Published in final edited form as:

Genet Epidemiol. 2011 September ; 35(6): 506–514. doi:10.1002/gepi.20600.

Evaluation of polygenic risk scores for predicting breast and prostate cancer risk

Mitchell J. Machiela, MPH^{#1}, Chia-Yen Chen, MS^{#1}, Constance Chen¹, Stephen J. Chanock, M.D.², David J. Hunter, MBBS ScD¹, and Peter Kraft, PhD¹

¹ Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts

² Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

These authors contributed equally to this work.

Abstract

Recently, polygenic risk scores have been shown to be associated with certain complex diseases. The approach has been based on the contribution of counting multiple alleles associated with disease across independent loci, without requiring compelling evidence that every locus had already achieved definitive genome-wide statistical significance. Whether polygenic risk scores assist in the prediction of risk of common cancers is unknown. We built polygenic risk scores from lists of genetic markers prioritized by their association with breast or prostate cancer in a training data set and evaluated whether these scores could improve current genetic prediction of these specific cancers in independent test samples. We used genome-wide association data on 1,145 breast cancer cases and 1,142 controls from the Nurses' Health Study and 1,164 prostate cancer cases and 1,113 controls from the Prostate Lung Colorectal and Ovarian Cancer Screening Trial. Ten-fold cross validation was used to build and evaluate polygenic risk scores with 10 to 60,000 independent single nucleotide polymorphisms (SNPs). For both breast and prostate cancer, the models that included only published risk alleles maximized the cross-validation estimate of the area under the ROC curve (0.53 for breast and 0.57 for prostate). We found no significant evidence that polygenic risk scores using common variants improved risk prediction for breast and prostate cancer over replicated SNP scores.

Keywords

single nucleotide polymorphisms; genome-wide association study; human genetics

Introduction

Since 2007, a wealth of genomic data investigating the germline contribution to cancer risk has been collected from genome-wide association studies (GWAS)[Ioannidis, Castaldi and

Evangelou, 2010]. While this rapidly expanding repository of germline genetic information has been instrumental in discovering disease-associated loci not previously identified by linkage nor candidate gene approaches, it is remarkable that nearly all variants discovered so far confer relatively low risk (e.g., estimated odds ratios less than 1.3 per risk allele). This observation coupled with the small sample sizes in first generation GWAS has resulted in underpowered studies with limited ability to detect low risk variants at stringent significance levels required to compensate for multiple testing. The current state of dissecting the underlying genetic factors that contribute to most complex diseases is far from complete with a substantial portion of heritability remaining to be explained.

Although other forms of genetic variation not tagged by single nucleotide polymorphisms (SNPs) are expected to explain the proportion not explained by GWAS with common variants, [Manolio et al., 2009], the possibility remains there is unutilized information in GWAS data. In fact, a multitude of loci may be awaiting discovery, which current studies are underpowered to detect [Park et al., 2010]. Refined methods for exploring the available genomic information are needed to maximize the information that can be gleaned from GWAS.

Evans [Evans, Visscher and Wray, 2009] and Purcell [International Schizophrenia Consortium et al., 2009] have proposed methods of aggregating information on a large number of SNP alleles associated with a trait that do not achieve stringent genome-wide statistical significance or even nominal statistical significance of $p < 0.05$. These models create polygenic risk scores by summing risk alleles from thousands or tens of thousands of loci spanning the genome to predict an individual's genetic risk of developing disease. Moreover, these scores could be used to predict an individual's outcome even without knowing which of the SNPs in the score are conclusively associated with disease.

