactions at genome-wide significance levels ($\alpha = 1.0 \times 10^{-7}$ and 1.85×10^{-6} for Stage I and II, respectively; Figure 4). In these calculations, we focused only on quantitative interactions where the effect of one locus can be modified, but not reversed, by the other locus and vice versa. Stage I of CGEMS had virtually no power to detect interaction odds ratios in the scenarios we examined which ranged 1.13–2.05. The larger Stage II, in contrast, had high power for detecting modest to large interaction odds ratio (≥ 1.7) even after accounting for the fact that some power is lost due to selection of the SNPs at Stage I by main effect only. It is notable that under a model of quantitative interaction, a larger interaction odds ratio also corresponds to larger main effects. Thus, the loss of power at Stage I due to the selection by main effect is often small when the interaction odds ratio and MAFs are reasonably large. In contrast, in the presence of qualitative interaction, the main effects of loci could be very weak or even non-existent even when the interaction odds ratio is large. The power for detecting such loci in our analysis is low, as the probability of selecting them for Stage II is low.

We conclude that the susceptibility SNPs we have studied are unlikely to have quantitative interactions of large magnitude with other SNPs in the genome. Theoretical calculations (48), as well as a lack of findings of epistasis for other diseases, also point towards the possibility that large non-multiplicative or non-additive effects may not be abundant in the etiology of complex traits. It is possible that our study has missed qualitative interactions, but the biologic plausibility of the presence of many such extreme types of interaction is questionable. It is also possible that epistasis, if it plays an important role in the etiology of CaP, will have a much more complex form than the pairwise SNP-SNP interactions we studied. Finding such higher order interactions in large scale studies, however, will remain an intrinsically challenging problem because of both the computationally daunting task of exploring all possible multi-locus models and the requirement for extremely large sample sizes that will be necessary to achieve sufficient power while minimizing the chance of false positives.

In the future, detecting evidence of gene-gene interactions through study of statistical interactions between SNP markers will likely require very large sample sizes that are achievable only by sharing individual level data in consortiums of GWAS. For smaller scale studies, the exercise of exploring gene-gene interactions is unlikely to lead to definitive findings, but it can be useful in generating lists of loci that require follow-up in replication studies. Incorporating biological knowledge from reliable pathway and network databases could enhance the power for detection, validation, and interpretation of interaction.

Our exploration of gene-gene interactions in CGEMS identified a list of SNPs that require future replication effort with varying degrees of priority (Table 2). We hope its public availability will motivate replication studies. The EB method we highlight is appropriate for those analyses, as it is both powerful and robust. We consider our most notable finding to be an interaction between SNPs in *EPAS1* and 8q24 Region 3 because it generates a preliminary hypothesis about the poorly understood association of 8q24 Region 3 with multiple epithelial cancers that centers on the recently characterized gene *POU5F1B*. A second result with high priority for follow-up is an interaction between *PRRX2* and the *KLK2-KLK3* region. It suggests that the functional relevance of the *KLK2-KLK3* susceptibility region in the etiology CaP may involve cellular proliferation.

Acknowledgments

This study utilized the high-performance computation capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD.

Reference List

1. Duggan D, Zheng SL, Knowlton M, Benitez D, Dimitrov L, Wiklund F, et al. Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. *J Natl Cancer Inst.* 2007; 19(99):1836–44. [PubMed: 18073375]
2. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet.* 2008; 40:316–21. [PubMed: 18264097]
3. Gudmundsson J, Sulem P, Rafnar T, Bergthorsson JT, Manolescu A, Gudbjartsson D, et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet.* 2008; 40:281–3. [PubMed: 18264098]
4. Gudmundsson J, Sulem P, Gudbjartsson DF, Blondal T, Gylfason A, Agnarsson BA, et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet.* 2009; 41:1122–6. [PubMed: 19767754]
5. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet.* 2008; 40:310–5. [PubMed: 18264096]
6. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* 2007; 39:645–9. [PubMed: 17401363]
7. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, Hayes RB, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet.* 2009; 41:1055–7. [PubMed: 19767755]
8. Zheng SL, Stevens VL, Wiklund F, Isaacs SD, Sun J, Smith S, et al. Two independent prostate cancer risk-associated Loci at 11q13. *Cancer Epidemiol Biomarkers Prev.* 2009; 18:1815–20. [PubMed: 19505914]
9. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet.* 2009; 41:882–4. [PubMed: 19561607]
10. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A.* 2010; 107:9742–6. [PubMed: 20453196]
11. Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, et al. Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A.* 2010; 107:3001–5. [PubMed: 20133699]
12. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a Large Chromatin Loop Linked to a Cancer Risk-Associated Single-Nucleotide Polymorphism in Colorectal Cancer Cells. *Mol Cell Biol.* 2010; 30:1411–20. [PubMed: 20065031]

