Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer Follow-up Scan 1

A genome-wide association study (GWAS) with 527,869 SNPs on 1,172 cases of prostate cancer and 1,157 controls of European origin nested in a prospective study (PLCO) was followed by testing of 26,958 SNPs in four independent studies (total: 3,941 cases and 3,964 controls).

Previous whole-genome scan

The CGEMS data portal provides public access to association results for approximately 550,000 SNPs genotyped in the CGEMS prostate cancer whole-genome scan (Phase 1A with HumanHap300 and Phase 1B HumanHap240 assays, both from Illumina Corp., San Diego, CA) in more than 1,100 prostate cancer patients and an equivalent number of controls from the PLCO Cancer Screening Trial. The materials and methods of the initial genome-wide scan have been reported[1] and are also available in a PDF format at http://cgems.cancer.gov/data/. Analysis of nearly 550,000 SNP genotypes resulted in the choice of 29,018 SNPs to type in additional 4,020 prostate cancer cases and 4,028 controls drawn from four additional studies using a custom Illumina iSelect™ assay chip.

Studies

1. Prostate, Lung, Colon and Ovarian (PLCO)

The Prostate, Lung, Colon and Ovarian (PLCO) Cancer Screening Trial is a large, randomized controlled trial of approximately 155,000 men and women[2,3]. 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].

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 testing) 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 sufficient to meet one of the following criteria:

   a. at least 11 µg DNA after extraction
   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 ≥ 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 following 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.

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
3. Observed through the year of follow-up in the case set with no prostate cancer diagnosis.

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.

The analysis of PLCO subjects for the CGEMS whole-genome scan applied an incidence density sampling approach to obtain a random sample of controls from the corresponding risk set with size equal to the number of cases in the set. 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. However in this follow up analysis, these data were analyzed by the simpler cumulative incidence sampling, where each subject is included once in a phenotype group at the last of the study follow up. The ranking of the most significant SNPs changed little when shifting from incidence to cumulative density sampling.

2. Health Professionals Follow-up Study

The Health Professionals Follow-up Study (HPFS) began in 1986 and is an ongoing prospective cohort study of 51,529 United States male dentists, optometrists, osteopaths, podiatrists, pharmacists, and veterinarians 40 to 75 years of age[5]. The baseline questionnaire provided information on age, marital status, height and weight, ancestry, medications, smoking history, disease history, physical activity, and diet. At baseline the cohort was 97% white, 2% Asian American, and 1% African American. The median follow-up through 2005 was 10.5 years (range 2-19 years). Self-reported prostate cancer diagnoses were confirmed by obtaining medical and/or pathology records. Prostate cancer deaths are either reported by family members in response to follow-up questionnaires, discovered by the postal system, or the National Death Index. Questionnaires are sent every two years to surviving men to update exposure and medical history. In 1993 and 1994, a blood specimen was collected from 18,018 men without a prior diagnosis of cancer.

Prostate cancer cases are matched to controls on birth year (+/-1) and ethnicity. Controls are selected from those who are cancer-free at the time of the case's diagnosis, and had a prostate-specific antigen test after the date of blood draw.

3. Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC)

The Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC Study) was initiated in 1985 as a randomized controlled trial of 29,133 Caucasian male smokers 50-69 years old in southwestern Finland that tested the cancer preventive effects of vitamin E and beta-carotene on cancer incidence[6]. It was continued as a longitudinal cohort, following the intervention period. Current follow-up exists for up to 20 years (16 years at the time the nested case-control set was created for the Breast Prostate Cancer Cohort Consortium, known as BPC3). Cancer incidence and overall mortality are 100% ascertained follow-up through national registries in Finland. Serum and questionnaire data were collected for all participants at baseline, and

whole blood (for DNA) in 1992 from 20,243 men. The ATBC Study nested set for the BPC3 included 1,058 prostate cancer cases and 1,058 controls without prostate cancer who were matched based on age, intervention assignment, date of blood draw.

4. The American Cancer Society Cancer Prevention Study II Nutrition Cohort (CPS-II)

The American Cancer Society Cancer (ACS) Prevention Study II Nutrition Cohort (CPS-II) was established in 1992; the cohort includes over 86,000 men and 97,000 women from 21 U.S. states who completed a mailed questionnaire in 1992. At baseline, the cohort was 97% white and the median age of participants was 63 (range: 40-92)[7]. Starting in 1997, follow-up questionnaires have been sent to surviving cohort members every other year to update exposure information and to ascertain occurrence of new cases of cancer; a >90% response rate has been achieved for each follow-up questionnaire. Incident cancers are verified through medical records, state cancer registries, or death certificates. From 1998 - 2001, blood samples were collected from a subgroup of 39,376 cohort members. To further supplement the DNA resources, during 2000 - 2001, buccal cell samples were collected by mail from an additional 70,004 cohort members.

Cases of prostate cancer were frequency matched to controls on single year of age, ethnicity, date of sample collection and specimen type (blood or buccal cell sample). All controls were selected from individuals who were cancer-free (except for non-melanoma skin cancer) at the beginning of the interval preceding the diagnosis of each case. Similarly, cases were not eligible if they had a history of another cancer other than non-melanoma skin cancer prior to their diagnosis of prostate cancer. Oversampling of aggressive prostate cancer cases included mainly buccal samples and controls were matched as above using buccal samples.

For all cases, exposure information was collected by questionnaire before the cancer diagnosis. However, collection of some biologic samples occurred after cancer diagnosis or immediately preceding cancer diagnosis. DNA from buccal samples were extracted with a standard method[8].

