Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer Genome-Wide Association Scan

The CGEMS data portal provides public access to summary results for approximately 550,000 SNPs genotyped in the CGEMS prostate cancer scan (Phase 1A with HumanHap300 and Phase 1B HumanHap240, both from Illumina, San Diego, CA) in more than 1,100 prostate cancer patients and an equivalent number of controls from the PLCO study. Analysis of nearly 550,000 SNP genotypes per subject provides approximately 90% coverage of common SNPs based on HapMap Phase 2 with minor allele frequency (MAF) greater than 0.05 in the European population and a linkage disequilibrium coefficient of $r^2>0.8$ with Tagzilla (http://tagzilla.nci.nih.gov/)[1-3].

The summary data can be viewed via the CGEMS data portal and downloaded in bulk. Future data releases at this portal will include results on additional genotype data with limited phenotype information. Access to a subset of the PLCO individual raw phenotype and genotype data will be possible for research scientific purposes only after registration by the individual investigator and the supporting institution. The accessible data will include genotypes from the WGAS and a set of covariates, namely, age (in categories of 5 years, 55-60, 60-65 and 65-70), family history of cancer (yes/no), and disease phenotype (control, non-aggressive prostate cancer, aggressive prostate cancer). Access to additional covariate data will be possible through established data sharing policies of PLCO (http://www.parplco.org/; contact Danielle Carrick, PLCO EEMS Study Coordinator, Westat, Inc., Rockville, MD, tel: 240-314-5896).

Study Population

The Prostate, Lung, Colon and Ovarian (PLCO, http://www.cancer.gov/prevention/plco/) Cancer Screening Trial is a large, randomized controlled trial of approximately 155,000 men and women. Participants are randomized to either a screening or control arm. Each year after enrollment, subjects are asked to notify the study of any cancers diagnosed in the past year using the Annual Study Update (ASU).

The trial is designed to test the efficacy of cancer screening to prevent early death from prostate, lung, colorectal and ovarian cancer. The collection of questionnaire data and biospecimens (e.g., repeated blood samples and in some instances, buccal cell samples) allows investigation of early markers for cancer as well as etiology of common cancers[4-6].

PLCO enrollment began in 1993 and ended in 2001. Recruitment included men and women, aged 55 to 74 with no reported history of prostate, lung, colon and ovarian cancer, although prior diagnoses of other cancers were acceptable.

The CGEMS cohort consisted of men enrolled in the screening arm of the PLCO Trial who:

1. were White and non-Hispanics;
2. had no prior history of prostate of cancer before randomization;
3. had at least one PLCO prostate cancer screen (PSA) before October 1, 2003;

4. had completed a Baseline Questionnaire about risk factors for cancer;
5. had signed informed consent;
6. had provided a blood sample with
    a. at least 11 μg DNA
    b. at least 1 vial of buffy coat, or
    c. at least 7 vials of whole blood was available; *and*
7. for controls, had returned at least one Annual Study Update (ASU).

Based on these criteria, 28,521 men were included in the CGEMS sub-cohort.

CGEMS distinguishes between non-aggressive and aggressive cases of prostate cancer at the time of diagnosis. The two subtypes are defined as follows:

1. ***Non-aggressive***: cases with a Gleason Score < 7 and Stage $<$ III.
2. ***Aggressive***: cases with a Gleason Sore ≥ 7 or Stage $\geq$ III.

Study enrollment began on October 1, 1993. Consequently, study years in the PLCO Trial are counted according to the Federal fiscal year, Oct 1 to the next September 30.

All men diagnosed with prostate cancer between enrollment and the end of FY2001 were considered for inclusion in CGEMS. Because of our interest in the clinically more significant, but less common aggressive form of prostate cancer, we increased the fraction of aggressive cases in the CGEMS case series by extending eligibility for cases diagnosed with aggressive prostate cancer through the end of FY2003.

A total of 1,361 subjects with prostate cancer met the eligibility criteria and were considered for the CGEMS project; 737 cancers were aggressive 624 cancers were non-aggressive. Of the eligible cases, all aggressive cases (n=737) were chosen to be cases in the CGEMS prostate cancer study. Of the 624 men found to have non-aggressive tumors, 493 men (70.4%) whose diagnosis was temporally closest to the first screening were included in this study.