13. Kastler S, Honold L, Luedeke M, Kuefer R, Moller P, Hoegel J, et al. POU5F1P1, a putative cancer susceptibility gene, is overexpressed in prostatic carcinoma. *Prostate*. 2009; 70:666–74. [PubMed: 20017164]
14. Panagopoulos I, Moller E, Collin A, Mertens F. The POU5F1P1 pseudogene encodes a putative protein similar to POU5F1 isoform 1. *Oncol Rep*. 2008; 20:1029–33. [PubMed: 18949397]
15. Suo G, Han J, Wang X, Zhang J, Zhao Y, Zhao Y, et al. Oct4 pseudogenes are transcribed in cancers. *Biochem Biophys Res Commun*. 2005; 337:1047–51. [PubMed: 16229821]
16. Zheng SL, Sun J, Cheng Y, Li G, Hsu FC, Zhu Y, et al. Association between two unlinked loci at 8q24 and prostate cancer risk among European Americans. *J Natl Cancer Inst*. 2007; 99:1525–33. [PubMed: 17925536]
17. Pan GJ, Chang ZY, Scholer HR, Pei D. Stem cell pluripotency and transcription factor Oct4. *Cell Res*. 2002; 12:321–9. [PubMed: 12528890]
18. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol*. 1996; 144:207–13. [PubMed: 8686689]
19. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994; 13:153–62. [PubMed: 8122051]
20. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*. 2001; 154:687–93. [PubMed: 11590080]
21. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007; 63:111–9. [PubMed: 17283440]
22. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92:399–418.
23. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med*. 1997; 16:1731–43. [PubMed: 9265696]
24. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008; 64:685–94. [PubMed: 18162111]
25. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol*. 2008; 32:615–26. [PubMed: 18473390]
26. CGEN. [Internet]. Maryland: National Cancer Institute; 2010. [cited 19 July 2010]. Available from <http://dceg.cancer.gov/about/staff-bios/chatterjee-nilanjan>
27. Zhang X, Zhang J, Wang T, Esteban MA, Pei D. Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J Biol Chem*. 2008; 283:35825–33. [PubMed: 18957414]
28. Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, et al. Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat Cell Biol*. 2006; 8:1114–23. [PubMed: 16980957]
29. Sharov AA, Masui S, Sharova LV, Piao Y, Aiba K, Matoba R, et al. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*. 2008; 9:269. [PubMed: 18522731]
30. Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, et al. Using Principal Components of Genetic Variation for Robust and Powerful Detection of Gene-Gene Interactions in Case-Control and Case-Only Studies. *Am J Hum Genet*. 2010; 86:331–342. [PubMed: 20206333]
31. Rankin EB, Giaccia AJ. The role of hypoxia-inducible factors in tumorigenesis. *Cell Death Differ*. 2008; 15:678–85. [PubMed: 18259193]
32. Simon MC, Keith B. The role of oxygen availability in embryonic development and stem cell function. *Nat Rev Mol Cell Biol*. 2008; 9:285–96. [PubMed: 18285802]

