Large-scale Exploration of Gene–Gene Interactions in Prostate Cancer Using a Multistage Genome-wide Association Study

Julia Ciampa1, Meredith Yeager2, Laufey Amundadottir2, Kevin Jacobs2, Peter Kraft3, Charles Chung2, Sholom Wacholder1, Kai Yu1, William Wheeler1, Michael J. Thun3, W. Ryan Divers5, Susan Gapstur5, Demetrius Albanes1, Jarmo Virtamo6, Stephanie Weinstein1, Edward Giovannucci4, Walter C. Willett4, Gilles Thomas1, Stephen Chanock2, and Nilanjan Chatterjee1

Abstract

Recent genome-wide association studies have identified independent susceptibility loci for prostate cancer 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 (SNP) across the genome to generate hypotheses about the functionality of prostate cancer susceptibility regions. We used data from Cancer Genetic Markers of Susceptibility: Stage I included 523,841 SNPs in 1,175 cases and 1,100 controls; Stage II included 27,383 SNPs in an additional 3,941 cases and 3,964 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 prostate cancer 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 prostate cancer 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 prostate cancer but also in other epithelial cancers that are associated with the 8q24 locus. Cancer Res; 71(9); 1–9. ©2011 AACR.

Introduction

Recent genome-wide association studies (GWAS) have identified multiple single nucleotide polymorphisms (SNP) associated with risk of prostate cancer (1–8). Functional variants in linkage disequilibrium (LD) with the established markers may directly contribute to prostate cancer risk through interactions with other, yet undetected, susceptibility loci. Large data sets 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 prostate cancer. In this report, we present the first results from a study of gene–gene interactions in the etiology of prostate cancer 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 prostate cancer risk, we conducted a series of conditional genome scans. The susceptibility (conditioning) regions included 9 gene regions and 4 independent regions within the chromosomal region 8q24. All have demonstrated strong associations with prostate cancer 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 (Fig. 1). Despite its strong associations with multiple
Table 1. Summary of 13 conditioning regions individually studied in conditional genome scans

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

NOTE: Reported values are based on single-SNP analyses of CGEMS Stages I and II combined. Abbreviations: Chr, chromosome; CI, confidence interval; MAF, minor allele frequency among controls in sample.

Materials and Methods

CGEMS Stage I

The details of CGEMS Stage I have been published previously (6). Briefly, the subjects available for this study included 1,175 cases and 1,100 controls of European ancestry from the Prostate, Lung, Colon, and Ovarian Screening Trial. They were genotyped using 2 Illumina chips (HumanHap300 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 3,941 cases and 3,964 controls of European ancestry. They represent 5 studies (case/control): Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study in Finland (929/921), Health Professionals Follow-up Study in the United States (596/611), American Cancer Society Cancer Prevention Study II Nutrition Cohort (1,760/1,775), 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 prostate cancer in CGEMS or other recent GWAS (Table 1). For each region, except EEFSSEC we used the SNP from the original publication. For EEFSSEC, 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 nonmultiplicative effect between the 2 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 3 alternative methods: (i) unconstrained maximum-likelihood (UML), (ii) constrained maximum-likelihood (CML), and (iii) empirical Bayes (EB). The UML method corresponds to the traditional prospective logistic regression analysis that...
obtains maximum-likelihood estimates of the OR 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 OR coefficients assuming gene–gene independence in the underlying population between the scan and conditioning SNPs (22). Although 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 500 Kb of the conditioning SNP to minimize gene–gene dependence due to physical proximity. For assessment and estimation of OR 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 that 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 2 key quantities: (i) 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 (ii) 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 CaseControlGenetics (26) that implements all 3 methods.

For all conditional regions, we ran 2 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 $P < 0.05$ level. To give higher priority for potential “cis” effects, we separately examined the significance of interaction for scan SNPs within 500 Kb 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 prostate cancer, 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-2α), they are positive regulators of POU5F1. We examined the top SNP from each gene based on Stage I main effects results (ESRRB $P = 0.001$, SALL4 $P = 0.28$). Both SNPs are located in an intron of their respective gene.