Controls were selected by incidence-density sampling. The first step was creation of non-overlapping sets of cases characterized by:

1. Calendar year (FY) of entry into the cohort,
2. Age at entry in five-year intervals (55-59, 60-64, 65-69, 70-74)
3. Number of years under follow-up between enrollment and diagnosis of prostate cancer.

Next, for each case set, we identified eligible men among all 28,251 men in the CGEMS cohort who met each of the following three criteria:

1. Same year of entry into the cohort as the case set;
2. Same five-year age-at-entry interval (55-59, 60-64, 65-69, 70-74) as the case set; and
3. Observed through the year of follow-up in the case set with no prostate cancer diagnosis.

In incidence density sampling, a male subject is included as a control for a given case set independently of eligibility and selection as a control for other case sets and independently of future diagnosis as a case. Our goal was to obtain a random sample of controls from the corresponding risk set with size equal to the number of cases in the set. For cases diagnosed before 2002, some samples had already been extracted for another prostate cancer study within the PLCO cohort. To assure that all eligible subjects had equal inclusion probability for this study, we replaced some of the previously selected controls with newly enrolled cohort members according to a random selection rule that ensured that the chance of inclusion as a control in the CGEMS study for a given case was the same for each man eligible to be a control.

1,204 different men, representing 1,230 control selections, were identified as controls at least once (1,179 subjects sampled once, 24 subjects sampled twice, and one subject sampled three times). Characteristics of the 1,230 controls selected are in Table 1 below. Forty-six control subject selections, derived from 44 eligible subjects, subsequently developed prostate cancer; 32 of these cases were included as cases and 12 were among the men with non-aggressive prostate cancer not included in the analysis.

Table 1. Potential CGEMS participants from PLCO cohort

Case status	Number of times selected as a control				Total
	0	1	2	3	
No prostate cancer diagnosis	26,000	1,136	23	1	27,160
Prostate ca dx, but not selected as a case	119	11	1	0	131
Prostate ca dx, and selected as a case	1198	32	0	0	1,230
All subjects	27,317	1,179	24	1	28,521

Sample handling:

DNA samples were received from the PLCO bio-repository and visually inspected for adequate fluid in individual tubes. Three measurements of DNA quantification were performed according to the standard procedures at the Core Genotyping Facility of the National Cancer Institute[7]. These include pico-green analysis, optical density spectrophotometry and real time PCR (http://cgf.nci.nih.gov/dnaquant.cfm). Samples were also analyzed with 15 short tandem repeats and the Amelogenin marker in the Identifiler™ Assay (ABI, Foster City, CA). Samples that completed less than 13 of the 15 micro-satellite markers were excluded and not deemed suitable for additional genotyping.

Table 2. Final set of PLCO samples genotyped in CGEMS

Status at initiation of CGEMS project	Number of times selected as a control				Total
	0	1	2	3	
Prostate cancer-negative during follow-up	0	1,087	22	1	1,110
Diagnosed with non-aggressive cancer	466	26	1	0	493
Diagnosed with aggressive cancer	679	16	0	0	695
All subjects	1,145	1,129	23	1	2,298

After final review and sample handling, 1,188 of 1,361 (87.2%) of eligible cases were genotyped in CGEMS. A total of 1,188 men with prostate cancer and 1,110 men not diagnosed with cancer are included in our analysis.

For quality control analysis, 49 DNA samples from PLCO were genotyped in duplicate. We also genotyped 100 DNAs from Centre d'Etude du Polymorphisme Humain (CEPH, Paris, France) families of which 77 were genotyped in duplicate.

Selection of SNPs:

Genotyping of the CGEMS Prostate Cancer Study was performed under contract by Illumina Corporation in two parts, Phase 1A used the Sentrix® HumanHap300 genotyping assay and Phase 1B used the Sentrix® HumanHap240[8-10]. Together, the chips constitute a fixed panel of 561,494 tagSNPs identified following the method initially described by Carlson et al[11]. This selection was performed using the data from the International HapMap project (http://www.hapmap.org/) that included a threshold for the linkage disequilibrium statistic r^2>0.7 for non-genic regions and r^2>0.8 for genic regions[9]. For the European population, this panel is expected to cover close to 90% of the common SNPs in HapMap phase 1 at a threshold of r^2>0.8 as evaluated by the TagZilla (http://tagzilla.nci.nih.gov/) program.