33. Forristal CE, Wright KL, Hanley NA, Oreffo RO, Houghton FD. Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions. *Reproduction*. 2010; 139:85–97. [PubMed: 19755485]
34. Covello KL, Kehler J, Yu H, Gordan JD, Arsham AM, Hu CJ, et al. HIF-2 α regulates Oct-4: effects of hypoxia on stem cell function, embryonic development, and tumor growth. *Genes Dev*. 2006; 20:557–70. [PubMed: 16510872]
35. Hochedlinger K, Yamada Y, Beard C, Jaenisch R. Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell*. 2005; 121:465–77. [PubMed: 15882627]
36. Schaeffer EM, Marchionni L, Huang Z, Simons B, Blackman A, Yu W, et al. Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. *Oncogene*. 2008; 27:7180–91. [PubMed: 18794802]
37. van Leenders GJ, Schalken JA. Epithelial cell differentiation in the human prostate epithelium: implications for the pathogenesis and therapy of prostate cancer. *Crit Rev Oncol Hematol*. 2003; 46 (Suppl):S3–10. [PubMed: 12850522]
38. Ricci-Vitiani L, Lombardi DG, Pilozzi E, Biffoni M, Todaro M, Peschle C, et al. Identification and expansion of human colon-cancer-initiating cells. *Nature*. 2007; 445:111–5. [PubMed: 17122771]
39. Mize GJ, Wang W, Takayama TK. Prostate-specific kallikreins-2 and -4 enhance the proliferation of DU-145 prostate cancer cells through protease-activated receptors-1 and -2. *Mol Cancer Res*. 2008; 6:1043–51. [PubMed: 18567807]
40. Gao J, Collard RL, Bui L, Herington AC, Nicol DL, Clements JA. Kallikrein 4 is a potential mediator of cellular interactions between cancer cells and osteoblasts in metastatic prostate cancer. *Prostate*. 2007; 67:348–60. [PubMed: 17221837]
41. Wang W, Mize GJ, Zhang X, Takayama TK. Kallikrein-related peptidase-4 initiates tumor-stroma interactions in prostate cancer through protease-activated receptor-1. *Int J Cancer*. 2009
42. Whitbread AK, Veveris-Lowe TL, Lawrence MG, Nicol DL, Clements JA. The role of kallikrein-related peptidases in prostate cancer: potential involvement in an epithelial to mesenchymal transition. *Biol Chem*. 2006; 387:707–14. [PubMed: 16800731]
43. Stelnicki EJ, Arbeit J, Cass DL, Saner C, Harrison M, Largman C. Modulation of the human homeobox genes PRX-2 and HOXB13 in scarless fetal wounds. *J Invest Dermatol*. 1998; 111:57–63. [PubMed: 9665387]
44. Kiemeny LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet*. 2008; 40:1307–12. [PubMed: 18794855]
45. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*. 2006; 124:47–59. [PubMed: 16413481]
46. Wacholder S, Chanock S, Garcia-Closas M, El GL, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004; 96:434–42. [PubMed: 15026468]
47. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*. 2007; 81:208–27. [PubMed: 17668372]
48. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*. 2008; 4:e1000008. [PubMed: 18454194]
49. Boiani M, Scholer HR. Regulatory networks in embryo-derived pluripotent stem cells. *Nat Rev Mol Cell Biol*. 2005; 6:872–84. [PubMed: 16227977]
50. Kang J, Shakya A, Tantin D. Stem cells, stress, metabolism and cancer: a drama in two Acts. *Trends Biochem Sci*. 2009; 34:491–9. [PubMed: 19733480]

Figure 1. Linkage disequilibrium and cancer susceptibility pattern for 8q24 region

The linkage disequilibrium heat map was drawn using HapMap I+II release 22 CEU data from 127,948Kb to 128,950Kb genomic region (reference build 36.3). The arrowheads indicate probable recombination hotspots according to the HapMap I+II. Five distinct regions have been associated with prostate cancer risk (Regions 1–5). Region 3 is also conclusively associated with colorectal cancer. Separate regions have been associated with breast (Region B) and bladder (Region BL) cancers. The oncogene *MYC* flanks the region telomerically. Regions 1–4 were studied in our conditional genome scans, with conditioning SNPs labeled in the plot. SNP rs431620 of the BL region demonstrated interaction separately with Regions 2–4. An omnibus test on its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant ($p\text{-value}=2.48\text{e-}5$).

Figure 2. Simplified schematics of the pluripotency network with emphasis on *POU5F1*

Panel A presents a subset of genes in this complex network and highlights those (green) that demonstrate interaction with 8q24 Region 3 in CGEMS Stage II. Sharp (blunt) arrows represent excitatory (inhibitory) relationships between genes; WNT, FZD and SMAD respectively represent the genes families of wingless type protein, frizzled, and similar to mothers against decapentaplegi; Ids = inhibitors of differentiation. Adapted from Boaini et al (49). Panel B highlights the known functional similarity of EPAS1, ESRRB and SALL4, transcription factors that serve as promoters of *POU5F1*. Dotted arrows represent binding. CR1-3 represent highly conserved promoter and enhancer regions of *POU5F1*. Adapted from Kang et al (50).