Evans et al. have shown that the discriminatory ability of risk models using both known variants and a polygenic score is higher than a model using known variants only for bipolar disorder, coronary heart disease, hypertension, Crohn's disease, rheumatoid arthritis, and type I and II diabetes. Purcell et al. have demonstrated that polygenic scores for schizophrenia and bipolar disorder can explain up to 3% of disease variance, and that these scores are specific to those diseases, since they do not explain variance in other non-psychiatric disorders. Recently, a multiple sclerosis consortium also showed evidence they could explain up to 3% of the variance in multiple sclerosis with a polygenic risk score [International Multiple Sclerosis Genetics Consortium (IMSGC) et al., 2010]. For these phenotypes, the variance explained by a polygenic risk score adds to that already accounted for by known genetic variants. This suggests there are additional risk loci beyond those that have achieved genome-wide significance. Moreover, the variance explained by not-yet-known markers may be underestimated by these analyses due to the noise associated with including potentially non-associated markers into the polygenic risk score.

Although an array of complex chronic diseases have been investigated using polygenic techniques, to our knowledge, only one study applied polygenic methods to existing GWAS studies of cancer [Witte and Hoffmann, 2011]. While that study found a modestly significant association for polygenic risk scores at liberal SNP inclusion thresholds for PCa ($p < 0.01$),

the authors did not fully investigate the added variance that can be explained by polygenic risk scores that is independent of established genome-wide significant SNPs.

Our goal is to investigate currently available GWAS to examine genetic models for improving risk prediction for breast cancer (BCa) and prostate cancer (PCa) and also to investigate whether additional undiscovered disease susceptibility loci may be awaiting discovery. So far, the largest number of risk variants identified through GWAS have been for PCa and then BCa. To date, at least 13 independent loci have been associated with BCa[Thomas et al., 2009, Stacey et al., 2007, Ahmed et al., 2009, Stacey et al., 2008, Gold et al., 2008, Zheng et al., 2009b, Hunter et al., 2007, Easton et al., 2007, Turnbull et al., 2010] and over 30 independent genetic loci have been associated with PCa risk[Duggan et al., 2007, Eeles et al., 2009, Eeles et al., 2008, Thomas et al., 2008, Gudmundsson et al., 2007b, Gudmundsson et al., 2008, Sun et al., 2008, Gudmundsson et al., 2009, Yeager et al., 2009, Zheng et al., 2009a, Gudmundsson et al., 2007a]. Our first aim is to investigate the discrimination of models using established SNPs for prediction. Next, we aim to employ polygenic methods to compile prioritized lists of variants to be used in constructing polygenic risk scores. We further explore the added discriminative ability of the polygenic scores and assess whether there is significant information added to the overall prediction of BCa and PCa by these scores.

Methods

Data for this study are based upon the Cancer Genetic Markers of Susceptibility (CGEMS) initiative. Breast cancer genetic data originated from the Nurses' Health Study (NHS) and prostate cancer data came from the Prostate, Lung, Colorectal, and Ovarian (PLCO) study[Hunter et al., 2007, Yeager et al., 2007].

Briefly, the NHS is a prospective cohort study established in 1976 with a total of 121,700 women recruited at baseline. During 1989 to 1990, 32,826 participants without previous BCa diagnosis provided a blood sample and were followed until June 1, 2004 for incident BCa. These participants were genotyped using the Illumina HumanHap500 array. Breast cancer cases were identified through personal mailings and searches of the National Death Index. Over 98% of BCa cases were ascertained in this cohort[TwoRoger et al., 2007]. Controls were individually matched with cases on age at diagnosis, timing of blood collection, use of postmenopausal hormones, ethnicity, and menopausal status. All cases and controls were self-reported Caucasian and postmenopausal at diagnosis. After quality control, 1,145 BCa cases and 1,142 controls were used in the following analyses[Hunter et al., 2007]. All study participants provided informed consent and the study was reviewed by the Institutional Review Board of the Brigham and Women's Hospital, Boston, MA.

The PLCO Cancer Screening Trial, is a large randomized control trial designed to investigate the efficacy of cancer screening on early mortality from prostate, lung, colorectal, or ovarian cancer[Gohagan et al., 2000, Prorok et al., 2000]. Blood samples were collected from participants at screening visits and DNA extracted according to standard protocols. Incident PCa cases were identified by screening exams, self or physician report, or link with the state cancer registries or the National Death Index. All cases of PCa were

pathologically confirmed by a trained pathologist. Genotypes for non-Hispanic white men randomized to the PCa screening arm of the PLCO were obtained using Illumina HumanHap300 and HumanHap240 genotyping platforms at separate times as explained elsewhere [Yeager et al., 2007]. A total of 1,164 PCa cases oversampled for aggressive disease and 1,113 PCa free incidence-density sampled controls were eligible for inclusion into the study. Written informed consent was required for participation in the study and all protocols were reviewed by the Institutional Review Boards of both the National Cancer Institute and each of the 10 participating study centers.