5. The CeRePP French Prostate Case-Control Study

The French Prostate Cancer Case Control Study (CeRePP), also known as ProGene study began in July 1994. Patients treated in three French Departments of Urology (Paris, Brest and Nancy) were recruited for the study[9]. Patients who had histologically confirmed prostate cancer were included after receiving their informed consent. At this time, a blood sample was collected. Clinical and histological data in the probands were obtained from medical records. Each participating patient received a questionnaire by mail to collect family data on their first-degree relatives (date of birth, and if deceased, age at death and cause of death) and the occurrence of cancers in their families.

Controls were recruited as participating in a systematic health screening program funded by the French National Health Insurance and found unaffected (asymptomatic for prostate cancer, normal digital rectal examination and total PSA < 4 ng/ml). They were recruited in the same

geographical areas as the three Hospitals. They also completed a questionnaire to collect information on occurrence of cancers in their first-degree relatives.

Additional information for both cases and controls included the month and year of blood donation, height & weight at diagnosis/screening, history of smoking at blood donation and family history for prostate cancer in first-degree relatives. Additional information for cases included tumor stage (TNM), tumor grade (Gleason score), tumor histology and death (year, date and whether due to prostate cancer). Cases were matched with controls within each recruiting location and by 5 year interval at age of interview/diagnostic.

Replication Design

A total of 29,018 SNPs were identified for the replication study based on several criteria (Table 1)[10].

A. **Whole genome scan replication, 1-SNP model:** The majority (87.07%) of SNPs were chosen for testing in follow-up based on a single-locus test of association (see "Association Testing" below). The p value cut-off for inclusion was 0.068. In order to maximize the number of regions monitored in replication, a linkage disequilibrium (LD) filter was applied using TagZilla (http://tagzilla.nci.nih.gov/) such that only the most significant SNP in a set of highly correlated SNPs ($r^2 \geq 0.8$) was advanced to replication.

B. **Whole genome scan replication, 2-SNP model**: 1,348 SNPs were chosen based on a 2-SNP model. Two locus tests were selected by taking each of SNP with a marginal $p<0.05$ and constructing a model with each SNP up to 25 SNPs upstream and 25 SNPs downstream within 200 kb. An unadjusted stratified analysis was performed by constructing three 3x3 contingency tables of the phenotype by genotype counts of the non-reference SNP stratified on the reference SNP. This is essentially the same as fitting a logistic model including the genotype effects of both loci, all possible interactions, but no other covariates, and testing for the main effects of the non-reference SNPs and all possible interactions with 12 degrees of freedom. Of the approximately 50 models tested for each reference SNP, the one with the smallest p-value was retained. These 2-SNP models where ranked and the top 1,348 pairs were taken, provided that both SNPs were not already included by the single-SNP selection criteria.

C. **Population substructure**: 1,508 SNPs were chosen to monitor population stratification (see below).

D. **Candidate genes/regions**: To perform fine mapping at the 8q24 candidate region, the region was tagged in HapMap CEU[11,12] using TagZilla (http://tagzilla.nci.nih.gov/) using SNPs with a minor allele frequency (MAF) greater than 5% and $r^2>0.80$; 184 SNPs were identified by this method. Two hundred SNPs were chosen to follow-up a preliminary scan of 91 patients with prostate cancer from the Framingham Heart Study Collaboration[13]. In order to investigate possible associations with interesting candidate genes, 5 genes were fine-mapped using a strategy similar to that which was used for the 8q24 region. Tags were selected from each of 5 genic regions 20 kb 5′ of the start of transcription and 10 kb 3′ of the end of the last exon.

Of the 29,018 SNPs that were originally selected, a total of 26,958 SNPs were successfully genotyped. Of the failed genotype assays, 1,811 (6.24%) failed manufacturing (see below). Specifically, 24,748 SNPs were tested in the single locus replication phase (see below for discussion of SNPs that failed genotyping quality control).

Table 1. SNPs selected for replication

Hypothesis	Ordered	Failed Manufacturing	Failed QC	Final Count
Whole-genome scan replication	26,613	1,668	197	24,748
Single locus (p<0.069)	*25,265*	*1,600*	*180*	*23,485*
Two locus	*1,348*	*68*	*17*	*1263*
Population Stratification	1,508	83	12	1413
Framingham Heart Study Collaboration	200	20	6	174
8q24 Fine Mapping	184	15	4	165
Candidate gene fine mapping	513	25	30	458
ATBF1	*64*	*3*	*0*	*61*
ERG	*105*	*4*	*0*	*101*
KLK3	*16*	*2*	*0*	*14*
MSR1	*222*	*15*	*30*	*177*
RUNX1	*106*	*1*	*0*	*105*
Total	**29,018**	**1,811**	**249**	**26,958**

Sample handling

DNA samples were received from each cohort and visually inspected for adequate fluid in individual tubes. Three measurements of quantification were performed according to the standard procedures at the Core Genotyping Facility of the National Cancer Institute. 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). All samples advanced to genotype analysis completed no less than 13 of the 15 micro-satellite markers. 332 DNAs were aliquoted in duplicate for quality control purposes, and are discussed in the genotype concordance section below. After final review and sample handling, a total of 8,693 DNA samples were selected for genotyping. See Table 3 for a detailed breakdown by study and phenotype.

