Application of a Novel Score Test for Genetic Association Incorporating Gene-Gene Interaction Suggests Functionality for Prostate Cancer Susceptibility Regions

Julia Ciampa^{a,b} Meredith Yeager^c Kevin Jacobs^c Michael J. Thun^d
Susan Gapstur^d Demetrius Albanes^a Jarmo Virtamo^e Stephanie J. Weinstein^a
Edward Giovannucci^f Walter C. Willett^f Geraldine Cancel-Tassin^g
Olivier Cussenot^g Antoine Valeri^g David Hunter^c Robert Hoover^a
Gilles Thomas^h Stephen Chanockⁱ Chris Holmes^b Nilanjan Chatterjee^a

^aEpidemiology and Biostatistics Program, Division of Cancer Epidemiology and Genetics, Department of Health and Human Services, National Cancer Institute, National Institutes of Health, Bethesda, Md., USA; ^bDepartment of Statistics, University of Oxford, Oxford, UK; ^cCore Genotype Facility, SAIC-Frederick, Inc., National Cancer Institute – Frederick, Frederick, Md.; ^dDepartment of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Ga., USA; ^eDepartment of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Helsinki, Finland; ^fDepartment of Nutrition, Harvard School of Public Health, Boston, Mass., USA; ^gCeRePP Hopital Tenon, Assistance Publique-Hôpitaux de Paris, Paris; ^hSynergie-Lyon-Cancer, INSERM U590, Centre Leon Berard, Lyon, France; ⁱLaboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute – Frederick, Frederick, Md., USA

Key Words

Gene-gene interaction · Score test · Prostate cancer

Abstract

Aims: We introduce an innovative multilocus test for disease association. It is an extension of an existing score test that gains power over alternative methods by incorporating a parsimonious one-degree-of-freedom model for interaction. We use our method in applications designed to detect interactions that generate hypotheses about the functionality of prostate cancer (PRCA) susceptibility regions. **Methods:** Our proposed score test is designed to gain additional power through the use of a retrospective likelihood that exploits an assumption of independence between unlinked

loci in the underlying population. Its performance is validated through simulation. The method is used in conditional scans with data from stage II of the Cancer Genetic Markers of Susceptibility PRCA genome-wide association study. **Results:** Our proposed method increases power to detect susceptibility loci in diverse settings. It identified two high-ranking, biologically interesting interactions: (1) rs748120 of *NR2C2* and subregions of 8q24 that contain independent susceptibility loci specific to PRCA and (2) rs4810671 of *SULF2* and both *JAZF1* and *HNF1B* that are associated with PRCA and type 2 diabetes. **Conclusions:** Our score test is a promising multilocus tool for genetic epidemiology. The results of our applications suggest functionality for poorly understood PRCA susceptibility regions. They motivate replication study.

Copyright © 2011 S. Karger AG, Basel

Introduction

Data from recent genome-wide association studies (GWAS) provide a tremendous opportunity to understand how different genetic loci influence the risk of complex diseases through both their individual main effects and genetic interactions, known as epistasis. Although evidence of gene-gene interaction on human traits so far has been sparse [1, 2], both animal studies [3–8] and statistical investigations [9–12] suggest that genetic interactions can mask the importance of susceptibility loci when they are studied one at a time. The investigation of gene-gene interactions in GWAS can lead to an understanding of the complex biological process through which multiple susceptibility loci influence the pathogenesis of a disease.

In this article, we introduce the retrospective Tukey score test that assesses disease association of genetic loci in the presence of potential interaction with established susceptibility loci. The retrospective Tukey score test builds upon an existing test of genetic association that can incorporate interactions using a single parameter in the venerable Tukey model [13] and can avoid loss of power typically incurred by standard regression due to excess degrees of freedom. Tukey's model has been used successfully to assess interactions between sets of genetic variables when each set may represent a latent, underlying biological mechanism that affects disease risk [14]. The latent variable framework may be most natural for SNPs that are similar in terms of genomic location or biological function. For example, a hypothesized common mechanism for SNPs in genes of the same metabolic pathway could be reduced levels of the pathway's endproduct. In this regard, an advantage of Tukey's model is its ability to capture modular biology, a phenomenon in which complex biological functions result from connections between discrete units of diverse molecules that accomplish relatively autonomous tasks [15]. Modular biology has clear implications for epistasis, and studies in yeast suggest epistasis should be investigated on a system level between functional modules rather than between individual genetic markers [16, 17]. Sets of genetic variables that may be appropriate for Tukey's model include multiple loci from a single gene or multiple genes from a single pathway.

In this article, we consider the application of Tukey's model to discover new susceptibility loci for prostate cancer (PRCA) that could affect disease risk through interaction with established susceptibility SNPs. In one analysis, we explore SNPs in the genome that may interact glob-

ally with PRCA susceptibility SNPs in the gene-poor 8q24 region. In a second analysis, we search for SNPs that may simultaneously interact with two susceptibility regions for PRCA that are also associated with type 2 diabetes, which itself is associated with PRCA [18–23]. In both analyses, identification of modular interaction could generate hypotheses about the poorly understood functionality of existing susceptibility regions.

For case-control studies, it has long been recognized that the power of tests for interaction can be greatly enhanced by exploiting an assumption of independence between the factors in the underlying population. Under this assumption, for example, the case-only method has been proposed as a powerful alternative to standard logistic regression analysis for tests of interaction on the odds ratio scale when the disease is rare [24]. More generally, it has been shown that an assumption of independence in the underlying population can be exploited in a case-control study through a 'retrospective' likelihood of the data that properly accounts for the underlying design [25, 26].

In this report, we consider developing a score test for genetic association under Tukey's model for gene-gene interaction using the retrospective likelihood so that power can be gained by exploiting an assumption of gene-gene independence between physically distant loci. Previously, an analogous 'prospective' Tukey score test under an unconstrained likelihood was developed that analyzes case-control data as if they were collected using a prospective design, a practice routinely applied in epidemiologic studies [14]. Although the classical theory [27] for case-control studies suggests that the use of a prospective likelihood is valid, recent investigations [25, 28–30] suggest that the use of prospective likelihood can be inefficient if certain assumptions can be made about the distribution of risk factors in the underlying population. In this article, we develop the retrospective Tukey score test that exploits the power advantage of both the parsimonious Tukey model and a retrospective likelihood, thereby offering a more powerful alternative to existing methods.

This paper has the following structure: in the next section, we present fundamentals of the derivation of the retrospective Tukey score test. We also describe the design of our simulation studies and PRCA applications. We present the results with an emphasis on the applied project that offers original insight into PRCA etiology. We discuss the results of our applications in the context of existing literature on PRCA pathogenesis. We also address the merits of the retrospective Tukey score test with suggestions for future work.