A total of 528,173 SNPs were genotyped and passed quality control for BCa and 546,593 SNPs were genotyped on the combined platforms for PCa [Hunter et al., 2007, Yeager et al., 2007].

Previously replicated genome-wide significant SNPs for BCa and PCa were extracted from the literature (Table I). These SNPs were moved to the top of a minor allele frequency ordered SNP list for each cancer, which was further thinned by using linkage disequilibrium filtration to remove redundant SNPs with correlation of an R^2 greater than 0.20. The thinning program was an in-house Python script that referenced the HapMap Rel.23a (NCBI B36) database. The result was a list of 161,702 LD thinned SNPs for BCa and 165,508 LD thinned SNPs for PCa.

For purposes of cross-validation, cases and controls for each respective disease were randomly divided into 10 approximately equal subgroups. Cross validation was performed by setting aside one subgroup for the purpose of testing, while using the remaining 9 subgroups (training set) to construct a model to test in the testing subset. This procedure was repeated 9 additional times with a different subgroup set aside for testing.

Models were constructed using the list of LD filtered SNPs. The list was divided into a list of previously replicated SNPs and a list of non-replicated SNPs. Associations were calculated for each non-replicated SNP on disease outcome in the training set using an allelic association analysis in PLINK [Purcell et al., 2007]. Risk alleles for each non-replicated SNP were determined from the association test. These non-replicated SNPs were then ordered by increasing association p-value and broken into lists of the top $k=10$ to the top $k=60,000$ SNPs with increasing interval size as k increases for inclusion into the polygenic risk score. Replicated SNP scores (RSS) and polygenic risk scores (PRS) were constructed separately. Risk scores were calculated as the sum of the number of risk alleles an individual carried at each locus summed across all loci included in the list. This assumes an additive model for each known SNP with equal weight applied to each locus and ignores any interactions that may exist between SNPs. For each cross validation set, PLINK was used to calculate one risk score for the replicated SNPs and a separate risk scores for the polygenic SNPs. This procedure was continued repeatedly for increasing k until the top 60,000 SNPs showing the strongest evidence for association with cancer were included into the PRS.

Unconditional logistic regression models were fit to the data in the training set with case status as the outcome and two predictor variables, normalized RSS and normalized PRS.

Scores were normalized by subtracting the mean and dividing by the standard error. Models were fit for each cross-validation training set for each k . The same models and parameter estimates from the training sets were used to define predicted disease probability in the testing set. Estimates of the area under the receiver-operating characteristic curve (AUC) were obtained for training and testing sets from each of the 10 cross-validation sets. Results were averaged over the 10 cross-validation sets to determine a final estimated average AUC for the testing sets. Data manipulation and statistical analyses were carried out using R 2.8.0 (R Foundation for Statistical Computing, Vienna, Austria) and SAS 9.1 (SAS Institute Inc., Cary, NC) on a UNIX platform.

Results

Breast Cancer

Thirteen SNPs that were found to be robustly associated with breast cancer in previous studies were used to construct the RSS here. Background information regarding the replicated SNPs is in Table I. For the entire dataset, the RSS ranged from 4 to 19 with an average of 11.6 for the cases and 10.8 for the controls ($p=5.83 \times 10^{-17}$). Logistic regression estimated an OR of 1.16 (95% CI: 1.12-1.20) for each additional risk allele in the RSS with an AUC of 0.59 and a Nagelkerke pseudo- R^2 of 0.039.