Quality control

Initial Assessment of Call Rates

A total of 561,494 SNP genotype assays were attempted on 2,540 DNA samples in two phases, Phase 1A (HumanHap300) and Phase 1B (HumanHap240) (see Table 3). If the completion rate for any sample was below 90%, the sample was excluded from further analysis; 18 CEPH samples and 7 PLCO samples were excluded by this criterion. 175 CEPH samples (including 73 duplicamte DNAs including one set of a quadruplet) and 2,340 PLCO samples (including 49 duplicate DNAs) passed.

A total of 14,901 SNPs (~2.6% overall) failed to provide accurate genotype results due to either low locus call rates (<90%) or were monomorphic across the study. Subsequent quality control analysis was performed on the remaining 546,593 SNPs.

Table 3. Sample completion rates

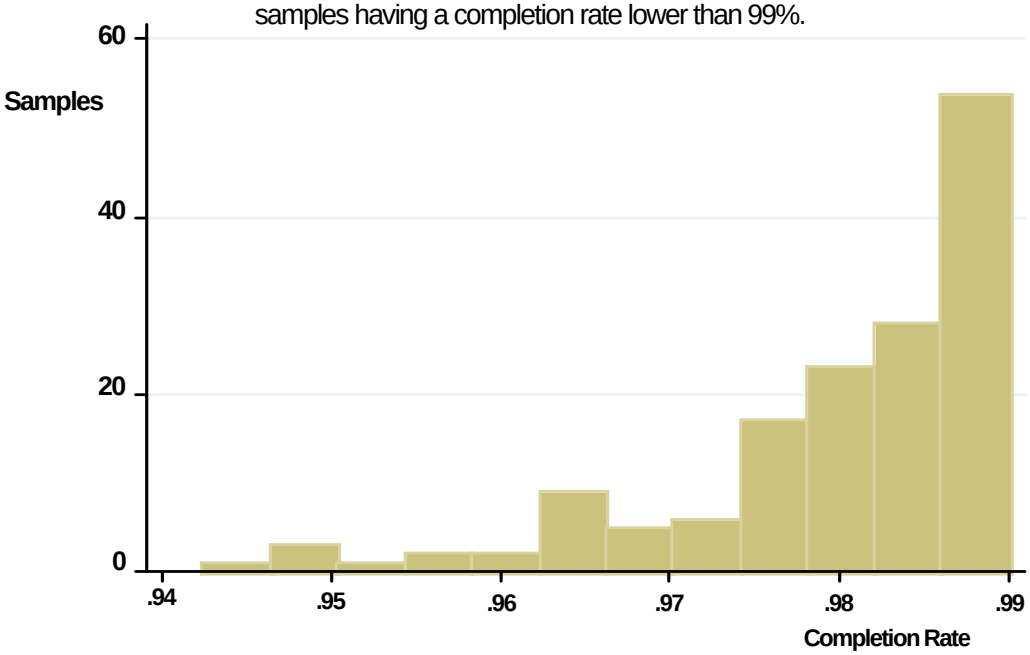
Phase	Sample Subset	Typed on Human Hap300	Typed on Human Hap240	Completion Rate Successful SNPs*	Completion Rate All SNPs
Phase 1A	All	X		99.731	96.523
Phase 1A	PLCO	X		99.742	96.534
Phase 1A	CEPH	X		99.602	96.399
Phase 1B	All		X	99.850	98.351
Phase 1B	PLCO		X	99.852	98.801
Phase 1B	CEPH		X	99.789	98.738
Phase 1A+1B	All	X		99.448	96.578
Phase 1A+1B	All		X	99.687	98.637
Phase 1A+1B	All	X	X	99.796	98.222
Phase 1A+1B	PLCO	X		99.799	98.255

Phase	Sample Subset	Typed on Human Hap300	Typed on Human Hap240	Completion Rate Successful SNPs*	Completion Rate All SNPs
Phase 1A+1B	PLCO		X	N/A**	N/A**
Phase 1A+1B	PLCO	X	X	99.798	98.224
Phase 1A+1B	CEPH	X		99.600	96.397
Phase 1A+1B	CEPH		X	99.702	98.653
Phase 1A+1B	CEPH	X	X	99.589	98.018

* Successful SNPs are those with >90% completion across all attempted samples
** Samples that failed in Phase 1A were not attempted in Phase 1B.