Methods

Let $\mathbf{S} = (S_1, \dots, S_M)$ represent the minor allele count for a set of independent 'scan' SNPs for which one desires to test the association with a binary disease outcome D , and let $\mathbf{C} = (C_1, \dots, C_Z)$ represent the risk allele count for a set of 'conditioning' SNPs that are associated with the disease outcome and may interact with the SNP set $\mathbf{S}$, with $M \geq 1$ and $Z \geq 1$ for $M + Z > 2$. Each SNP set is grouped by a common known or hypothesized biological function, such as a candidate pathway for $\mathbf{S}$ and a susceptibility region for $\mathbf{C}$. The assumption that $\mathbf{C}$ contains established susceptibility SNPs implies that preliminary analyses are required before an analysis (i.e. genome scan) can be performed using the Tukey score test. For example, the results of single-SNP GWAS analyses can influence the choice of $\mathbf{C}$ in conjunction with relevant literature reviews on biological function. In addition, we assume the analysis needs to be adjusted for a set of covariates $\mathbf{A} = (A_1, \dots, A_W)$, such as age, sex, ethnicity and population stratification eigenvalues. We assume disease risk can be described as a function of $\mathbf{S}$, $\mathbf{C}$ and $\mathbf{A}$ using 'Tukey's' model:

$$\text{logit}[P(D=1|\mathbf{S}, \mathbf{C}, \mathbf{A})] = \beta_0 + \sum_{m=1}^M \beta_{1m} S_m + \sum_{z=1}^Z \beta_{2z} C_z + \theta \left(\sum_{m=1}^M \beta_{3m} S_m \sum_{z=1}^Z \beta_{2z} C_z \right) + \sum_{w=1}^W \beta_{3w} A_w \quad (1)$$

The above model includes main effect terms for each of the individual factors for $\mathbf{S}$, $\mathbf{C}$ and $\mathbf{A}$, analogous to its standard logistic counterpart. However, the interaction between the scan and conditioning SNPs is captured by a single interaction parameter θ . Under the above model, the null hypothesis of no association between the disease and scan SNPs is equivalent to $\beta_1 = (\beta_{11}, \dots, \beta_{1M}) = 0$, under which both the main effects of the scan SNPs and their interaction with conditioning SNPs disappear. The parameter β_1 can be used to perform a global test of association that can capture both main effects and interaction, as in omnibus testing [11].

We make two assumptions of gene-gene independence in the underlying population, which is assumed to be represented by controls for rare diseases, such as cancer. We assume gene-gene independence between the scan and conditioning SNPs in an attempt to gain power in a fashion similar to case-only studies of epistasis. We assume independence among scan SNPs in a non-singular $\mathbf{S}$ in an attempt to limit the sample size requirement of our score test. The method, in principle, can be extended to account for dependence via estimation of joint genotype frequencies for correlated scan SNPs. The sample size needed for standard asymptotic theory to be valid in such a test may be large due to the requirement of sufficient sample size in each multivariate genotype frequency category. We advise that if correlated scan SNPs are to be included, then one may need to avoid sparsity by collapsing multivariate genotype categories appropriately. These gene-gene independence assumptions are natural for unlinked loci. We also assume independence for $\mathbf{S}$ and $\mathbf{A}$ at the outset. We use Bayes' theorem to express the standard prospective likelihood in a retrospective framework. The resultant retrospective likelihood that incorporates these constraints for Tukey's model with N total subjects is:

$$l = \prod_{i=1}^N \frac{P(D|\mathbf{S} = \mathbf{s}_i, \mathbf{C} = \mathbf{c}_i, \mathbf{A} = \mathbf{a}_i) P(\mathbf{C} = \mathbf{c}_i, \mathbf{A} = \mathbf{a}_i) \prod_{m=1}^M P(S_m = s_{mi})}{P(D)} \quad (2)$$

Note that equation 1 can be used to solve for conditional and marginal probabilities on $\mathbf{D}$. To guard against population stratification that induces dependence between $\mathbf{S}$ and $\mathbf{C}$, we introduce a stratification variable $\mathbf{U}$ and replace $P(S_m = s_{mi})$ with $P(S_m = s_{mi} | \mathbf{U} = \mathbf{u}_i)$ in equation 2 [25, 26]. One may select proxies for population strata, such as ethnicity or study country, for $\mathbf{U}$. In that way, $\mathbf{U}$ may also be a subset of $\mathbf{A}$. We set a scalar U to be a binary stratification variable in our derivation. In practice, more complex stratification is possible. For example, a trichotomous regression allows estimation of $P(\mathbf{S} | \mathbf{U})$ for continuous stratification variables.

It may be infeasible to maximize the corresponding log-likelihood of equation 2 when there is a high dimensional nuisance parameter $\delta = \{P(\mathbf{C} = \mathbf{c}, \mathbf{A} = \mathbf{a})\}$. It has been shown in the case of a standard logistic regression model that the log-likelihood can be re-parameterized and retain proper maximum likelihood estimates. Let $\mu = \{\mu_d = P(R = 1 | D = d); d = 0, 1\}$, which can be computed explicitly when disease prevalence is known for $P(R = 1)$, be the probability of inclusion in the study [14]. The resulting profile log-likelihood in this setting is:

$$L = \sum_{dsc} n_{dsc} \left[h_{ds} - \ln \left(\sum_{d's'} \exp(h_{d's'}) \right) \right]$$

where n_{dsc} = number of subjects with $D = d$, $\mathbf{S} = \mathbf{s}$, $\mathbf{C} = \mathbf{c}$ and

$$h_{ds} = d \left[\ln \left(\frac{\mu_1}{\mu_0} \right) + \beta_0 + m(\mathbf{s}, \mathbf{c}, \mathbf{a}; \beta_1, \beta_2, \beta_3, \theta) \right] + \sum_{m=1}^M \ln \left(\frac{q_{jmu}}{q_{0mu}} \right)$$

with $m(\mathbf{x}; \boldsymbol{\gamma})$ = an arbitrary function of the covariates ($\mathbf{x}$) and parameters ($\boldsymbol{\gamma}$) that define the regression model under study and

$$q_{jmu} = \{f = f_{jmu} = P(S_m = j | U = u), f_{0mu} = 1 - (f_{1mu} + f_{2mu}); j = 1, 2; u = 0, 1\},$$

as Hardy-Weinberg equilibrium is not assumed. The score function for a given θ is:

$$S(\boldsymbol{\eta}) = \sum_{i=1}^N \left[\frac{\partial h_{ds}}{\partial \boldsymbol{\eta}} - E \left[\frac{\partial h_{ds}}{\partial \boldsymbol{\eta}} | \mathbf{C} = \mathbf{c}_i, \mathbf{A} = \mathbf{a}_i, R = 1 \right] \right]$$

with

$$\boldsymbol{\eta} = \{\beta_0, \beta_1, \beta_2, \beta_3, \theta\}.$$

Under the null hypothesis, we define

$$S_{\beta_1}(\boldsymbol{\eta}_0) = \sum_{i=1}^N \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi} \right) (\mathbf{s}_i d_i - E(\mathbf{S} | \mathbf{C}, U, R) p_i)$$

where

$$p_i = E[D | \mathbf{C} = \mathbf{c}_i, R = 1] \text{ and } E(S_m | \mathbf{C}, U, R) = I_{U_i = u} (f_{1mu} + 2f_{2mu}).$$

The score function differs from that of the prospective Tukey score test through the substitution of the observed for expected value of $\mathbf{S}$ in the final term. The asymptotic variance of score function $\mathbf{I}^{\beta_1 \beta_1}$ can be estimated through the information matrix, as detailed in Appendix I.