For each of the 10 cross validation sets, PRSs were constructed with increasing number of previously non-replicated SNPs added to predict the risk of BCa. We observed no statistically significant BCa association ($p>0.05$) between any PRS in any test set (Figure 1). Additionally, the mean AUC from the 10 cross validation sets were calculated for both the training and testing sets (Figure 2). For the training sets, average AUCs increased rapidly when the number of SNPs included in the score increased and approached 1.0 when 1,000 SNPs were included in the polygenic score. This phenomenon reflects overfitting in the training set and is used by algorithms to query whether a particular individual is in a genome-wide association study [Jacobs et al., 2009]. For testing sets, average AUCs initially dropped when the PRS was added to the RSS model and then remained relatively stable as the number of SNPs in the risk score increased. The average AUCs in the testing set ranged between 0.50 and 0.53 when both the RSS and GRS were in the model.

Prostate Cancer

A total of 30 independent SNPs associated with PCa were extracted from the literature. The SNPs, their genetic location, and risk alleles are in Table I. Individual PCa RSS ranged between 12 and 38. On average, cases carried 25.0 risk alleles and controls carried 23.5 ($p=2.51 \times 10^{-20}$). A logistic regression model fit to the entire dataset with the RSS as a predictor and PCa status as outcome indicated the RSS is a significant predictor of PCa status with every one unit in risk score associated with a 1.12 increase in the odds of incident PCa (95% CI: 1.09-1.14). The resulting model had an AUC of 0.614 and a Nagelkerke pseudo- R^2 of 0.051.

Ten cross validation datasets were produced to observe the effect of including additional nominally significant SNPs to a PRS, adjusting for the known associated SNPs in the RSS.

None of the PRS was statistically significantly associated with prostate cancer in the testing sets (Figure 1). For AUC, the overarching pattern observed in the training sets was that of a sharp increase as additional SNPs were added to the PRS, which leveled off near an AUC of 1.0 at 1,500 SNPs. In the test sets, the primary pattern observed was a notable decrease in discrimination followed by minimal change as additional loci were included (Figure 2). The maximum AUC in the testing set of 0.569 was obtained at the first iteration when 10 loci were included into the PRS. After the inclusion of 60,000 SNPs into the PRS (approximately 35% of the SNPs in the thinned dataset) the testing set had an AUC of 0.564.

Discussion

Our results did not demonstrate that the approach of polygenic risk scores provides a more robust discrimination of risk for either breast or prostate cancer using common SNPs in a current single GWAS. The shift of the AUC was limited (less than 0.60) even when common SNPs not achieving genome-wide association were added to the score. Moreover, we did not find any evidence that polygenic risk scores were significantly associated with breast or prostate cancer risk in our sample sets. This suggests that the polygenic contributions to breast and prostate cancer beyond the known loci are subtle, and will require larger sample sizes or more sophisticated analytic approaches to discover.

In our study, it is not surprising the replicated SNPs performed better than the polygenic SNPs. The replicated SNPs are a cleaner set of signals that has been robustly replicated across several independent sample sets. These SNPs likely represent surrogates in high linkage disequilibrium with the variants that biologically contribute to specific cancer risk. Conversely, the genetic signal from the polygenic SNPs includes considerable noise, as a result of including genetic variants that are false positive markers in the PRS. While prioritizing SNPs by association p-value favors inclusion of genetic loci that are more likely to have true associations with disease causing loci, chance distribution of SNPs across the association p-value distribution allows for SNPs to be included in the PRS that are not associated with disease in the specific studies being tested. In this regard, the PRS is limited by the problem of a low signal-to-noise ratio in which, with the inclusion of each additional set of SNPs into the polygenic score, there is a possibility that additional disease-associated signal can overcome the high background, non-associated noise. The signal-to-noise ratio may be particularly problematic when the training sets are small (due to low power), or when the overall number of associated markers is small: e.g. several score instead of hundreds or thousands. Notably, using nonparametric modeling on existing GWAS data Park and colleagues [Park et al., 2010] estimate that the total number of breast, prostate, and colorectal cancer risk loci with effect sizes similar to those already discovered is 67 (95% CI 31, 173).