Genotype Quality control

Assessment of Call Rates

A total of 26,958 SNP genotype assays were attempted on the 8,693 DNA samples using assays drawn from the Illumina HumanHap550. Samples that did not meet a 90% completion threshold were excluded from further analysis. See Table 2 for the number of samples from

each cohort that were excluded based on these criteria. The remaining 8,498 DNAs were retained for the subsequent analyses.

SNPs were excluded for 2 reasons: 1) The assay failed manufacturing at Illumina (n=1,811) or 2) The assay exhibited low completion rate in the laboratory. A total of 249 SNPs failed to provide reliable genotype results due to either no call or low call rates (<90%; see Table 1 for distribution among SNP categories of inclusion).

Table 2. Samples genotyped

	Attempted	Failed QC	Passed QC
CPS-II (blood)	2,820	55	2,765
CPS-II (buccal)	1,128	40	1,088
ATBC	2,021	40	1,981
CeRePP	1,408	30	1,378
HPFS	1,316	30	1,286
Total	8,693	195	8,498

Table 3. Subject counts by study that passed sample-level genotype QC

		Cases			
	Controls	All Cases	Non-aggressive tumor	Aggressive tumor	Unknown stage tumor
PLCO*	1,105	1,180	489	691	0
CPS-II	1,797	1,790	699	926	165
ATBC	940	940	516	240	184
CeRePP	671	671	0	671	0
HPFS	620	619	405	123	91
Replication	4,028	4,020	1,620	1,960	**440**
All	5,133	5,200	2,109	2,651	**440**

* Previously genotyped in the genome-wide scan of > 527,000 SNPs.

Assessment of unique subjects

After removal of sample and locus data due to low completion rates, genotypes for each sample that appeared in duplicate were merged to form consensus genotypes for each study subject; for any observed genotype discordances were henceforth considered as missing observations. Table 3 contains the detailed numbers for each study of the distribution of subjects by phenotype.

Assay concordance

The genotype concordance rate for SNP assays was evaluated using the 332 pairs of known duplicated DNA samples. These pairs of samples were separate aliquots from the same DNA preparation and all met quality control criteria requested for the other samples, thereby, providing reliable data for comparison. An average discordance rate of 0.0058% was observed.

No SNPs or samples were excluded from further analysis as a result of this analysis. Table 4 shows the individual study discordance rates in more detail.

Table 4. Intra-cohort sample genotype concordance

Concordance					Discordance Rate	
Cohort	**Pairs**	**Discordant**	**Concordant**	**Total Comparisons**	**Mean**	**Max**
CPS-II (blood)	110	171	2,998,477	2,998,648	0.0057%	0.267%
CPS-II (buccal)	37	13	1,011,119	1,011,132	0.0013%	0.011%
ATBC	101	105	2,751,688	2,751,793	0.0038%	0.060%
CeRePP	36	134	976,440	976,574	0.0137%	0.213%
HPFS	48	102	1,312,424	1,312,526	0.0078%	0.160%
Total	332	525	9,050,148	9,050,673	0.0058%	0.267%

Concordance analysis also revealed monozygotic twins or individuals enrolled unexpectedly in more than one study. 18 pairs of such subjects were found and were verified to have nearly identical age and phenotypes. Table 5 shows the discordance rates per inter-cohort comparison.

Table 5. Inter-cohort sample genotype concordance

			Genotype Counts			Discordance Rate	
Cohort 1	**Cohort 2**	**Pairs**	**Dis-cordant**	**Con-cordant**	**Total Comparisons**	**Mean**	**Max**
PLCO	CPS-II (blood)	5	69	132,998	133,067	0.0519%	0.181%
PLCO	CPS-II (buccal)	3	9	79,972	79,981	0.0113%	0.015%
PLCO	HPFS	3	9	80,093	80,102	0.0112%	0.011%
CPS-II (blood)	HPFS	5	2	135,973	135,975	0.0015%	0.004%
CPS-II (buccal)	HPFS	2	3	54,423	54,426	0.0055%	0.007%

Hardy –Weinberg Proportions in control DNA

Genotype data were tested for deviation from Hardy-Weinberg proportions using an exact test[14]. The analysis was conducted in each cohort's control group. Significant deviations were observed for an average of 5.34% of all SNPs at the level of $p<0.05$ and 0.39% at $p<0.001$. Table 6 contains the proportion of SNPs per study that deviate from Hardy-Weinberg proportions. None of these SNPs were excluded from analysis since the tests for association applied to such data are valid in the presence of departure from Hardy-Weinberg proportions, although with potentially reduced power when these deviations are due to systematic genotyping errors with equal effects among cases and controls.

Table 6. SNPs exhibiting deviation from Hardy-Weinberg Proportions

	Proportion of SNPs	
	p<0.05	**p<0.001**
PLCO	0.055	0.0047
CPS-II	0.058	0.0046
ATBC	0.052	0.0038
CeRePP	0.052	0.0034
HPFS	0.050	0.0030

Subject exclusions

Subjects with valid genotypes were excluded from analysis based on the following (summarized in Table 7):

1. For unanticipated inter-study duplicates, the subject was excluded from the larger study. An exception was made retaining all duplicate subjects in the PLCO cohort, since their genotypes were already published as part of the initial genome-wide scan.
2. Imputed race, where only subject with imputed European background were retained, regardless of their self-reported race.
3. Sparse groups. Only 2 PLCO subjects were available from one study center and were excluded. Also, only 5 non-aggressive cases were part of the CPS-II buccal group, so they were also excluded.
4. Missing covariates. These subjects were missing one or more covariates necessary for association analysis.