We define the retrospective Tukey score test statistic for a given value of θ as:

$$T(\theta) = S_{\beta_1}^T(\boldsymbol{\eta}_0) \mathbf{I}^{\beta_1 \beta_1}(\boldsymbol{\eta}_0) S_{\beta_1}(\boldsymbol{\eta}_0)$$

where

$$\mathbf{I}^{\beta_1 \beta_1}(\boldsymbol{\eta}_0) = [\mathbf{I}_{\beta_1 \beta_1}(\boldsymbol{\eta}_0) - \mathbf{I}_{\beta_1 \psi}(\boldsymbol{\eta}_0) [\mathbf{I}_{\psi \psi}(\boldsymbol{\eta}_0)]^{-1} \mathbf{I}_{\psi \beta_1}(\boldsymbol{\eta}_0)]^{-1}.$$

Table 1. Summary of conditioning SNPs used in two genome scans of CGEMS stage II data

Scan	Conditioning SNP (gene)	Frequency ^a (allele)	OR	p value
			per allele	
8q24	rs17446916	0.43 (T)	1.07	0.04
Region P	rs6999921	0.08 (G)	1.15	0.02
	rs1447293	0.37 (G)	1.14	1.54 ⁻⁴
	rs921146	0.22 (C)	1.21	2.13 ⁻⁶
	rs13253127	0.46 (T)	1.14	1.22 ⁻⁴
	rs4242382	0.10 (A)	1.47	3.74 ⁻¹³
	rs6991990	0.66 (C)	1.15	1.02 ⁻⁴
Type 2 diabetes	rs4430796 ^b (<i>HNF1B</i>)	0.52 (A)	1.26	1.85 ⁻¹¹
	rs1164973 (<i>HNF1B</i>)	0.81 (G)	1.18	2.70 ⁻⁴
	rs10486567 (<i>JAZF1</i>)	0.75 (G)	1.24	2.21 ⁻⁷

^a Risk allele frequency is based on study controls. ^b SNP allele is protective for type 2 diabetes mellitus. OR = Odds ratio.

In practice, maximum likelihood estimates are computed for $\{\beta_0, \beta_2, \beta_3\}$ using the null hypothesis model and for $\mathbf{f}$ using all subjects' data. We base inference on disease association of the scan SNPs on $T = \max_{\theta} T(\theta)$ for a prespecified range of θ (-5 to 5 on a 0.2 grid), as has been advocated in similar settings [9, 14].

Cancer Genetic Markers of Susceptibility Application

Details of the multi-stage Cancer Genetic Markers of Susceptibility (CGEMS) GWAS have been published elsewhere [20, 31, 32]. Our analyses involve only stage II data. They include 3,941 cases and 3,964 controls, representing distinct individuals from stage I, and 27,383 autosomal SNPs that met quality control metrics after the selection for promising disease association in the stage I single-SNP analyses ($p < 0.05$). The subjects represent four international studies (case/control): the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study in Finland (ATBC, 929/921), the Health Professionals Follow-Up Study in America (HPFS, 596/611), the American Cancer Society Cancer Prevention Study II Nutrition Cohort (ACS, 1,760/1,775) and the CeRePP French Prostate Case-Control Study (FPCC, 656/657).

Conditional Scans Involving 8q24, JAZF1 and HNF1B SNPs

Our CGEMS application involved two conditional genome scans for separate sets of conditioning SNPs (table 1). In each conditional scan, we tested individual scan SNPs for disease association, allowing for interaction with the specified set of conditioning SNPs. The scan SNPs included all CGEMS stage II SNPs at least 500 kb from any of the conditioning SNPs to minimize violations of the gene-gene independence assumption. It is an accepted practice to assume SNPs that distant in the genome are independent. However, there is some evidence suggesting rare instances of long-range SNP-SNP correlation, representing a very small subset of the total set of SNPs [33, 34]. Approximately 27,000 SNPs remained, setting a Bonferroni threshold of

$0.05/27,000 = 1.83 \times 10^{-6}$ for declaring genome-wide significance. The first set of conditioning SNPs included seven loci from the three subregions of 8q24 that are associated with only PRCA [35] and will jointly be referred to as Region P (prostate) (fig. 1). The SNP set included the most significant marker in the region, rs4242382, and six SNPs in low linkage disequilibrium (LD) with it to maximize coverage of tagged SNPs and minimize multi-collinearity in the null hypothesis model. We assumed the Region P SNPs represent a single biological mechanism that affects PRCA risk in contrast to SNPs residing in other 8q24 subregions that are associated with different cancers. The second set of conditioning SNPs included two independent intronic SNPs in *HNF1B* [19] and one in *JAZF1* [20] that have been conclusively established as susceptibility loci for PRCA. SNPs in these genes have also been associated with type 2 diabetes [18, 19], which is itself associated with PRCA [21], suggesting our conditioning SNPs share a causal link to risk of PRCA that could include one of the susceptibility pathways for type 2 diabetes. Our regression models included study identifiers as adjusting covariates. To avoid gene-gene dependence that may arise due to population stratification, we additionally stratified the analysis for ATBC participation because the minor allele frequency (MAF) of conditioning SNPs substantially differed among controls relative to all others, which is not surprising because the ATBC study was conducted in Finland. Noteworthy results were followed-up using the traditional logistic regression model to explore individual interactions between specific combinations of conditioning and scan SNPs. We examined a 'saturated' model that included main effects of the scan and conditioning SNPs and all pair-wise interactions between the scan and conditioning SNPs. We analyzed this model using both standard methods and a retrospective analysis that imposed gene-gene independence. The omnibus tests that jointly assess main and interaction effects of all scan SNPs in the constrained analysis most closely resemble the retrospective Tukey score test of general disease association. We computed joint odds ratios to further characterize the modular interactions we detected.