Our study avoided design artifact—a particular concern for these polygenic analyses [Evans, Visscher and Wray, 2009, International Schizophrenia Consortium et al., 2009]—by carrying out analyses on a homogenous population with European ancestry. Cases and controls were randomly distributed across genotyping plates, guarding our study from spurious genotyping artifacts between cases and controls. We included only high quality genotyped SNPs that passed a rigorous quality control assessment when constructing the

PRSs, thus eliminating misclassification due to cross platform differences in performance. In addition, by using only genotyped SNPs, we avoided the uncertainty and error implicit in analyzing imputed genotypes into a PRS. We LD filtered the genotyped SNPs to remove an increase or decrease in discriminative ability that may be due to adding highly correlated redundant markers. Such an addition of an area of high LD to a PRS could result in substantial changes in the PRS. Furthermore, ten-fold cross validation was implemented in our dataset to minimize any potential chance findings or overtraining and improve external validity of these findings to other data sets. As evidenced by Figure 2, overtraining is a major concern when a large number of SNPs are compiled into a risk score.

The method implemented to compile risk scores in our dataset consisted of summing the number of risk alleles an individual carries. This method gives equal weight to all loci included in the risk score, regardless of the locus' respective strength of association. Other methods of weighting the loci such as using the log of the odds ratio are possible, but produced little differences in our dataset (results not included) as well as in others [Evans, Visscher and Wray, 2009]. Such a weighting method would require large sample sizes to obtain more refined odds ratio estimates used as weights and ignores uncertainty in the estimated odds ratio. Because most odds ratio estimates are near one, we chose to present the risk allele counting method.

The AUC was the primary metric used to assess the discriminative ability of the resulting polygenic models because it can be used to assess the probability that a model can correctly distinguish a case from a control in a pair of individuals. The measure is also a summary estimate for the entire model and gives a quantitative estimate of how well the included parameters of a model are separating out cases from controls. By averaging estimates of AUC over the cross-validation testing sets, an informative metric was obtained in which to track the change in discriminative ability of each addition to the PRS. Other metrics such as R^2 or net reclassification indexes could have been used. We chose the AUC primarily for ease of interpretation—AUC values can directly compare samples with different case to control ratios, where R^2 and reclassification indexes cannot—and its close relationship to the sensitivity and specificity of a test.

In general, the results were similar for both BCa and PCa when the pool of SNPs was expanded in the PRS. Overall, in the testing sets as additional common variants (not achieving genome-wide significance) were included in the PRS there was a pattern of a noticeable drop in mean AUC initially followed by little change thereafter. Even at early intervals for the PRS, it appears more noise than signal is added to the model as evidenced by the noticeable decrease in discriminative ability of the overall model and insignificant PRS p-values in the testing set. After the initial decline, there seems to be little change in discriminative ability and PRS p-values remain insignificant. This indicates the model is saturated with noise from the PRS and any additional inclusion of SNPs into the risk score has little effect on the overall prediction ability of the model. Here, any discrimination of the model is primarily being driven by the RSS. It is noteworthy that some of the BCa and PCa loci used in the RSS were discovered in the sample sets we used. This may result in an artificial inflation of the AUC estimate for the RSS.

Although our study used the same CGEMS dataset accessed by Witte et al. [Witte and Hoffmann, 2011], our analysis produced slightly different conclusions for prostate cancer. Qualitative differences in analytic parameter settings within the polygenic risk score framework may be responsible for the observed differences. For example, Witte et al. chose not to remove known PCa SNPs from the polygenic risk score, whereas in our approach known PCa SNPs were removed and adjusted for as a separate parameter (RSS). Failing to adjust for known loci may result in known signal being detected in the PRS and, therefore, differences in overall conclusions surrounding the added significance of the PRS. The technique used to split the data into training and testing sets also differed. Witte et al. employed a 50:50 training to testing set split, while our analysis used a 90:10 split. The 50:50 split method is less powered to select SNPs for inclusion in the PRS, while the 90:10 method loses power to detect a PRS-trait association. However, when we repeated our analyses using a 50:50 split we did not see evidence for an association between GRS and either BCa or PCa ($p > 0.05$ for all thresholds, detailed results not shown). Additionally, we pruned for linkage disequilibrium at an R^2 threshold of 0.20 and Witte et al. used a threshold of 0.50. Our conservative method for pruning may lose power by removing modestly tagging SNPs from the PRS, while less stringent LD pruning may lose power by upweighting large regions of high LD in the PRS that are not associated with disease. The above mentioned differences are all analytic settings for which there are currently no theoretical guidelines. The resulting slight differences in analytical conclusions suggest polygenic signals for PCa are likely weak since minor differences in analytic preferences produce conflicting qualitative results.