Table 7. Excluded Subjects

	Inter-study Duplicates	Non-European Origin*	Sparse Group	Missing Covariates	Total
PLCO	0	5	2	3	10
CPS-II	8	34	5	5	52
ATBC	0	27	0	3	30
CeRePP	0	27	0	2	29
HPFS	10	19	0	3	32
Replication	18	107	5	13	143
All	18	112	7	16	153

Non-European origin is defined as having less than 0.85 European origin as inferred using the program STRUCTURE[15].

After the exclusion of subjects based on various criteria, the total number of cases and controls for association analyses were 5,064 and 5,116, respectively. Table 8 contains the final subject counts per study for each phenotypic state.

Table 8. Final subject counts for association analysis

		Cases			
	Controls	All Cases	Non-aggressive tumor	Aggressive tumor	Unknown stage tumor
PLCO	1,100	1,175	488	687	0
CPS-II	1,775	1,760	684	913	163
ATBC	921	929	510	238	181
CeRePP	657	656	0	656	0
HPFS	611	596	389	118	89
Replication	3,964	3,941	1,583	1,925	443
All	5,064	5,116	2,071	2,612	443

In order to confirm the validity of the genotype call using the Infinium™ Technology (Illumina Corp., San Diego, CA), notable SNPs were assayed by TaqMan™ (ABI, Foster City, CA) platform in two studies, HPFS and CPS-II (buccal component) to assess the concordance between Infinium™-generated genotype calls (see Figures 1 and 2 for cluster plots for the SNPs for HPFS and CPS-II, respectively) and TaqMan™. Table 9 displays % concordance between Infinium and TaqMan genotypes and GenTrain scores for the Infinium assays[10]. Table 10 contains the primer and probe sequences for each assay.

Table 9. Concordance between Infinium (iSelect) and TaqMan-derived genotypes for the SNPs in Table 1 (main text) and GenTrain Scores

	Concordance (%)		GenTrain Score#	
	CPS-II (buccal)	HPFS	CPS-II (buccal)	HPFS
rs4242382*			0.8065	0.8043
rs6983267*			0.8223	0.8380
rs4430796	99.19	99.57	0.6666	0.6983
rs10993994	98.29	98.89	0.8454	0.8244
rs10896449	98.65		0.8348	0.8831
rs4962416+			0.9039	0.8793
rs10486567	100.00	100.00	0.8917	0.8454
rs12771728	99.04	100.00	0.8111	0.7988
rs4961199	99.23	100.00	0.8965	0.8738
rs4782726	99.33	99.92	0.7701	0.8000
rs4072111	100.00	100.00	0.8074	0.8013
rs6982080	98.75	99.57	0.7972	0.7904

* Previously reported for both Infinium and TaqMan technologies[1]

\+ Assay failed in manufacturing

\# A number from 0 to 1 indicating how well the samples clustered for a particular locus. Score is based on cluster tightness as well as relative separation between clusters.

Table 10. Primer and Probe Sequences for TaqMan Assays

	Forward/Reverse Primer	Fam/Vic Probe
rs4430796	CTCCTTCTCCTTTCTGAAGAAATTCTCA CCTGCCCAATTTAAGCTTTATGCA	ACAGACTGGAGATGC CAGACTGGAAATGC
rs10993994	TCAACCCTCTGTAAACACTCAATGTG CTTTTATACTGGAGAGGGCAACCA	CATTCGACATCATTGG CATTCGACGTCATTGG
rs10896449	CATAGCTGAAAGATGAGGAGTCAAGA CCAAGGTTCAGCCTCATCTCTT	CCCTTCTTAAACTTTCAG CCCTTCTTAAATTTTCAG
rs10486567	CCTTTCCTGCTTTAATCATAGCCAAAT CATTAAAACTGTAAAGCCCAAGCATCT	CTCTAAGTTTAAGTATCCC CTCTAAGTTTAAATATCCC
rs12771728	GGTTTGGGAGTCAGCCAGATG GGCCAAGAGACCACATGGA	CTGAATTCTATCCATGGCCA CTGAATTCTATCCGTGGCCA
rs4961199	AGTTAGCAAAAGCAAATTAATGTAGTTTGACAA ACCTATAAGCTGTCCTCTGGTCTT	TAGAGCTGGGCATTT AAATAGAGCTGGACATTT
rs4782726	ATGATGTTTGGTGCAGTCTTGAAAG AGAAAAGCTCAGGACATAGAAAGCA	CGCAGTCACTGTGGAG CGCAGTCACTATGGAG
rs4072111	CACGATCCTGAGTCACTGTGAT GGGTCTGGATGTCGGCTAAC	TCCAGTCTCCATCATT CCAGTCCCCATCATT
rs6982080	CAGGGAAGTCCATTTCTTCAACTGA GAGGACCCTGCAGTAACAGTAA	CAGACTTCAAGTCTCG TCAGACTTCACGTCTCG