Simulation Study

We conducted simulations to evaluate the size and power of the retrospective Tukey score test. We incorporated CGEMS data into the simulations so that the set of conditioning SNPs would retain the LD pattern and disease association of known susceptibility regions. In each simulation, we randomly selected an equal number of cases and controls, recording their disease status and genotypes for eight 8q24 SNPs selected to minimize r^2 with the two most significant SNPs in distinct LD blocks, rs4242382 and rs6983267, and maximize coverage of all subregions (fig. 1). We assumed no adjusting covariates and simulated a single scan SNP. The scan SNP data were generated for all subjects under the null hypothesis, assuming Hardy-Weinberg equilibrium with a MAF of 0.12. The same procedure was used for controls under the alternative hypothesis, except MAF was set to 0.15 and $\beta_1 = \ln(1.15)$. The scan SNP data for cases under the alternative hypothesis were randomly sampled from a multinomial distribution with probabilities that accounted for the relationship of $\mathbf{S}$ and $\mathbf{C}$ in the Tukey model:

Fig. 1. LD plot for 24 SNPs in PRCA susceptibility region 8q24 genotyped in CGEMS stage II. The seven boxed loci represent conditioning SNPs in the 8q24 Region P genome scan. The larger outlined LD block that contains five conditioning SNPs contains three binding sites for the androgen receptor [42]. The smaller outlined LD block is associated with several epithelial cancers,

including prostate and colon [35]. This LD plot spans 354,599 base pairs. The color scheme reflects LD with white corresponding to low r^2 and black to high. Pairwise r^2 based on 3,887 study controls are written as percentages in the respective blocks. Black squares identify the conditioning SNPs used in joint odd ratios calculations (fig. 3).

$$P(S=s|\mathbf{C}, \mathbf{A}, U, D=1) = \frac{OR_D(S=s|\mathbf{C}, \mathbf{A})P(S=s|U)}{\sum_{s=0}^2 OR_D(S=s|\mathbf{C}, \mathbf{A})P(S=s|U)}; s=0,1,2$$

where

$$OR_D(S=s|\mathbf{C}, \mathbf{A}) = \exp \left\{ s\beta_1 \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_z \right) \right\} \quad (3)$$

In each simulation, we generated 1,000 test statistics and computed corresponding p values in three ways. Standard calculations did not apply because the retrospective Tukey score test statistic is a function of unknown parameters estimated from all subjects' data. The first method was a χ^2 approximation that was suggested through the empirical study of another omnibus-type testing procedure [9]. The second was a permutation method in which scan SNP genotypes were shuffled among sub-

jects to preserve the disease association of the conditioning SNPs. The third method is based on an asymptotic approximation introduced by Lin and Zou [65] that is computationally less intensive than permutation. The latter two methods were used previously for the prospective Tukey score test to overcome the challenge that a null distribution of the Tukey score test statistics is unknown [14].

Empirical alpha and power calculations were calculated for an alpha level of 0.01. Power was compared to the prospective Tukey score test. It was assessed over a range of θ for a total sample size of 1,000 so that the relative contribution of the epistatic effect to the retrospective Tukey score test signal could range from non-existent ($\theta=0$) to dominant. Power was also assessed over a range of total sample sizes with equal numbers of cases and controls under moderate epistasis ($\theta=1.2$). The robustness of the retrospective Tukey score test to model misspecification with respect to the interaction term was assessed through simulations that

Fig. 2. Quantile-quantile plots of χ^2_2 approximate p values for retrospective Tukey score test statistics for the 8q24 Region P (a) and type 2 diabetes (b) genome scans of ~27,000 SNPs.

Fig. 3. Empirical joint odds ratios (ORs). Joint ORs were computed for the most promising SNP in each CGEMS analysis (a 8q24 Region P; b type 2 diabetes). The ORs are calculated based on the minor allele count of the putative susceptibility SNP and the total risk allele count of the established susceptibility (conditioning) SNPs that demonstrated interactions ($p < 0.05$) with it in logistic models saturated with pair-wise interactions. The conditioning SNPs for the 8q24 Region P application were rs4242382 and rs6991990 and for the diabetes application they were rs4430796 (*HNF1B*) and rs10486567 (*JAZF1*). Values are given for ORs significant at the 5% level. Absent values indicate an empty cell. The MAF of rs748120 was 0.23 and of rs4810671 it was 0.48.

generated scan SNP data for cases under the alternative hypothesis based on varied subsets of conditioning SNPs in the summation term of equation 3. The figure legends provide further details on parameter values used in simulations.

Results

Quantile-quantile plots of the p values from the CGEMS application are given in figure 2. The top hits represent established susceptibility regions that were detected through single-SNP tests, reflecting the omnibus feature of the retrospective Tukey score test. Further analysis focused on high-ranking SNPs that have not been reported previously in PRCA GWAS. The 8q24 Region P application identified one biologically interesting SNP: rs748120 of *NR2C2* (also known as *TR4*, $p = 4.98 \times 10^{-5}$, MAF = 0.23). It ranked 7th overall and 1st amongst 'novel' susceptibility SNPs, in contrast to its rank of 191 in standard single-SNP analyses. The 'diabetes' application also identified a biologically interesting SNP: rs4810671 of *SULF2* ($p = 4.84 \times 10^{-5}$, MAF = 0.48). It ranked 23rd overall and 4th amongst novel susceptibility SNPs, in contrast to its rank of 252 in standard single-SNP analyses. These SNPs remain high-ranking but have attenuated association signals in a prospective analysis via the Tukey score test (rs748120: rank 157, $p = 4.45 \times 10^{-3}$; rs4810671: rank 105, $p = 2.34 \times 10^{-3}$).

More traditional methods support our results and provide further information on the nature of interactions we detected. First, we considered a standard logistic model saturated with main effects for all scan and conditioning SNPs, interactions between the scan SNP and

Table 2. Results of regression that involved rs748120 and seven susceptibility SNPs from the gene-poor 8q24 Region P

Covariate	Main effect term		Interaction with rs748120 term	
	OR (95% CI)	OR p value	OR (95% CI)	OR p value
rs748120 (<i>NR2C2</i>)	0.99 (0.84, 1.16)	0.90		
rs17446916	1.08 (1.01, 1.16)	0.03	0.98 (0.90, 1.06)	0.58
rs6999921	1.17 (1.04, 1.33)	0.01	1.02 (0.89, 1.16)	0.82
rs1447293	1.01 (0.91, 1.12)	0.81	1.11 (0.99, 1.24)	0.07
rs921146	0.90 (0.77, 1.05)	0.18	1.06 (0.90, 1.24)	0.51
rs13253127	0.98 (0.88, 1.09)	0.67	1.05 (0.94, 1.18)	0.39
rs4242382	1.70 (1.48, 1.96)	2.56 ⁻¹³	0.71 (0.61, 0.84)	2.29 ⁻⁵
rs6991990	1.13 (1.03, 1.25)	0.01	0.89 (0.80, 0.99)	0.03

An omnibus test on all rs748120 parameters produced a p value of 2.67⁻⁵. The regression model also adjusted for and stratified by study. OR = Odds ratio; CI = confidence interval.