Our study had potential limitations related to power and the lack of a comprehensive set of common variants. The sample size of our analysis consisted of 1,145 cases and 1,142 controls for BCa, and 1,164 cases and 1,113 controls for PCa. In comparison to other studies that found significant associations of PRS with complex diseases, our total sample sizes were smaller. The Evans et al. study investigated the contribution of polygenic risk scores to several diseases with sample sizes ranging from 1,748 to 1,963 cases and a set of 1,480 common controls [Evans, Visscher and Wray, 2009]. Purcell et al. used a schizophrenia sample of 3,322 cases and 3,587 controls [International Schizophrenia Consortium et al., 2009]. The International Multiple Sclerosis Genetics Consortium used a training sample of 931 cases and 2,431 controls and a test sample of 806 cases and 1,720 controls [International Multiple Sclerosis Genetics Consortium (IMSGC) et al., 2010]. All of these studies found significant associations between PRS and disease. With a reduced sample size, our study had limited power to detect true signals and prioritize these signals for inclusion into the PRS. Furthermore, splitting the dataset into 10 equally-sized cross validation sets leaves a restricted number of cases and controls for our testing set and may explain some of the differences in the PCa results between our study and the study by Witte and Hoffman [Witte and Hoffmann, 2011]. Other, more sophisticated analytic approaches to estimating the polygenic contribution to traits may be better powered than the polygenic risk score approach. The linear mixed model approach implemented by Yang et al. [Yang et al., 2011], for example, uses all markers to estimate the additive genetic contribution to a trait (not just those below a significance threshold) and does not require the available data be split into training and test sets. Finally, our data sets included cases unselected for family history of

these cancers; it is possible that it may be easier to detect a polygenic signal in data from cases with a higher proportion of affected relatives.

We conducted additional simulations to see what range of polygenic models our study was powered to detect (detailed results not shown). We had over 50% and over 90% power to detect two of the models consistent with the polygenic score results for schizophrenia (Models M1 and M2 in Purcell et al. [International Schizophrenia Consortium et al., 2009], respectively). While our results indicate there is no significant evidence for a strong polygenic effect for BCa and PCa, we cannot rule out the possibility of a weak polygenic effect. Results from our study suggests that if there is a polygenic component to breast and prostate cancer the effect is subtle and therefore the utility of using such polygenic models for genetic risk prediction may be limited.

In conclusion, our analysis found no strong evidence that polygenic risk scores improve risk prediction of breast or prostate cancer over current replicated SNP scores. With increases in sample sizes and formation of consortia, it will be possible to revisit this approach in expectation of improving risk prediction of complex diseases such as breast and prostate cancer. Innovative analytical techniques will be needed to extract genetic predictors of disease from GWAS.

Acknowledgments

Special thanks to the Cancer Genetic Markers of Susceptibility project investigators, staff, and participants. Support for this project was received from CA09001-34 and T32GM074897.