Overall, 219 of 243 (90%) CEPH samples and 2,243 of 2,340 (96%) PLCO samples generated high performance genotype calls (>99% completion rate). The lowest accepted call rate was 94.24% in Phase 1A and 97.27% for Phase 1B. Overall, the average completion rate for the PLCO samples for 546,593 SNPs is 99.8%. This rate was not significantly different in the controls compared to the combined aggressive and non-aggressive case groups (status at initiation of CGEMS project, t test p=0.30, Kruskal-Wallis p=0.46) and remained non-significant when the two case phenotypes were separated (Kruskal-Wallis p=0.76). No additional subjects were excluded from the association analysis based on completion rate. Figure 1 provides the distribution of the completion rate for the 151 remaining samples in Phase 1A with a completion rate below 99%.

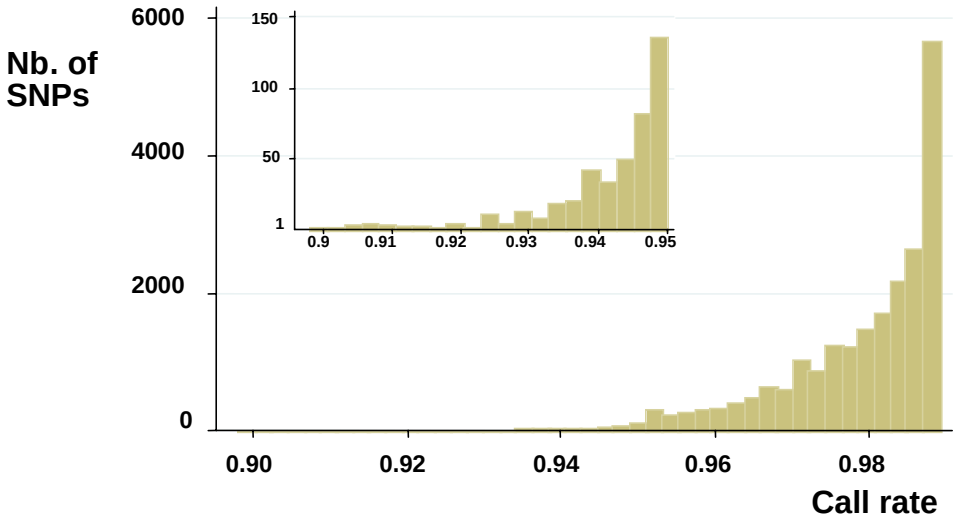
Figure1 : Distribution of the completion rates for the 151 PLCO samples having a completion rate lower than 99%.



Call rate for SNPs

A total of 538,548 (96%) SNPs yielded a call rate higher than 99%. For example, Figure 2 provides the distribution of the call rates for SNPs in the Phase 1A that fell below this threshold. The lowest call rate retained for our analysis was 89%.

Figure 2 Distribution of the call rates of the 22 064 SNPs having a call rate lower than 99%.



Concordance rate

The genotype concordance rate for SNP assays was evaluated based on three comparisons:

1. Genotype generated on CEPH DNA.

Of the CEPH DNAs, 77 had been provided twice as separate aliquots from the same DNA preparation, then genotyped and evaluated for genotype concordance. An average of 305,400 genotypes comparisons were performed for each DNA pair. Between 0 and 1,342 discordances (average 56) were observed within each pair comparison, yielding a discordance rate of 1.85×10^{-4}.

2. Genotypes generated on duplicate DNA from PLCO.

From PLCO, 56 pairs of controls DNA were provided twice and analyzed as separate aliquots from the same DNA preparation but performed comparably, thereby, providing reliable data. The DNA samples were selected randomly. Analysis of the discrepancies within these pairs revealed similar results to the CEPH DNA duplicates. An average discordance rate 1.88×10^{-4} was observed.