Table 3. Results of regression that involved rs4810671 and susceptibility SNPs of *HNF1B* and *JAZF1*

Covariate	Main effect term		Interaction with rs4810671 term	
	OR (95% CI)	OR p value	OR (95% CI)	OR p value
rs4810671 (<i>SULF2</i>)	0.75 (0.56, 1.00)	0.053		
rs4430796 (<i>HNF1B</i>)	1.20 (1.11, 1.29)	2.45 ⁻⁶	1.13 (1.03, 1.23)	0.010
rs11649743 (<i>HNF1B</i>)	1.14 (1.04, 1.26)	0.007	1.00 (0.88, 1.13)	0.974
rs10486567 (<i>JAZF1</i>)	1.19 (1.09, 1.30)	8.30 ⁻⁵	1.14 (1.02, 1.28)	0.019

An omnibus test on all rs4810671 parameters produced a p value of 4.32⁻⁴. The regression model also adjusted for and stratified by study. OR = Odds ratio; CI = confidence interval.

each conditioning SNP, and adjusting covariates. We analyzed these models assuming gene-gene independence, and we performed omnibus tests of all parameters involving the scan SNP to resemble the general disease association assessed by the retrospective Tukey score test. Omnibus tests for analyses that assumed gene-gene independence support our findings for both rs748120 ($p = 2.67^{-5}$) and rs4810671 ($p = 4.32^{-4}$), as do those from unconstrained analyses. In the respective saturated logistic model, these scan SNPs demonstrate interaction ($p < 0.05$) of similar magnitude with two conditioning SNPs (tables 2, 3). Empirical joint odds ratios were calculated based on the minor allele counts of the scan SNP and the total risk allele counts of these conditioning SNPs (fig. 3).

Simulation studies reinforce the merits of the retrospective Tukey score test. It maintains type I error when the independence assumption is valid (table 4), using all p value computation methods we considered. Due to its

Table 4. Empirical alpha level for the retrospective Tukey score test at two nominal alpha levels for 1,000 simulations, using three methods to compute p values

	Nominal alpha	
	0.05	0.01
χ^2_2 approximation	0.050	0.010
Permutation	0.051	0.008
Asymptotics	0.041	0.009

ease of implementation, we used χ^2_2 approximate p values (fig. 4) in both our power calculations and the CGEMS application. Our power analysis demonstrates that the retrospective Tukey score test gains power over the prospective Tukey score test when independence holds.

Fig. 4. χ^2 quantile-quantile plot of 1,000 test statistics generated under the null hypothesis for the retrospective Tukey score test.

These gains are maintained over a range of sample sizes likely to be observed in GWAS (fig. 5a). They are more pronounced as the epistatic effect of the scan SNP increases, as represented by an increasing θ value (fig. 5b). This result is expected because retrospective analyses of logistic models gain power over prospective analyses primarily through assessment of interaction rather than main effect terms. Both Tukey score tests are robust to modest misspecifications of the conditioning SNP set (i.e. simulated interaction excludes less than half the conditioning SNPs) (fig. 5c, unpubl. data). This feature is advantageous because a priori information can be limited for the selection of conditioning SNPs.

Discussion

Our simulation studies suggest that the retrospective Tukey score test is a promising analytic tool for case-control studies that allows for interaction with known risk factors in its test of disease association. Tukey's model incorporates patterns of epistasis that are likely to contribute to complex human disease, offering a more flexible alternative to single-SNP models and a more parsimonious alternative to logistic models saturated with interactions. Its test of general disease association offers a robust alternative to either pure main effects or interaction tests. In an analysis of the CGEMS dataset, we identified biologically interesting SNPs that suggest functionality for established PRCA susceptibility regions, motivating replication studies.

Fig. 5. Comparison of empirical power in a variety of settings. Each power calculation was based on 1,000 simulations. In all simulations the scan SNP MAF = 0.15 and $\beta_1 = \ln(1.15)$. **a** Power was compared between retrospective (RTS) and prospective (PTS) Tukey score tests over a range of total sample sizes (with equal numbers of cases and controls) for $\theta = 3.5$ and a nominal alpha level of 0.01. **b** Power was compared between the RTS and the PTS over a range of θ for a nominal alpha level of 0.05, using random samples of 500 cases and 500 controls. **c** Power was considered for the RTS when subsets of the eight conditioning SNPs contributed to the epistatic element of the Tukey model that generated the data for the scan SNP. Each simulation included a random sample of 500 cases and 500 controls and set nominal alpha to 0.05.

Simulations indicate that the retrospective Tukey score test gains power over the prospective Tukey score test. The gene-gene independence assumption that constrains the likelihood function improves efficiency by reducing variance in the score function. This reduction is due to substitution of observed for expected values of $\mathbf{S}$, which is different from the decrease in variance of point estimates on interaction parameters in case-only versus case-control designs [36]. The increased efficiency is contingent on a valid gene-gene independence constraint. The retrospective Tukey score test is likely to be sensitive to violations of its independence constraints similar to other retrospective logistic analyses [37, 38]. A number of recent reports have suggested methods [33, 38–40] that can protect against bias in retrospective methods when the underlying assumption of independence is violated, for example, due to population stratification. Future work on the retrospective Tukey score test could focus on relaxing the independence assumption. At present, we advise that analyses exclude scan SNPs within 500 kb of each other for non-singular $\mathbf{S}$ and of conditioning SNPs in order to minimize the effect of residual LD. Alternatively, the gene-gene independence could be investigated using publicly available external data. We also advise that χ^2_2 approximate p values be used for retrospective Tukey score test statistics for analysis of a single scan SNP. We caution against deterioration of this approximation for larger sets of scan SNPs (unpubl. data) and advise that researchers compute p values using a recently proposed, efficient permutation algorithm that over-samples the tails of non-standard test statistics [41].