Figure 3. Quantile-quantile plots for interaction p-values from genome scan conditional on one susceptibility SNP near *MSMB*

P-values were computed through Wald tests on the estimate of a multiplicative interaction between a susceptibility locus near *MSMB* and each of 27,053 scan SNPs in CGEMS Stage II. Estimates were computed via three methods: unconstrained maximum likelihood (UML, left), constrained maximum likelihood (CML, middle), and an empirical Bayes method (EB, right). Plots exclude SNPs within 500Kb of the conditioning *MSMB* SNP. P-values less than 1.0×10^{-6} were treated as 1.0×10^{-6} .

Figure 4. Power curves for detecting interactions in CGEMS at genome wide significance
 Power calculations assume: a) independent scan and conditioning SNPs, b) no main effect in interaction model for scan or conditioning SNP, c) dominant disease model and d) equal number of cases and controls (Stage I, 1100; Stage II, 4000). Stage II calculations incorporate the reduction in power due to selection of SNPs based on main effects in Stage I (significance threshold=0.05). Plots are presented for three values of the marginal odds ratio for the conditioning SNP (1.1, left; 1.15, middle; 1.2, right). Power curves are plotted against minor allele frequency (MAF) of the conditioning SNP for five values for MAF of the scan SNP and corresponding interaction odds ratio, given in the legend for each column. Note, under the assumed model, the MAF of the conditioning SNP restricts the magnitude of interaction since the marginal odds ratio for the conditioning SNP is fixed. The significance thresholds for Stage I (top) and Stage II (bottom) scans were 1.0×10^{-7} and 1.85×10^{-6} , respectively.

Table 1
Summary of thirteen conditioning regions individually studied in conditional genome scans.

Conditioning SNP	Conditioning Region	SNP-Region Proximity	Chr	Minor Allele	MAF	OR (95%CI)	P-value
rs4962416	<i>CTBP2</i>	Intron	10	C	0.25	1.18 (1.11, 1.26)	1.05E-07
rs1571801	<i>DAB2IP</i>	Intron	9	A	0.26	1.10 (1.04, 1.18)	2.01E-03
rs4857841	<i>EEFSEC</i>	Intron	3	A	0.28	1.15 (1.08, 1.22)	6.79E-06
rs17432497	<i>EHBP1</i>	Intron	2	G	0.18	1.10 (1.03, 1.18)	6.49E-03
rs4430796	<i>HNF1B</i>	Intron	17	G	0.48	0.82 (0.77, 0.87)	4.56E-12
rs10486567	<i>JAZF1</i>	Intron	7	A	0.24	0.83 (0.78, 0.89)	4.89E-08
rs2735839	<i>KLK2, KLK3</i>	Near	19	A	0.14	0.92 (0.85, 1.00)	5.38E-02
rs10993994	<i>MSMB</i>	Near	10	T	0.38	1.23 (1.16, 1.30)	7.47E-13
rs10896449	<i>MYEOV</i>	Near	11	A	0.50	0.84 (0.80, 0.89)	8.30E-10
rs4242382	8q24, Region 1	Within	8	A	0.10	1.48 (1.36, 1.61)	<1.0E-15
rs7841060	8q24, Region 2	Within	8	G	0.20	1.21 (1.13, 1.30)	1.08E-07
rs6983267	8q24, Region 3	Within	8	T	0.50	0.80 (0.76, 0.84)	1.55E-15
rs620861	8q24, Region 4	Within	8	T	0.37	0.90 (0.84, 0.95)	5.62E-04

Reported values are based on single-SNP analyses of CGEMS Stages I and II combined.

Chr=chromosome; MAF=minor allele frequency among controls in sample; OR=odds ratio; CI=confidence interval.

Table 2

Results for top SNPs in CGEMS Stage II grouped by conditional scans and ranked by interaction p-values ($<1.0\text{e-}4$).