Figure 1. Illumina Infinium HumanHap550 genotype clustering for HPFS cohort

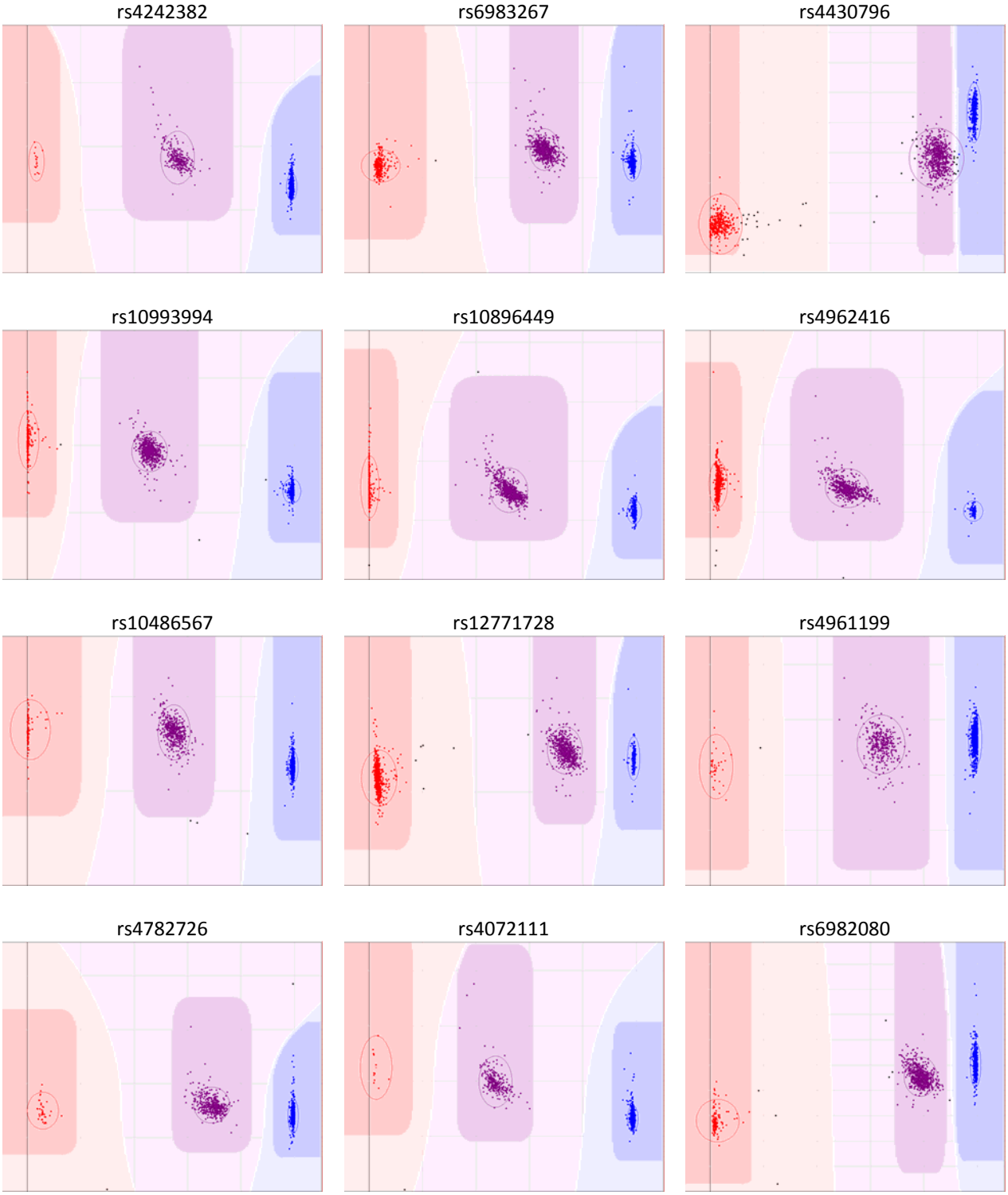
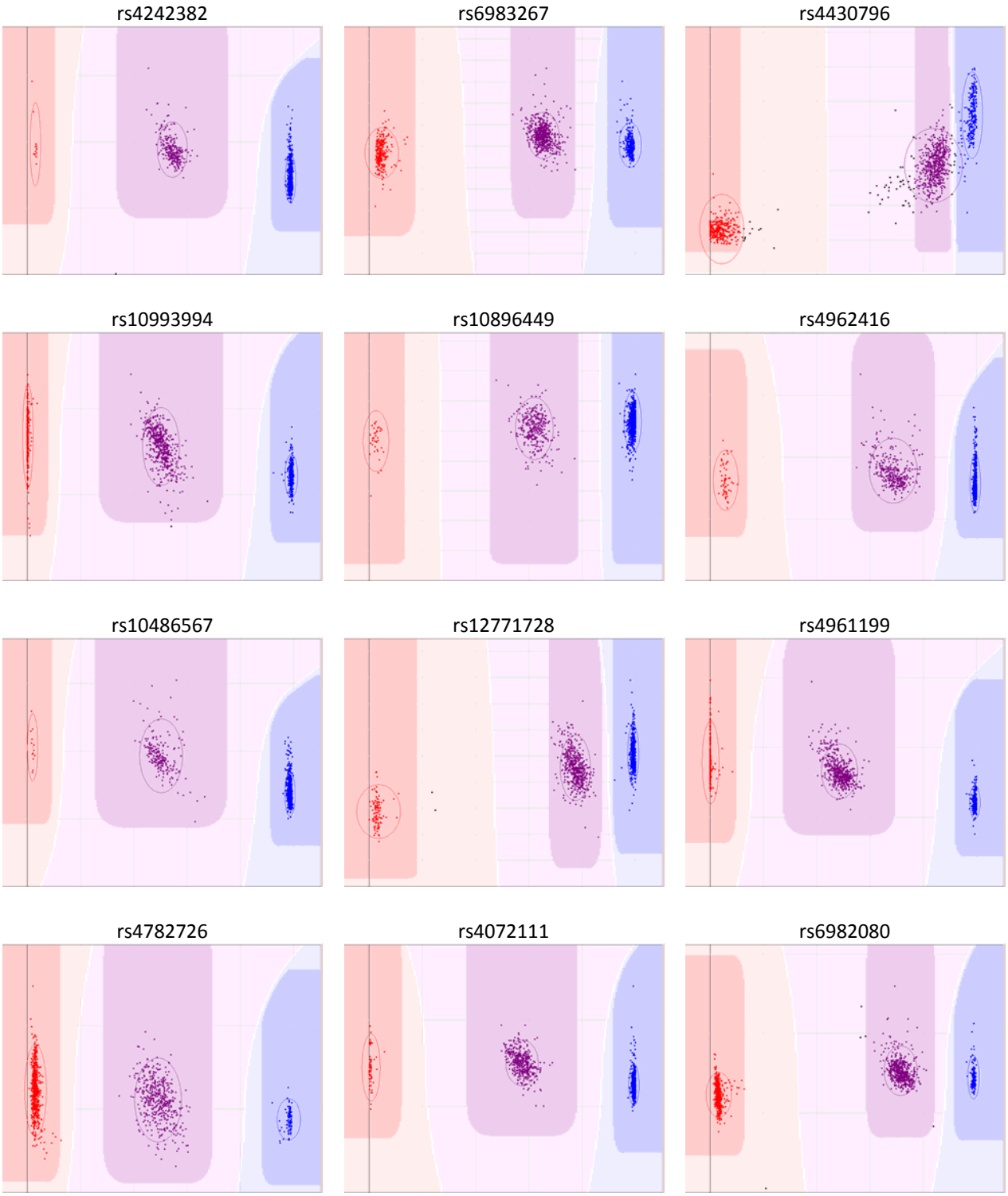