Our conditional scan with 8q24 Region P highlights rs748120 of *NR2C2*. The retrospective Tukey score test signal appears to be due primarily to an interaction since the rank is much higher than in main effects analysis. Multiple 8q24 SNPs demonstrate evidence of interaction ($p < 0.10$) with rs748120 in our analysis of a standard logistic model saturated with pair-wise interactions. These SNPs reside in the same subregion of Region P that has been reported to contain androgen receptor (AR) binding sites and enhancer elements responsive to androgens [42]. Two loci demonstrate epistasis ($p < 0.05$) with rs748120: rs4242382 and rs6991990. They are independent with $r^2 = 0.04$ in CGEMS stage II controls, suggesting there is a region-wide interaction with rs748120. The conditioning SNP rs4242382 is in complete LD with rs11986220 [43], which is a PRCA susceptibility SNP in Region P not genotyped in CGEMS. The rs11986220 risk allele is associated with both increased androgen-dependent enhanc-

er activity and altered AR binding in Region P [42]. The empirical joint odds ratios that incorporate the conditioning SNPs rs4242382 and rs6991990 suggest that the latent causal mechanism of 8q24 Region P acts almost exclusively in the absence of minor alleles for rs748120 of *NR2C2*.

NR2C2 is an AR co-regulator. Its protein can form a complex with AR that decreases expression of both their target genes [44]. One target gene of *NR2C2* is *NANOG*, which is one of the most important genes in stem cell pluripotency regulation [45]. This relationship is particularly noteworthy because PRCA pathogenesis is thought to involve the reactivation of embryonic pathways [46]. *NR2C2* has also been shown to reduce synthesis of vitamin D [47] that influences cellular differentiation and proliferation in the prostate [48, 49]. Observational studies suggest that vitamin D deficiency may be a risk factor for PRCA [50]. Accordingly, impaired AR-*NR2C2* binding could correspond to increased *NR2C2* activity that activates *NANOG* expression and represses vitamin D synthesis at levels outside the physiologic range. Both sequelae could increase risk of PRCA, which is in line with previous reports. We hypothesize that impaired AR-*NR2C2* binding contributes to the increased androgen responsiveness of 8q24 Region P in the presence of a risk allele at rs11986220 (or rs4242382 by complete LD).

Our second scan involves conditioning SNPs from *HNF1B* and *JAZF1*, selected to represent a type 2 diabetes phenotype. The analysis highlights rs4810671 of *SULF2*. This signal also appears to involve a relatively large epistatic effect, given the comparative single-SNP analysis rank. One SNP from each susceptibility gene region demonstrates interaction in our analysis of a standard logistic model saturated with scan SNP interactions: rs4430796 of *HNF1B* and rs10486567 of *JAZF1*. The empirical joint odds ratios that incorporate these conditioning SNPs illustrate that rs4810671 affects PRCA risk only when at least three of the conditioning SNP risk alleles are present.

SULF2 has demonstrated oncogenic properties in several cancers, including cancer of the pancreas, breast, lung and liver [51–54]. It encodes an enzyme that modifies the sulfation and, thereby, function of heparin sulfate proteoglycans [55]. Abnormalities in this family of sugars and particularly in HSPG2 (Perlecan) contribute to the disease progression of both type 2 diabetes [56] and PRCA [57, 58]. Heparin sulfate proteoglycans influence the FGF axis and WNT signaling, both of which are disrupted in PRCA [59–61]. A positive regula-

tor of *SULF2* is insulin [62], the hormone at the center of diabetes. Given the increasing evidence that cellular response to insulin mediates the association between type 2 diabetes and PRCA [21], we hypothesize that our observed interaction suggests that abnormal insulin function affects PRCA risk through dysregulation of *SULF2*.

In conclusion, we have introduced an innovative score test for disease association that simulation studies suggest is a promising multilocus tool for genetic epidemiol-

ogy. While we present the retrospective Tukey score test in the context of gene-gene interactions, it can also be used to investigate gene-environment interactions, provided the covariates are chosen with model assumptions in mind. Our GWAS applications of this score test identified interactions that warrant replication study. They suggest mechanisms, consistent with existing literature, for well-established but poorly understood PRCA susceptibility regions.

Appendix I

The asymptotic representation of the information matrix is used to compute the retrospective Tukey score test statistics:

$$\mathbf{I}^{\beta_1\beta_1}(\boldsymbol{\eta}_0) = [\mathbf{I}_{\beta_1\beta_1}(\boldsymbol{\eta}_0) - \mathbf{I}_{\beta_1\psi}(\boldsymbol{\eta}_0) [\mathbf{I}_{\psi\psi}(\boldsymbol{\eta}_0)]^{-1} \mathbf{I}_{\psi\beta_1}(\boldsymbol{\eta}_0)]^{-1}.$$

The information matrix is the inverse of the variance on the score function with general form:

$$\mathbf{I}(\boldsymbol{\eta}) = \sum_{i=1}^N \text{Var} \left(\frac{\partial h_{ds}}{\partial \boldsymbol{\eta}} | \mathbf{C}, \mathbf{A}, R \right).$$

Under the null hypothesis, its components can be written as follows.

$$\begin{aligned} \mathbf{I}_{\beta_1\beta_1}(\boldsymbol{\eta}_0) &= \sum_{i=1}^N \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi} \right)^2 \left(p_i E(\mathbf{S}^T \mathbf{S} | \mathbf{C}, \mathbf{A}, U, R) - p_i^2 E(\mathbf{S}^T | \mathbf{C}, \mathbf{A}, U, R) E(\mathbf{S} | \mathbf{C}, \mathbf{A}, U, R) \right) \\ \mathbf{I}_{\psi\beta_1}(\boldsymbol{\eta}_0) &= \sum_{i=1}^N \begin{pmatrix} p_i (1 - p_i) E(\mathbf{S} | \mathbf{C}, \mathbf{A}, U, R) \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi} \right) \\ p_i (1 - p_i) (\mathbf{c}_i^T, \mathbf{a}_i^T) E(\mathbf{S} | \mathbf{C}, \mathbf{A}, U, R) \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi} \right) \\ I_{U=u_i} p_i \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi} \right)_{2M \times M} \\ 2 * I_{U=u_i} p_i \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi} \right)_{2M \times M} \end{pmatrix} \\ \mathbf{I}_{\psi\psi}(\boldsymbol{\eta}_0) &= \sum_{i=1}^N \begin{pmatrix} (1, \mathbf{c}_i^T, \mathbf{a}_i^T) (1, \mathbf{c}_i^T, \mathbf{a}_i^T)^T p_i (1 - p_i) & \mathbf{0}_{(Z+W+1) \times 4M} \\ \mathbf{0}_{4M \times (Z+W+1)} & \begin{bmatrix} \left(\frac{1}{f_{0mu}} + \frac{1}{f_{1mu}} \right)_{2M \times 2M} & \left(\frac{1}{f_{0mu}} \right)_{2M \times 2M} \\ \left(\frac{1}{f_{0mu}} \right)_{2M \times 2M} & \left(\frac{1}{f_{0mu}} + \frac{1}{f_{2mu}} \right)_{2M \times 2M} \end{bmatrix} \end{pmatrix} \end{aligned}$$

In these matrices, p_i is the expected value of d_i under the null hypothesis model and the entries involving $\mathbf{f}$ that differ only in m or u subscripts are given by a general formula with appropriate matrix dimensions.