References

- Ahmed S, Thomas G, Ghoussaini M, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet.* 2009; 41:585–90. [PubMed: 19330027]
- Duggan D, Zheng SL, Knowlton M, et al. Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. *J Natl Cancer Inst.* 2007; 99:1836–44. [PubMed: 18073375]
- Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007; 447:1087–93. [PubMed: 17529967]
- Eeles RA, Kote-Jarai Z, Al Olama AA, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet.* 2009; 41:1116–21. [PubMed: 19767753]
- Eeles RA, Kote-Jarai Z, Giles GG, et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet.* 2008; 40:316–21. [PubMed: 18264097]
- Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet.* 2009; 18:3525–31. [PubMed: 19553258]
- Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials.* 2000; 21:251S–72S. [PubMed: 11189683]
- Gold B, Kirchhoff T, Stefanov S, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A.* 2008; 105:4340–5. [PubMed: 18326623]
- Gudmundsson J, Sulem P, Gudbjartsson DF, et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet.* 2009; 41:1122–6. [PubMed: 19767754]

- Gudmundsson J, Sulem P, Manolescu A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet.* 2007a; 39:631–7. [PubMed: 17401366]
- Gudmundsson J, Sulem P, Rafnar T, et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet.* 2008; 40:281–3. [PubMed: 18264098]
- Gudmundsson J, Sulem P, Steinthorsdottir V, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet.* 2007b; 39:977–83. [PubMed: 17603485]
- Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* 2007; 39:870–4. [PubMed: 17529973]
- International Multiple Sclerosis Genetics Consortium (IMSGC). Bush WS, Sawcer SJ, et al. Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. *Am J Hum Genet.* 2010; 86:621–5. [PubMed: 20362272]
- International Schizophrenia Consortium. Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009; 460:748–52. [PubMed: 19571811]
- Ioannidis JP, Castaldi P, Evangelou E. A Compendium of Genome-Wide Associations for Cancer: Critical Synopsis and Reappraisal. *J Natl Cancer Inst.* 2010
- Jacobs KB, Yeager M, Wacholder S, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat Genet.* 2009; 41:1253–7. [PubMed: 19801980]
- Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461:747–53. [PubMed: 19812666]
- Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet.* 2010; 42:570–5. [PubMed: 20562874]
- Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials.* 2000; 21:273S–309S. [PubMed: 11189684]
- Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–75. [PubMed: 17701901]
- Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet.* 2007; 39:865–9. [PubMed: 17529974]
- Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet.* 2008; 40:703–6. [PubMed: 18438407]
- Sun J, Zheng SL, Wiklund F, et al. Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. *Nat Genet.* 2008; 40:1153–5. [PubMed: 18758462]
- Thomas G, Jacobs KB, Kraft P, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet.* 2009; 41:579–84. [PubMed: 19330030]
- Thomas G, Jacobs KB, Yeager M, et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet.* 2008; 40:310–5. [PubMed: 18264096]
- Turnbull C, Ahmed S, Morrison J, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet.* 2010; 42:504–7. [PubMed: 20453838]
- TwoRoger SS, Eliassen AH, Sluss P, Hankinson SE. A prospective study of plasma prolactin concentrations and risk of premenopausal and postmenopausal breast cancer. *J Clin Oncol.* 2007; 25:1482–8. [PubMed: 17372279]
- Witte JS, Hoffmann TJ. Polygenic Modeling of Genome-Wide Association Studies: An Application to Prostate and Breast Cancer. *OMICS.* 2011
- Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011; 88:76–82. [PubMed: 21167468]
- Yeager M, Chatterjee N, Ciampa J, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet.* 2009; 41:1055–7. [PubMed: 19767755]

- Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* 2007; 39:645–9. [PubMed: 17401363]
- Zheng SL, Stevens VL, Wiklund F, et al. Two independent prostate cancer risk-associated Loci at 11q13. *Cancer Epidemiol Biomarkers Prev.* 2009a; 18:1815–20. [PubMed: 19505914]
- Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet.* 2009b; 41:324–8. [PubMed: 19219042]

Figure 1.

Average p-values for PRS in the testing set as k increases from 10 to 60,000. Results for BCa are shown in (A) and PCa in (B). The shaded region denotes a 95% confidence interval around the mean.

Figure 2.