Figure 2. Illumina Infinium HumanHap550 genotype clustering for CPS-II (buccal) cohort

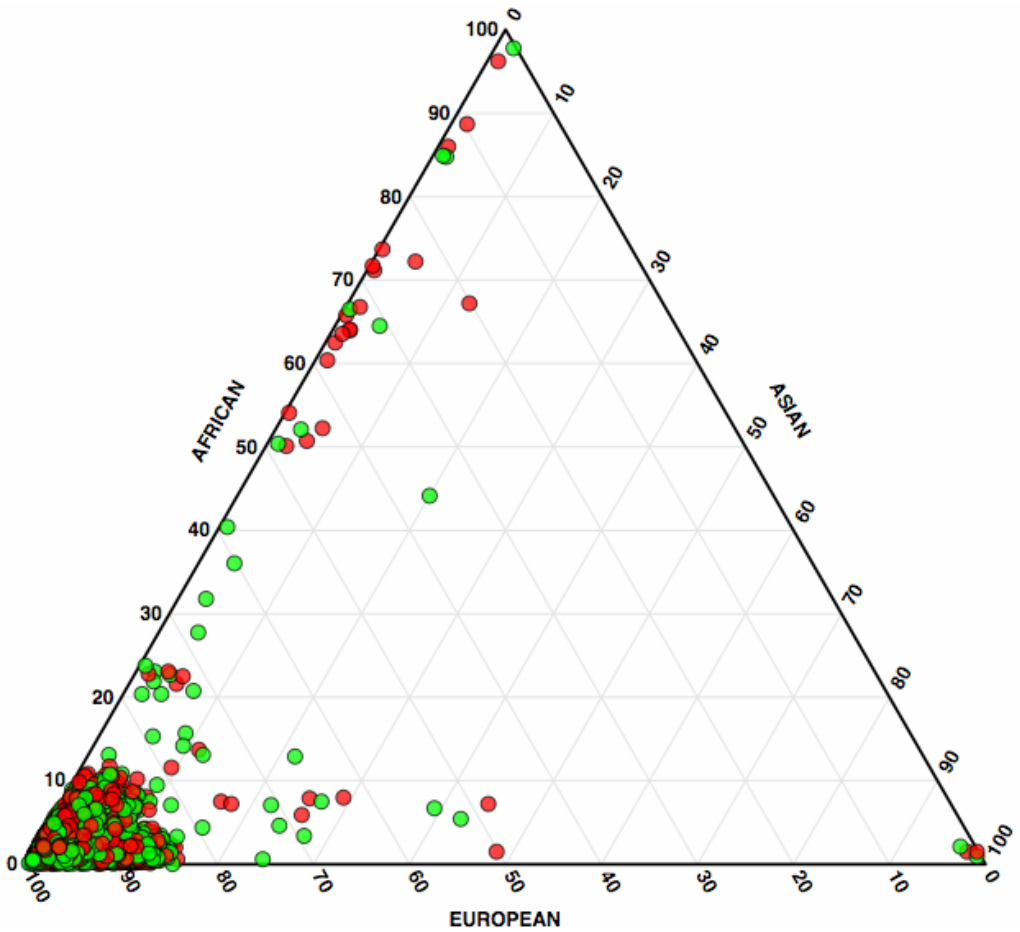


Assessment of Population Structure

Two independent sets of SNPs were used for the detection of population stratification in the four follow up samples. The first set of pre-selected 1,508 SNPs had been specifically chosen for typing in the follow-up studies for this purpose and were selected because of completion rates higher than 90%, a residual linkage disequilibrium r^2 value lower than 0.1 for any pair of SNPs less than 200 kb apart and enabling a stable detection of the first two principal components of population structure identified in the PLCO study[16]. The second set of 10,693 SNPs was selected from all of the SNPs typed in the follow-up studies because of indication of their possible associations in the genome-wide scan. The selection was based upon a completion rate higher than 90% and residual linkage disequilibrium r^2 value lower than 0.0075 for any pair of SNPs less than 200 kb apart.