We find that this analytic approach to the construction of an information matrix is appropriate for analyses of singular scan SNP sets. However, the construction of the information matrix for analyses of non-singular $\mathbf{S}$

requires an empirical variance-covariance matrix for the individual score functions. This difference arises because score tests that use observed (rather than expected) information matrices can yield negative test statistics [63, 64]. The contribution of each subject to the score function evaluated under the null hypothesis is:

$$S_i(\boldsymbol{\eta}_0) = \begin{pmatrix} (\mathbf{1}, \mathbf{c}_i^T, \mathbf{a}_i^T)(d_i - p_i) \\ \left(1 + \theta \sum_{z=1}^Z \beta_{2z} c_{zi}\right) (s_i d_i - E(\mathbf{S}|\mathbf{C}, \mathbf{A}, U, R) p_i) \\ \frac{I_{s_{mi} \neq 0} - f_{1mu} - f_{2mu}}{f_{0mu}} + \frac{I_{s_{mi}=1} - 1}{f_{1mu}} \\ \frac{I_{s_{mi} \neq 0} - f_u - f_{2mu}}{f_{0mu}} + \frac{I_{s_{mi}=2} - 1}{f_{2mu}} \end{pmatrix}^T$$

The empirical variance-covariance matrix for

$$S(\boldsymbol{\eta}_0) = \sum_{i=1}^N S_i(\boldsymbol{\eta}_0)$$

used to compute $\mathbf{I}^{\mathbf{B}_1 \mathbf{B}_1}(\boldsymbol{\eta}_0)$ for multiple scan SNPs has entries that satisfy the general formula:

$$\sum_{i=1}^{n_{ca}} \left(S_{i\eta_x} S_{i\eta_y} - \overline{S_{ca, \eta_x}} \overline{S_{ca, \eta_y}} \right) + \sum_{i=n_{ca}+1}^N \left(S_{i\eta_x} S_{i\eta_y} - \overline{S_{co, \eta_x}} \overline{S_{co, \eta_y}} \right)$$

where η_x represents a parameter in $\boldsymbol{\eta}_0$ and data are ordered with cases ('ca') before controls ('co').

References

- Liu Y, Xu H, Chen S, Chen X, Zhang Z, Zhu Z, Qin X, Hu L, Zhu J, Zhao GP, Kong X: Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genet* 2011;7:e1001338.
- Strange A, Capon F, Spencer CC, Knight J, Weale ME, et al: A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat Genet* 2010;42:985–990.
- Carlborg O, Kerje S, Schutz K, Jacobsson L, Jensen P, Andersson L: A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res* 2003;13:413–421.
- Hsueh WC, Cole SA, Shuldiner AR, Beamer BA, Blangero J, Hixson JE, MacCluer JW, Mitchell BD: Interactions between variants in the beta3-adrenergic receptor and peroxisome proliferator-activated receptor-gamma2 genes and obesity. *Diabetes Care* 2001;24:672–677.
- Leamy LJ, Routman EJ, Cheverud JM: An epistatic genetic basis for fluctuating asymmetry of mandible size in mice. *Evolution* 2002;56:642–653.
- Templeton AR, Sing CF, Brokaw B: The unit of selection in *Drosophila mercatorum*. I. The interaction of selection and meiosis in parthenogenetic strains. *Genetics* 1976;82:349–376.
- Tripodis N, Hart AA, Fijneman RJ, Demant P: Complexity of lung cancer modifiers: mapping of thirty genes and twenty-five interactions in half of the mouse genome. *J Natl Cancer Inst* 2001;93:1484–1491.
- van Wezel T, Ruivenkamp CA, Stassen AP, Moen CJ, Demant P: Four new colon cancer susceptibility loci, Scc6 to Scc9 in the mouse. *Cancer Res* 1999;59:4216–4218.
- Chapman J, Clayton D: Detecting association using epistatic information. *Genet Epidemiol* 2007;31:894–909.
- Evans DM, Marchini J, Morris AP, Cardon LR: Two-stage two-locus models in genome-wide association. *PLoS Genet* 2006;2:e157.
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ: Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 2007;63:111–119.
- Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;37:413–417.
- Tukey JW: One degree of freedom for non-additivity. *Biometrics* 1948;5:232–242.
- Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S: Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet* 2006;79:1002–1016.
- Hartwell LH, Hopfield JJ, Leibler S, Murray AW: From molecular to modular cell biology. *Nature* 1999;402:C47–C52.
- Segre D, Deluna A, Church GM, Kishony R: Modular epistasis in yeast metabolism. *Nat Genet* 2005;37:77–83.
- Tong AH, Lesage G, Bader GD, Ding H, Xu H, et al: Global mapping of the yeast genetic interaction network. *Science* 2004;303:808–813.
- Frayling TM, Colhoun H, Florez JC: A genetic link between type 2 diabetes and prostate cancer. *Diabetologia* 2008;51:1757–1760.
- Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, et al: Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* 2007;39:977–983.
- Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, et al: Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* 2008;40:310–315.
- Kaaks R, Stattin P: Obesity, endogenous hormone metabolism, and prostate cancer risk: a conundrum of 'highs' and 'lows'. *Cancer Prev Res (Phila)* 2010;3:259–262.
- Nicolucci A: Epidemiological aspects of neoplasms in diabetes. *Acta Diabetol* 2010;47:87–95.
- Pierce BL, Ahsan H: Genetic susceptibility to type 2 diabetes is associated with reduced prostate cancer risk. *Hum Hered* 2010;69:193–201.
- 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;13:153–162.