The most notable finding considering biologic plausibility was 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-2α), they are positive regulators of POU5F1. We examined the top SNP from each gene based on Stage I main effects results (ESRRB $P = 0.001$, SALL4 $P = 0.28$). 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 suggests 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, e.g., Fig. 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, which 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, 1 SNP, rs2002865, reached genome-wide significance for interaction after Bonferroni correction ($P = 9.14E-10$) 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 ($P = 3.79E-6$). Both interactions, however, failed to replicate ($P > 0.05$), when the SNP was followed up with further genotyping in a subset of the Stage II sample. Within the extended JAZF1 conditioning region, 1 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 ($P < 1.0E-4$) from each conditional scan is shown in Table 2. The most notable finding considering biologic plausibility was...
POU5F1B contains an interaction between rs6983267 in 8q24 Region 3, which for interaction analyses using case–control data. Interaction in the conditional scans for Region 2 (rs4314620 in the extended 8q24 region). It showed significant 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 = 0.0003), Region 3 (P = 0.0004), and Region 4 (P = 0.02; Fig. 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (P = 2.48E-5). The second result involves rs17714461 for the KLK2-KLK3 conditioning region: rs1558874 intronic to PRRX2 (P = 4.80E-5, multiplicative OR = 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 = 0.047, multiplicative OR = 1.27).

We highlight 2 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 = 0.0003), Region 3 (P = 0.0004), and Region 4 (P = 0.02; Fig. 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (P = 2.48E-5). The second result involves rs17714461 for the KLK2-KLK3 conditioning region: rs1558874 intronic to PRRX2 (P = 4.80E-5, multiplicative OR = 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 = 0.047, multiplicative OR = 1.27).

We highlight 2 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 = 0.0003), Region 3 (P = 0.0004), and Region 4 (P = 0.02; Fig. 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (P = 2.48E-5). The second result involves rs17714461 for the KLK2-KLK3 conditioning region: rs1558874 intronic to PRRX2 (P = 4.80E-5, multiplicative OR = 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 = 0.047, multiplicative OR = 1.27).

We highlight 2 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 = 0.0003), Region 3 (P = 0.0004), and Region 4 (P = 0.02; Fig. 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (P = 2.48E-5). The second result involves rs17714461 for the KLK2-KLK3 conditioning region: rs1558874 intronic to PRRX2 (P = 4.80E-5, multiplicative OR = 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 = 0.047, multiplicative OR = 1.27).

We highlight 2 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 = 0.0003), Region 3 (P = 0.0004), and Region 4 (P = 0.02; Fig. 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (P = 2.48E-5). The second result involves rs17714461 for the KLK2-KLK3 conditioning region: rs1558874 intronic to PRRX2 (P = 4.80E-5, multiplicative OR = 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 = 0.047, multiplicative OR = 1.27).

We highlight 2 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 = 0.0003), Region 3 (P = 0.0004), and Region 4 (P = 0.02; Fig. 1). An omnibus test for rs4314620 that includes its main effect and interaction with each of the 8q24 conditioning SNPs was highly significant (P = 2.48E-5). The second result involves rs17714461 for the KLK2-KLK3 conditioning region: rs1558874 intronic to PRRX2 (P = 4.80E-5, multiplicative OR = 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 = 0.047, multiplicative OR = 1.27).