3. Comparison of CGEMS and HapMap Genotypes

28 samples genotyped in CGEMS were also genotyped by the International HapMap Consortium[2,3]. The discordance rate between the CGEMS Phase 1A and HapMap genotypes for these individuals was 1.7×10^{-3}, which is an order of magnitude higher than that observed for PLCO duplicates in the CGEMS project. Notably, the International HapMap Consortium received its DNA from the Corriell Institute (Camden, New Jersey) and CGEMS received DNA from the CEPH. Although the DNAs were obtained from cell

lines derived from the same individual, they were extracted from cell lines propagated at different times and location and processed independently. The distribution of discordances was as follows: 94% between homozygote and heterozygote calls and only 6% showed homozygote/homozygote discordances.

Hardy–Weinberg Proportions in control DNA

Genotype data for all autosomal SNPs were tested on PLCO controls for deviation from Hardy-Weinberg proportions[12]. The analysis was conducted in the PLCO control group. Significant p values ($p < 0.05$) were observed for 30,887 SNPs (5.8%).

Final sample and SNP selection for association analysis

Three DNA samples revealed a large number of heterozygous loci on the X chromosome, suggesting sample mix-up. In addition, three pairs of samples were found to be unexpected duplicates. These 9 subjects were removed from subsequent analysis. Thus, association analysis was performed on the final set of 2,282 subjects described in table 4

Table 4. Final set of PLCO samples analyzed for association

Status at initiation of CGEMS project	Number of times selected as a control				Total
	0	1	2	3	
Never developed prostate cancer	0	1,082	22	1	1,105
Diagnosed with non-aggressive cancer	461	26	1	0	488
Diagnosed with aggressive cancer	673	16	0	0	689
All subjects	1,134	1,124	23	1	2,282

Association Analysis

The primary analysis of the CGEMS prostate GWAS study explores the association between single SNPs and prostate cancer susceptibility in 561,494 SNPs per subject.

By maximizing genome coverage for a given number of SNPs, we increase the opportunity to pursue different working hypotheses and different regions of interest now and in the future. Similarly, our 'agnostic' approach to the analysis does not take gene function or prior information in prostate or other cancers into consideration.

The analytic approach assumes no structure to the risk across the 3 possible genotypes at each locus. This approach maintains power to detect recessive or over-dominant alleles at the cost of a small decrease in power relative to an Armitage trend test for the detection of alleles with multiplicative effect.

Prostate cancer stage and grade at diagnosis are important predictors of survival; they may also have different etiologic factors. Therefore, we distinguish between non-aggressive and aggressive prostate cancer in the analysis. Essentially, our analysis combines the effect from looking separately at the two case phenotypes. Our analysis has power to identify susceptibility loci specific to aggressive or non-aggressive prostate cancer, at a small cost of power for loci with the same odds ratio in aggressive and non-aggressive cases.

Analytic approaches

We present results from two distinct analytic approaches. The first scheme is more frequently used in case control studies. The second scheme takes full advantage of the prospective nature of the PLCO cohort and the power from incidence density sampling.

Analysis with Single selection

For this scheme, which will be more familiar to non-epidemiologists, does not account for the dynamic nature of the cohort. Genotypes of individuals that have been selected as a case in the relevant phenotype case group are counted once as a case and never as a control. Individuals who have been selected several times as controls but had not developed prostate cancer during follow-up are counted only once in the control group.

Analysis accounting for incidence density sampling

Selection of controls from cases identified in a cohort that accounts for the dynamic nature of the cohort including development of disease during follow-up and timing of entry to and exit from follow-up may have more power to detect an association than the single-selection method. The main feature of incidence-density sampling, as used for control selection here, is that controls are selected independently for each case among those who are at risk at the time of the diagnosis of the case; i.e., among those who would become a case in the study had they developed disease at the same time. Inclusion as a control for a given case set is independent of future diagnosis as a case, of selection as a control for other case sets, and of entry and exit times. Thus, individuals may be included as a case and as a control. Genotypes of individuals who have been selected multiples times are taken into account each time he is selected; the man's covariates that vary with time, such as age are defined differently each time, depending on the characteristics of the case set for which he was selected as a control[13].