In an attempt to maximize genetic homogeneity, subjects with significant estimated non-European ancestry were excluded from analysis. This estimation was done using the STRUCTURE program by merging the genotypes from the follow-up studies with those of the reference HapMap population downloaded from the HapMap web site. The number of clusters (the "k" parameter) was set to three and the CEU, YRI and JPT+CHB samples were each specified to a different cluster schematically representing populations of European, African and Asian origin respectively. The CGEMS samples were left unspecified. A total of 112 subjects were estimated to have less than 85% European ancestry and were excluded from analysis (Figure 1). All individuals that had at greater than 85% European ancestry were retained for the replication study, regardless of their reported origin (Table 7).

Figure 3. Admixture plot



We defined the imputed origin as follows: a subject is considered to have a single continental origin if the estimated admixture portion was greater than 85% with any single seed population. Subjects were labeled "admixed" if their estimated proportion of admixture was greater than 15% with only one population, but less than 85%. Subjects with greater than 15% estimated admixture with more than one seed population was labeled as the concatenation of all such seed populations.

Table 11. Cross-tabulation of reported versus imputed race

		Cohort Reported Origin				
STUDY	Imputed origin	Caucasian	African American	Asian	Hispanic	Unknown
PLCO	Caucasian	**2,277**				
	Admixed Caucasian	2				
	African + Caucasian	1				
	Asian + Caucasian	2				
CPS-II	Caucasian	**3,530**	2	1	1	1
	Admixed African		1			
	Admixed Caucasian	7			2	
	African		**2**			
	Asian			**2**		
	African + Caucasian	4	6			
	Asian + Caucasian	4			4	2
ATBC	Caucasian	**1,850**				
	Admixed Caucasian	25				
	Asian + Caucasian	2				
CeRePP	Caucasian	**1,313**				
	Admixed Caucasian	1				
	African	3				
	African + Caucasian	19				
	Asian + Caucasian	3				
	African + Asian + Caucasian	1				
HPFS	Caucasian	**1,143**				64
	Admixed Caucasian	4				2
	Asian					2
	African + Caucasian					8
	Asian + Caucasian	2				1

To adjust for any remaining finer-grained population differences, a principal component analysis of all DNAs used in this study was performed using the EIGENSTRAT program[16]. Both sets of SNPs detected essentially the same three most significant principal components. The correlation coefficients of the DNA coordinates along these directions were 0.84, 0.78 and 0.32. The fourth component was not correlated. As the density of the clustering appeared more clearly with the second set of SNPs, this set was used for correction of population stratification using four components. The use of one additional component, as compared to the analysis that was done for the first stage (PLCO study), which used three components, accounts for the distinct difference of the Finnish subjects with respect to the rest of the cases and controls[1]. Thus four eigenvectors were included as quantitative covariates when testing for association between prostate cancer phenotypes and SNPs typed as part of the replication scan.

Association Analysis

The primary analysis of the CGEMS prostate GWAS study explores the association between single SNPs and prostate cancer susceptibility in 26,985 SNPs per subject overall. 25,265 SNPs were selected based on single-SNP association tests and 1,348 SNPs were selected based on two-SNP association tests, in addition to the 1,508 SNPs chosen to aid in the characterization of population structure. Additional SNPs were chosen to explore high-profile candidate genes, fine-mapping, and non-CGEMS hypotheses and will not be discussed here.

For the follow-up replication studies, all one- and two- SNPs analyses were conducted using unconditional logistic regression, adjusted for age (in ten-year categories), study, and center for the two studies in which center for recruitment was available (i.e., PLCO and the French Prostate Case Control). We analyzed each study separately, the pooled replication studies (all except PLCO), and all studies combined.

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 the Cochrane-Armitage trend test[17,18] 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.

Genetic Models

In order to maintain high power to detect SNPs that are involved in non multiplicative models (such as complete recessive 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 in the initial genome-wide scan, a single degree of freedom is used per case phenotype.

Design of follow-up studies

We compared two follow-up designs to evaluate our ability to identify SNPs that would demonstrate the genome wide significance threshold of 10^{-7} based on our available resources. The initial genome wide scan had been performed on 1,150 pairs of cases and controls; an additional 8,800 pairs of cases and controls were available for follow-up. Follow-up of the initial scan was simulated either as a single stage, which would involve the typing of all SNPs that reached a p-value lower than 0.068 at the initial screen on all 8,800 case-control pairs, or as two stages, with a first stage involving the typing of the same SNPs on 3,800 pairs of cases and controls followed by a second stage which would type only those SNPs that reached a p value lower than 10^{-3} on the remaining 5,000 pairs.

Simulation was performed under the following assumptions: the frequency of the at-risk allele is uniformly distributed between 0.05 and 0.95; odds ratio, OR, of the at-risk homozygote varied between 1.1 and 2 (Table 12). The heterozygote was assumed to have an OR equal either to 1 (recessive model), or to the square root of the OR of the at-risk homozygote (multiplicative model of risk – i.e. additive on the log OR scale) or to the OR of the at-risk homozygote (dominant model).