Discussion

Our study presents one of the first large-scale explorations of gene–gene interactions in the setting of a multistage GWAS. Our analysis identifies a list (Table 2) of pair-wise SNP interactions that, through follow-up study, may elucidate the functional relevance of prostate cancer susceptibility SNPs. Our analysis also provides insights into future methodologic 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 POU5F1 are overexpressed in prostate cancer, but POU5F1 is not expressed in either healthy or malignant prostate tissue (13, 31). Given these data, we propose that POU5F1 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 and colleagues suggested the overexpression of POU5F1B in prostate cancer may mimic the ectopic expression of POU5F1 (13), which has been shown to promote epithelial tumors (35). That hypothesis aligns well with reports that prostate cancer 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 prostate cancer (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 prostate cancer association of 8q24 Region 3 involves a type of pluripotency network centered on \textit{POU5F1B} rather than on \textit{POU5F1}. Our follow-up analysis of 8q24 Region 3 with \textit{ESRRB} and \textit{SALL4} offers preliminary support of the hypothesis because those genes, which function as both regulators and target genes of \textit{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 \textit{EPAS1} and 8q24 Region 3 in independent studies. To obtain conclusive evidence, the sample size for those studies needs to be large due to the modest magnitude of the anticipated interaction. Even if our findings can be replicated, they 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 prostate cancer but rather all epithelial cancers associated with 8q24 Region 3 because, in subsets of those cancers, \textit{EPAS1} is overexpressed (31), mRNA

### Table 3. Results from multivariate logistic regression analysis of SNPs in 8q24 Region 3, \textit{ESRRB} and \textit{SALL4}

<table>
<thead>
<tr>
<th>Effect</th>
<th>OR (95% CI)</th>
<th>OR, P</th>
</tr>
</thead>
<tbody>
<tr>
<td>8q24 Region 3</td>
<td>1.17 (1.05, 1.30)</td>
<td>0.004</td>
</tr>
<tr>
<td>\textit{ESRRB}</td>
<td>1.25 (1.06, 1.48)</td>
<td>0.009</td>
</tr>
<tr>
<td>\textit{SALL4}</td>
<td>0.89 (0.79, 1.00)</td>
<td>0.04</td>
</tr>
<tr>
<td>8q24 Region 3:\textit{ESRRB}</td>
<td>0.86 (0.76, 0.99)</td>
<td>0.03</td>
</tr>
<tr>
<td>8q24 Region 3:\textit{SALL4}</td>
<td>1.11 (1.01, 1.21)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**NOTE:** \textit{ESRRB} and \textit{SALL4} are transcription factors that serve as positive regulators and target genes of \textit{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 3 SNPs and interaction terms (indicated by colon) of 8q24 Region 3 with \textit{ESRRB} and \textit{SALL4}.
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 2 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 prostate cancer in conjunction with KLK2 (39), and additional reports have linked KLK4 to various aspects of prostate cancer 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 was replicated in our relatively small, independent CGEMS Stage I analysis. These preliminary results suggest that the KLK2–KLK3 susceptibility region contributes to prostate cancer 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 3 independent known risk alleles for prostate cancer within the 8q24 region suggest that 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 subregion of 8q24 that is associated with bladder cancer (Fig. 1; ref. 44) and contains 2 regulatory regions for the oncogene MYC that flanks 8q24 Region 4 (45).

In our analysis of CGEMS Stage I data, 1 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 2,439 cases and 2,241 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 1,175 cases and 1,100 controls, was underpowered for study of interactions (Fig. 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 researchers to increase sample size while retaining the full array of SNPs.

Because of 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 ($\alpha = 1.0E^{-7}$ and $1.85E^{-6}$ for Stages I and II, respectively; Fig. 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 ORs in the scenarios we examined, which ranged from 1.13 to 2.05. The larger Stage II, in contrast, had high power for detecting modest to large interaction OR ($\geq 1.7$) 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 OR 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 OR and MAFs are reasonably large. In contrast, in the presence of qualitative interaction, the main effects of loci could be very weak or even nonexistent even when the interaction OR 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 nonmultiplicative or nonadditive 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 prostate cancer, will have a much more complex form than the pair-wise 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 multilocus 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 interactions.

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. This finding suggests that the functional relevance of the KLK2–KLK3 susceptibility region in the etiology prostate cancer may involve cellular proliferation.
Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

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

References

Interaction Analysis for Prostate Cancer with Modern Methods

Large-scale Exploration of Gene–Gene Interactions in Prostate Cancer Using a Multistage Genome-wide Association Study

Julia Ciampa, Meredith Yeager, Laufey Amundadottir, et al.

Cancer Res  Published OnlineFirst March 3, 2011.

Updated version  Access the most recent version of this article at:
do:10.1158/0008-5472.CAN-10-2646