Genotypes

In order to maintain high power to detect SNPs that are involved in non multiplicative models (such as complete recessivity or over-dominance), we provide analyses of the data based on genotype frequencies. Each of the three possible genotype states are considered separately. Accordingly, for autosomal loci, analysis of each case phenotype uses a statistical test with two degrees of freedom for each case phenotype considered separately (aggressive and non aggressive separately yield 4 degrees of freedom with 3 genotypes). For tests involving X-linked loci, a single degree of freedom is used per case phenotype.

Single SNP statistics

In order to expedite public access to the data, the first-pass analysis of the CGEMS data aims at detecting association of single SNPs with prostate cancer susceptibility. Multi-SNP approaches, such as haplotype association, have not yet been performed.

Statistical tests.

We performed four sets of analyses.
For two tests using single selection, analysis included

- 561,494 **SNPs**,
- 488 ***cases*** diagnosed with ***non-aggressive*** tumors and
- 688 ***cases*** diagnosed with ***aggressive*** tumors.

For two tests using incidence density sampling, analysis included

- 561,494 **SNPs**,
- 476 ***cases*** diagnosed with ***non-aggressive*** tumors and
- 688 ***cases*** diagnosed with ***aggressive*** tumors.

The characteristics of the four tests are:

1. ***Single selection, unadjusted*** score test
   - 1,101 ***controls*** that were not diagnosed with prostate cancer at the start of follow-up.
   - ***3-by-3 contingency table*** of genotypes by phenotypes was constructed.
   - ***No adjustment*** for covariates.
   - The ***p-value*** from the standard test of independence was computed from the ***3-by-3 contingency table*** table, based on a chi-squared test with up to 4 degrees of freedom.

2. ***Single selection, adjusted*** score test
   - 1,101 ***controls*** that were not diagnosed with prostate cancer at the start of follow-up.
   - ***Polytomous logistic regression*** was performed.
     - Two case phenotypes with a common set of controls;
     - The regression variable was a two-indicator variable for genotype.
   - ***Adjustment*** for
     - Age group at randomization (4 groups),
     - Region of recruitment (9 regions) *and*
     - Indicator variable for cases diagnosed within one year of entry to the trial.
     - 3 sets of eigenvectors corresponding to three top principle components identified by Eigenstrat program.
   - The ***p-value*** was obtained from a score test with up to 4 degrees of freedom.

3. ***Incidence-density sampling, unadjusted*** score test
   - 1,168 ***controls*** using an incidence density sampling strategy.
   - ***3-by-3 contingency*** table of genotypes by phenotypes was constructed.
   - ***No adjustment*** for covariates
   - The ***p-value*** from the standard test of independence was computed from this table, based on a chi-squared test with 4 degrees of freedom (or fewer if there were empty cells).

4. ***Incidence density sampling, adjusted*** score test
   - 1,168 ***controls*** selected using an incidence density sampling strategy.
   - ***Polytomous logistic regression***
     - Two case phenotypes with a common set of controls;
     - The regression variable was a two-indicator variable for genotype.
   - ***Adjustment*** for
     - Age group at randomization (4 groups),
     - Region of recruitment (9 regions), *and*
     - 0-1 variable that indicates cases diagnosed within one year of entry to the trial.
     - 3 sets of eigenvectors corresponding to three top principle components identified by Eigenstrat program.
   - The ***p-value*** was obtained from a score test with up to 4 degrees of freedom.