Conditioning Region	Interacting Region			Marginal* OR (95% CI)	Multiplicative Interaction	
	SNP	MAF	Chr, Gene		OR (95% CI)	P-value
CTPB2	rs7765379	0.12	6	0.99 (0.90, 1.09)	0.80 (0.71, 0.89)	9.83E-05
	rs4242	0.32	22	1.03 (0.97, 1.11)	1.18 (1.10, 1.28)	1.37E-05
DAB2IP	rs12489404	0.42	3, GRM7	0.99 (0.93, 1.05)	0.85 (0.79, 0.91)	3.10E-06
	rs10458466	0.49	1	0.97 (0.91, 1.03)	1.15 (1.07, 1.24)	6.52E-05
EEFSEC	rs10514124	0.35	5	1.00 (0.93, 1.06)	0.83 (0.76, 0.90)	1.91E-05
	rs704638	0.31	4	0.97 (0.90, 1.03)	1.19 (1.10, 1.30)	4.13E-05
EHBP1	rs7604809	0.12	2, GTF3C3	1.00 (0.91, 1.10)	1.27 (1.13, 1.42)	5.05E-05
	rs10506678	0.43	12	1.00 (0.94, 1.07)	1.14 (1.07, 1.22)	5.35E-05
HNF1B	rs617182	0.46	17, PHOSPHO1	0.95 (0.89, 1.01)	1.15 (1.07, 1.23)	7.14E-05
	rs4691238	0.41	4	0.95 (0.90, 1.02)	1.14 (1.07, 1.22)	8.37E-05
JAZF1	rs745720	0.19	2	0.96 (0.88, 1.03)	1.23 (1.12, 1.35)	2.28E-05
	rs2899748	0.35	15, GLCE	0.96 (0.90, 1.02)	0.84 (0.77, 0.91)	3.04E-05
KLK2, KLK3	rs1558874	0.12	9, PRRX2	1.05 (0.95, 1.16)	1.33 (1.16, 1.53)	4.80E-05
	rs12196677	0.10	6, PNPLA1	0.94 (0.85, 1.05)	1.37 (1.18, 1.60)	5.88E-05
MSMB	rs10935317	0.42	3	0.98 (0.92, 1.05)	1.14 (1.07, 1.22)	3.93E-05
	rs12605415	0.38	18	0.99 (0.93, 1.06)	1.14 (1.07, 1.22)	6.20E-05
	rs11083271	0.32	18	1.02 (0.96, 1.09)	1.16 (1.08, 1.24)	6.67E-05
	rs9880831	0.31	3, LOC391524	1.04 (0.97, 1.12)	0.87 (0.82, 0.93)	6.90E-05
MYEOV	rs7921651	0.23	10, FAM107B	1.03 (0.97, 1.12)	0.86 (0.79, 0.93)	8.62E-05
	rs1240224	0.30	12	1.03 (0.96, 1.10)	1.15 (1.07, 1.23)	9.00E-05
8q24 Region 1	rs4660403	0.21	1, LOC388621	0.93 (0.86, 1.01)	1.24 (1.12, 1.38)	7.57E-05
8q24 Region 2	rs991000	0.26	2	1.03 (0.96, 1.11)	0.83 (0.76, 0.91)	4.97E-05
8q24 Region 3	rs943323	0.30	14, KIAA1409	1.01 (0.95, 1.08)	1.18 (1.09, 1.28)	9.60E-05
	rs4953347	0.48	2, EPAS1	0.97 (0.91, 1.04)	1.13 (1.06, 1.21)	9.69E-05
8q24 Region 4	rs11589338	0.10	1, TMCO4	0.94 (0.84, 1.05)	1.30 (1.15, 1.47)	1.57E-05

* Marginal results are based on single-SNP analyses. Chr=chromosome; OR=odds ratio; CI=confidence interval.

Table 3

Results from multivariate logistic regression analysis of SNPs in 8q24 Region 3, *ESRRB* and *SALL4*.

Effect*	OR (95% CI)	OR P-value
8q24 Region 3	1.17 (1.05, 1.30)	0.004
<i>ESRRB</i>	1.25 (1.06, 1.48)	0.009
<i>SALL4</i>	0.89 (0.79, 1.00)	0.04
8q24 Region 3: <i>ESRRB</i>	0.86 (0.76, 0.99)	0.03
8q24 Region 3: <i>SALL4</i>	1.11 (1.01, 1.21)	0.02

ESRRB and *SALL4* are transcription factors that serve as positive regulators and target genes of *POU5F1*. The regression model allowed for pair-wise interactions between each gene and 8q24 Region 3. An omnibus test on both interaction parameters produced a p-value of 0.007.

* Each effect is obtained from a single model that includes the main effect of three SNPs and interaction terms (indicated by colon) of 8q24 Region 3 with *ESRRB* and *SALL4*.

OR=odds ratio; CI=confidence interval.