For each model, genotypes were randomly generated based on their expected frequency and p values were computed on the merged genotypes that would have been available at the completion of each phase. A two-degree of freedom $\chi 2$ was used when all 6 cells had more than 5 individuals, otherwise the Fisher exact test was used. For each evaluation, 10,000 simulations were done using the Stata9 software package.

We present estimates and 95% confidence limits of the absolute power loss. The 2-stage follow-up test can only reject a hypothesis when the 1-stage follow-up test also rejects the hypothesis. Therefore, we first estimate the proportion not rejected by the 2-stage follow-up test among those hypotheses rejected by the 1-stage follow-up test. To calculate the 95% CI for absolute power loss, we first calculate the 95% CI for the proportion of hypotheses rejected by the 1-stage follow-up test that are not rejected by the 2-stage follow-up test, then multiply the CI limits by the power of the 1-stage follow-up in order to reflect the absolute power loss rather than the power loss conditional on rejection of the 1-stage follow-up test. These confidence limits are slightly too narrow, as they are conditional on the observed number rejected by the 1-stage test, because under the condition, the observed number rejected by the 2-stage test is binomial.

Table 12. Power estimation in a 1-stage and 2-stage follow-up strategy

"at risk" Homo zygote Odds Ratio	Recessive model			Additive model			Dominant model		
	1-stage follow-up	2-stage follow-up	power loss (95%CI)	1-stage follow-up	2-stage follow-up	power loss (95%CI)	1-stage follow-up	2-stage follow-up	power loss (95%CI)
1.10	0.2%	0.2%	0.0% (0.0-0.0)	0.0%	0.0%	0.0% (0.0-0.0)	0.2%	0.2%	0.0% (0.0-0.0)
1.15	4.4%	4.0%	0.4% (0.3-0.5)	0.6%	0.5%	0.1% (0.0-0.1)	4.9%	4.6%	0.4% (0.3-0.5)
1.20	21.3%	20.0%	1.3% (1.1-1.5)	3.5%	3.1%	0.4% (0.3-0.5)	19.6%	18.7%	0.8% (0.7-1.0)
1.25	38.1%	37.3%	0.8% (0.6-1.0)	13.9%	12.8%	1.2% (0.9-1.4)	36.6%	35.7%	0.9% (0.7-1.1)
1.30	51.3%	50.8%	0.5% (0.4-0. 60)	29.6%	28.1%	1.5% (1.2-1.7)	50.0%	49.4%	0.6% (0.4-0.7)
1.40	68.4%	68.1%	0.3% (0.2-0.4)	55.6%	54.7%	0.9% (0.7-1.1)	66.0%	65.7%	0.3% (0.2-0.4)
1.50	78.3%	78.1%	0.3% (0.2-0.3)	72.5%	72.1%	0.4% (0.3-0.5)	73.6%	73.4%	0.2% (0.1-0.3)
1.60	83.0%	82.8%	0.2% (0.1-0.3)	83.0%	82.7%	0.3% (0.2-0.4)	78.5%	78.3%	0.2% (0.1-0.3)
1.80	88.8%	88.7%	0.1% (0.1-0.2)	93.8%	93.6%	0.2% (0.1-0.2)	83.8%	83.7%	0.1% (0.1-0.2)
2.00	91.6%	91.5%	0.1% (0.0-0.2)	96.9%	96.9%	0.1% (0.0-0.1)	86.4%	86.3%	0.1% (0.1-0.2)

Thus, if the "at risk" homozygote genotypic odds ratio was 1.5, we estimate a loss of power of 0.5% or less under the 3 models shown. This represents the statistical "cost" of the 2-stage follow-up versus the 1-stage design follow-up.

References

1. Yeager, M. et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).
2. 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).
3. Prorok, P.C. et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials* **21**, 273S-309S (2000).
4. Hayes, R.B. et al. Methods for etiologic and early marker investigations in the PLCO trial. *Mutat Res* **592**, 147-54 (2005).
5. Chen, Y.C. et al. Sequence variants of Toll-like receptor 4 and susceptibility to prostate cancer. *Cancer Res* **65**, 11771-8 (2005).
6. Group., T.A.C.P.S. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Ann Epidemiol* **4**, 1-10 (1994).
7. Calle, E.E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490-501 (2002).
8. Feigelson, H.S. et al. Determinants of DNA yield and quality from buccal cell samples collected with mouthwash. *Cancer Epidemiol Biomarkers Prev* **10**, 1005-8 (2001).
9. Valeri, A. et al. Segregation analysis of prostate cancer in France: evidence for autosomal dominant inheritance and residual brother-brother dependence. *Ann Hum Genet* **67**, 125-37 (2003).
10. Chanock, S.J. et al. Replicating genotype-phenotype associations. *Nature* **447**, 655-60 (2007).
11. The International HapMap Project. *Nature* **426**, 789-96 (2003).
12. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
13. Murabito, J.M. et al. A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S6 (2007).
14. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887-93 (2005).
15. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).
16. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
17. Agresti, A. *Categorical Data Analysis*, (John Wiley & Sons, Inc., New York, 1990).
18. Margolin, B.H. Test for Trend in Proportions. in *Encyclopedia of Statistical Sciences*, Vol. 9 (eds. Klotz, S. & Johnson, N.L.) 334-336 (John Wiley & Sons, Inc., New York, 1988).