- 25 Chatterjee N, Carroll RJ: Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 2005;92:399–418.
- 26 Umbach DM, Weinberg CR: Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med* 1997;16:1731–1743.
- 27 Prentice RL, Pyke R: Logistic disease incidence models and case-control studies. *Biometrika* 1979;66:403–411.
- 28 Beuten J, Gelfond JA, Martinez-Fierro ML, Weldon KS, Crandall AC, Rojas-Martinez A, Thompson IM, Leach RJ: Association of chromosome 8q variants with prostate cancer risk in Caucasian and Hispanic men. *Carcinogenesis* 2009;30:1372–1379.
- 29 Lin DY, Zeng D: Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *J Am Stat Assoc* 2006;101:89–118.
- 30 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;73:1316–1329.
- 31 Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et al: Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007;39:645–649.
- 32 Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, et al: Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2009;41:1055–1057.
- 33 Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, Chatterjee N: 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.
- 34 Ciampa J, Yeager M, Amundadottir L, Jacobs K, Kraft P, Chung C, Wacholder S, Yu K, Wheeler W, Thun MJ, Divers WR, Gapstur S, Albanes D, Virtamo J, Weinstein S, Giovannucci E, Willet WC, Cancel-Tassin G, Cussenot O, Valeri A, Hunter D, Hoover R, Thomas G, Chanock S, Chatterjee N: Large scale exploration of gene-gene interactions in prostate cancer using a multi-stage genome-wide association study. *Cancer Res* 2011;71:3287–3295.
- 35 Ghousaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, et al: Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* 2008;100:962–966.
- 36 Khoury MJ, Flanders WD: Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 1996;144:207–213.
- 37 Albert PS, Ratnasinghe D, Tangrea J, Wacholder S: Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol* 2001;154:687–693.
- 38 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;64:685–694.
- 39 Li D, Conti DV: Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol* 2009;169:497–504.
- 40 Murcray CE, Lewinger JP, Gauderman WJ: Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 2009;169:219–226.
- 41 Yu K, Liang F, Ciampa J, Chatterjee N: Efficient p-value evaluation for resampling-based tests. *Biostatistics* 2011;12:582–593.
- 42 Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, Yan C, Khalid O, Kantoff P, Oh W, Manak JR, Berman BP, Henderson BE, Frenkel B, Haiman CA, Freedman M, Tanay A, Coetzee GA: Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* 2009;5:e1000597.
- 43 International HapMap Project: HapMap Genome Browser, phase 1 & 2 full dataset for Utah residents with ancestry from northern and western Europe; hapmap.ncbi.nlm.nih.gov. 2010.
- 44 Lee YF, Shyr CR, Thin TH, Lin WJ, Chang C: Convergence of two repressors through heterodimer formation of androgen receptor and testicular orphan receptor-4: a unique signaling pathway in the steroid receptor superfamily. *Proc Natl Acad Sci USA* 1999;96:14724–14729.
- 45 van den Berg DL, Zhang W, Yates A, Engelen E, Takacs K, Bezstarosti K, Demmers J, Chambers I, Poot RA: Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. *Mol Cell Biol* 2008;28:5986–5995.
- 46 Schaeffer EM, Marchionni L, Huang Z, Simons B, Blackman A, Yu W, Parmigiani G, Berman DM: Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. *Oncogene* 2008;27:7180–7191.
- 47 Lee YF, Young WJ, Lin WJ, Shyr CR, Chang C: Differential regulation of direct repeat 3 vitamin D3 and direct repeat 4 thyroid hormone signaling pathways by the human TR4 orphan receptor. *J Biol Chem* 1999;274:16198–16205.
- 48 Hansen CM, Binderup L, Hamberg KJ, Carlberg C: Vitamin D and cancer: effects of 1,25(OH)2D3 and its analogs on growth control and tumorigenesis. *Front Biosci* 2001;6:D820–D848.
- 49 Omdahl JL, Morris HA, May BK: Hydroxylase enzymes of the vitamin D pathway: expression, function, and regulation. *Annu Rev Nutr* 2002;22:139–166.
- 50 Schwartz GG, Hanchette CL: UV, latitude, and spatial trends in prostate cancer mortality: all sunlight is not the same (United States). *Cancer Causes Control* 2006;17:1091–1101.
- 51 Lai JP, Sandhu DS, Yu C, Han T, Moser CD, Jackson KK, Guerrero RB, Aderca I, Isomoto H, Garrity-Park MM, Zou H, Shire AM, Nagorney DM, Sanderson SO, Adjei AA, Lee JS, Thorgeirsson SS, Roberts LR: Sulfatase 2 up-regulates glypican 3, promotes fibroblast growth factor signaling, and decreases survival in hepatocellular carcinoma. *Hepatology* 2008;47:1211–1222.
- 52 Lemjabbar-Alaoui H, van Zante A, Singer MS, Xue Q, Wang YQ, Tsay D, He B, Jablons DM, Rosen SD: Sulf-2, a heparan sulfate endosulfatase, promotes human lung carcinogenesis. *Oncogene* 2010;29:635–646.
- 53 Morimoto-Tomita M, Uchimura K, Bistrup A, Lum DH, Egeblad M, Boudreau N, Werb Z, Rosen SD: Sulf-2, a proangiogenic heparan sulfate endosulfatase, is upregulated in breast cancer. *Neoplasia* 2005;7:1001–1010.
- 54 Nawroth R, van Zante A, Cervantes S, McManus M, Hebrok M, Rosen SD: Extracellular sulfatases, elements of the Wnt signaling pathway, positively regulate growth and tumorigenicity of human pancreatic cancer cells. *PLoS One* 2007;2:e392.
- 55 Dai Y, Yang Y, MacLeod V, Yue X, Rapraeger AC, Shriver Z, Venkataraman G, Sasisekharan R, Sanderson RD: HSulf-1 and HSulf-2 are potent inhibitors of myeloma tumor growth in vivo. *J Biol Chem* 2005;280:40066–40073.
- 56 Conde-Knappe K: Heparan sulfate proteoglycans in experimental models of diabetes: a role for perlecan in diabetes complications. *Diabetes Metab Res Rev* 2001;17:412–421.
- 57 Datta MW, Hernandez AM, Schlicht MJ, Kahler AJ, DeGueme AM, Dhir R, Shah RB, Farach-Carson C, Barrett A, Datta S: Perlecan, a candidate gene for the CAPB locus, regulates prostate cancer cell growth via the Sonic Hedgehog pathway. *Mol Cancer* 2006;5:9.
- 58 Kosir MA, Wang W, Zukowski KL, Tromp G, Barber J: Degradation of basement membrane by prostate tumor heparanase. *J Surg Res* 1999;81:42–47.
- 59 Murphy T, Darby S, Mathers ME, Gnana-pragasam VJ: Evidence for distinct alterations in the FGF axis in prostate cancer progression to an aggressive clinical phenotype. *J Pathol* 2010;220:452–460.
- 60 Verras M, Sun Z: Roles and regulation of Wnt signaling and beta-catenin in prostate cancer. *Cancer Lett* 2006;237:22–32.
- 61 Kwabi-Addo B, Ozen M, Ittmann M: The role of fibroblast growth factors and their receptors in prostate cancer. *Endocr Relat Cancer* 2004;11:709–724.
- 62 Wang P, Keijer J, Bunschoten A, Bouwman F, Renes J, Mariman E: Insulin modulates the secretion of proteins from mature 3T3-L1 adipocytes: a role for transcriptional regulation of processing. *Diabetologia* 2006;49:2453–2462.
- 63 Freedman D: How can the score test be inconsistent? *Am Stat* 2007;61:291–295.
- 64 Morgan B, Palmer K, Ridout M: Negative score test statistic. *Am Stat* 2007;61:285–288.
- 65 Lin DY, Zou F: Assessing genomewide statistical significance in linkage studies. *Genet Epidemiol* 2004;27:202–214.