Interpreting the results

In examining the results one should keep in mind the following points:

1. Markers were selected on genomic criteria, not on functional basis. In the absence of complementary information, each of the SNPs has a low *a priori* probability. Observation of a low p-value in these tables is not sufficient evidence to demonstrate an association for the marker; additional studies are required to confirm the association. For this analysis, we expected to observe roughly $\alpha \times 3 \times 10^5$ p-values lower than a specified α when there is one statistical test for each of 3×10^5 SNPs by chance alone; thus for $\alpha = 10^{-3}$ or $\alpha = 10^{-5}$, we expected to observe 300 and 3 SNPs, respectively, meeting the criterion by chance. In the pre-computed analysis presented we observed 314±13 (with a range of 301 to 327) depending on which of the four tests was selected for $\alpha = 10^{-3}$. For $\alpha = 10^{-5}$, we observed between 7 and 9 SNPs for each of the four tests. Nevertheless, the observation of a low p-value for a SNP in this GWAS alone does suggest that the associated gene or chromosomal region has an increased likelihood of harboring a prostate cancer susceptibility locus but follow-up analysis is required and is planned in the follow-up phases of CGEMS (http://cgems.cancer.gov/).

2. Many pairs of SNP markers may have substantial correlation between them. In fact, correlation may extend across several markers on the same chromosomal region. Before interpreting the observation of clustering of SNPs with low p-values in a small chromosomal region as a strong signal of the presence of susceptibility loci in the region, one must consider that the clustering may be a consequence of linkage disequilibrium between neighboring SNPs. Similarly, the p values across the 4 statistical tests are highly correlated.

3. The four tests we used for each SNP are strongly associated. It is probably best to choose one test for exploratory purposes. We recommend using the fourth one, ***Incidence density sampling, adjusted*** score test for exploratory purposes.

Citation of data used:

Please cite the website for publications related to data available on this website (http://cgems.cancer.gov/) and reference the full name of the study, Cancer Genetic Markers of Susceptibility.

Appendix for statistical test

COMPUTATION OF 4 D.F SCORE-TEST FOR CGEMS ANALYSIS

The association of each SNP with advanced and non-advanced prostate cancer was tested using a 4 d.f score-test based on a polytomous logistic regression model[14]. If $Y = 0, Y = 1$ and $Y = 2$ denote controls, non-advanced and advanced prostate cancer cases, respectively, we specify the probability of observing a phenotype category as a function of the genotype data (G) and a set of co-factors (Z) as

$$\begin{aligned} &\Pr(Y = 0 \mid G, Z) = K \\ &\Pr(Y = 1 \mid G, Z) = K \exp\{\alpha_1 + \gamma_1 Z + \beta_{11} I(G = 1) + \beta_{12} I(G = 2)\} \\ &\Pr(Y = 2 \mid G, Z) = K \exp\{\alpha_2 + \gamma_2 Z + \beta_{21} I(G = 1) + \beta_{22} I(G = 2)\} \end{aligned} \qquad (0.1)$$

where K denotes a normalizing constant, $I(G = 1)$ and $I(G = 2)$ denote indicator variables for heterozygous and homozygous variant genotypes for a given SNP. In (0.1), the parameters β_{11} and β_{12} denote the log-odds-ratios associated with heterozygous and homozygous variant genotypes for non-advanced cases and β_{21} and β_{22} denote those for advanced cases. The null hypothesis of interest is

$$H_0 : \beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = 0;$$

i.e., that carrying 1 or 2 variant alleles is not associated with aggressive or non-aggressive tumors. The parameters α_1 and α_2 determine the baseline probability of non-advanced and advanced prostate cancer for subjects with homozygous wild-type genotype $(G = 0)$ and some reference values for the covariates. The parameters γ_1 and γ_2 denote the log-odds-ratio associated with covariates (Z) for advanced and non-advanced cases, respectively.

We implemented an Iterated Re-weighted Least Square (IRLS) algorithm[15] for obtaining the maximum-likelihood estimates of the parameters from model (0.1).

Steps for score-test calculations

1) Obtain estimates of covariate effects under the null

Estimate of $\theta_1 = (\alpha_1, \gamma_1)$ and $\theta_2 = (\alpha_2, \gamma_2)$ using the IRLS algorithm by setting the design matrix as $X = [1\ Z]$

2) Compute the score-vector

The formula for score-function S_{ij} for β_{ij} is given by

$$S_{ij} = \sum_{m=1}^{N_0+N_1} I(G_m = j)\{I(Y_m = i) - p_i(Z_m)\},$$

where $p_i(Z) = \Pr_{\theta_i}(Y = i \mid Z)$ is computed using estimate of θ_i obtained from step (1).

Alternatively, the score can be written in terms of "cell frequencies" (as opposed to individual level data) as following. Suppose the covariate Z defines a total of K strata. Let n_{ijk} and N_{jk} be the number of subjects in the data with $(Y = i, G = j, Z = z_k)$ and $(G = j, Z = z_k)$, respectively. Then the score S_{ij} can be written as

$$S_{ij} = \sum_{k=1}^{K} \{n_{ijk} - n_{+jk}\, p_i(z_k)\},$$

which has the usual (O-E) (difference between observed and expected) form. Define the score vector $\boldsymbol{S} = (S_{11}, S_{12}, S_{21}, S_{22})$.
Now we need to find the variance of the score-vector.

3) Variance calculation

Steps for variance calculations

3.1) Define the design matrix $X = [1\ Z\ I(G=1)\ I(G=2)]$

3.2) Based on this design matrix and parameter values $(\theta_1 = \widehat{\theta}_1, \beta_{11} = 0, \beta_{21} = 0, \theta_2 = \widehat{\theta}_2, \beta_{21} = 0, \beta_{22} = 0)$ compute the information matrix which can be as part of the IRLS algorithm.

3.3) Get $V = I^{-1}$ and extract the 4 by 4 sub-matrix from V, say $V_{\beta\beta}$ that corresponds to the rows and columns for the four β_{ij} parameters.

4) The final test-statistics is now given by $T = \boldsymbol{S} V_{\beta\beta} \boldsymbol{S}'$

Handling missing data on genotypes

The null model in step (1) should be fitted using all subjects that have covariate data Z, irrespective of whether those subjects have missing genotypes. Note that, under the null, this gives statistically the most efficient estimate of the covariate effect parameters θ_1 and θ_2. Moreover, this will be computationally also very efficient as one has to fit the "null" polytomous regression model only once. Alternatively, for computing the test of association for a particular SNP, one can estimate the parameters of the null model using only the data from those subjects who have complete genotype data for that SNP, but this will require fitting up to 550,000 different "null" polytomous model.

Once θ_1 and θ_2 are estimated using the full data, the score-vector calculation for a particular SNP will remain as above, except that the "sums" would now involve only those subjects who has complete genotype data on that particular SNP. The information matrix (I) calculation for a particular SNP will similarly involve only those subjects with complete genotype data for that particular SNP, except that the sub-matrix of I that corresponds to the parameters (θ_1 and θ_2) should be computed based on all the subjects that went into estimation of these parameters. This sub-matrix is simply given by the information matrix computed in step (1).

Reference List

1. Barrett,J.C. & Cardon,L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659-662 (2006).

2. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

3. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796 (2003).

4. Gohagan,J.K., Prorok,P.C., Hayes,R.B. & Kramer,B.S. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin. Trials* **21**, 251S-272S (2000).

5. Hayes,R.B. *et al.* Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clin. Trials* **21**, 349S-355S (2000).

6. Hayes,R.B. *et al.* Methods for etiologic and early marker investigations in the PLCO trial. *Mutat. Res.* **592**, 147-154 (2005).

7. Haque,K.A. *et al.* Performance of high-throughput DNA quantification methods. *BMC. Biotechnol.* **3**, 20 (2003).

8. Gunderson,K.L. *et al.* Whole-genome genotyping. *Methods Enzymol.* **410**, 359-376 (2006).

9. Gunderson,K.L. *et al.* Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics.* **7**, 641-648 (2006).

10. Steemers,F.J. & Gunderson,K.L. Illumina, Inc. *Pharmacogenomics.* **6**, 777-782 (2005).

11. Carlson,C.S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106-120 (2004).

12. Wigginton,J.E., Cutler,D.J. & Abecasis,G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887-893 (2005).

13. Wacholder,S., Silverman,D.T., McLaughlin,J.K. & Mandel,J.S. Selection of controls in case-control studies. III. Design options. *Am. J. Epidemiol.* **135**, 1042-1050 (1992).

14. Hosmer,D.W. & Lemeshow,S. *Applied Logistic Regression.* John Wiley and Sons, Hoboken, NJ (2000).

15. McCullagh,P. & Nelder,J.A. *Generalized Linear Models.* Chapman and Hall, London (1989).