Graphical trace of the average area under the receiver operator characteristic curve (AUC) across the ten cross-validation datasets for prediction of incident BCa (A) and PCa (B). The dashed line is the AUC as additional prioritized SNPs are added to the PRS in the training set (90%) and the solid line is the AUC for the same model in the testing set (10%). The shaded regions denote a 95% confidence interval around the mean.

Table I

Genomic positions and nearby genes of replicated SNPs associated with BCa and PCa. Correlated and proxy SNPs with an R² greater than 0.20 are not included. RA denotes the risk allele for the SNP. P-values are for BCa and PCa association in CGEMS and PLCO, respectively.

	Locus	Reported Genes	SNP	RA	P-Value	Reference	PMID
BreastCancer	1p11	Intergenic	rs11249433	G	4.20E-04	Thomas et al. NG 2009	19330030
	2q35	Intergenic	rs13387042	A	3.02E-03	Stacey et al. NG 2007	17529974
	3p24	<i>SLC4A7/NEK10</i>	rs4973768	T	1.29E-01	Ahmed et al. NG 2009	19330027
	5p12	<i>MRPS30</i>	rs10941679	G	1.20E-03	Stacey et al. NG 2008	18438407
	5q11	<i>MAP3K1</i>	rs16886165	G	1.23E-02	Thomas et al. NG 2009	19330030
	6q22	<i>ECHDC1</i>	rs2180341	G	3.49E-01	Gold et al. 2008	18326623
	6q25	Intergenic	rs2046210	A	7.77E-02	Zheng et al. NG 2009	19219042
	8q24	Intergenic	rs1562430	T	1.19E-02	Turnball et al. NG 2010	20453838
	10q26	<i>FGFR2</i>	rs2981579	A	4.87E-06	Hunter et al. NG 2007	17529973
	11p15	<i>LSP1</i>	rs3817198	C	3.89E-01	Easton et al. Nature 2007	17529967
	14q24	<i>RAD 51 LI</i>	rs999737	C	1.51E-02	Thomas et al. NG 2009	19330030
	16q12	<i>TNRC9</i>	rs3803662	A	6.41E-02	Easton et al. Nature 2007	17529967
	17q23	<i>COX 11</i>	rs6504950	G	8.98E-02	Ahmed et al. NG 2009	19330027
	Locus	Reported Genes	SNP	RA	P-Value	Reference	PMID
Prostate Cancer	2p15	<i>EHBP1</i>	rs2710647	C	3.56E-02	Eeles et al. NG 2009	19767753
	2p15	<i>EHBP1</i>	rs721048	A	8.84E-04	Gudmundsson et al. NG 2008	18264098
	2p21	<i>THADA</i>	rs1465618	T	8.51E-01	Eeles et al. NG 2009	19767753
	2q31.1	<i>ITGA6</i>	rs12621278	A	6.18E-01	Eeles et al. NG 2009	19767753
	3p12.1	<i>LOC285232</i>	rs2660753	T	4.88E-01	Eeles et al. NG 2008	18264097
	3q21.3	<i>EEFSEC</i>	rs4857841	A	3.30E-03	Gudmundsson et al. NG 2009	19767754
	4q22.3	<i>PDLIM5</i>	rs12500426	A	7.30E-02	Eeles et al. NG 2009	19767753
	4q24	<i>TET2</i>	rs7679673	A	1.19E-02	Eeles et al. NG 2009	19767753
	6q25.3	<i>SLC22A3</i>	rs9364554	T	8.19E-01	Eeles et al. NG 2008	18264097
	7p15.2	<i>JAZF1</i>	rs10486567	G	7.87E-02	Thomas et al. NG 2008	18264096
	7q21.3	<i>LMTK2</i>	rs6465657	C	4.42E-01	Eeles et al. NG 2008	18264097
	8p21.2	<i>NKX3-1</i>	rs1512268	T	2.92E-01	Eeles et al. NG 2009	19767753
	8p21.2	<i>SLC25A37</i>	rs2928679	A	2.64E-01	Eeles et al. NG 2009	19767753
	8q21.3	<i>CPNE3,CNGB3</i>	rs4961199	A	2.79E-02	Thomas et al. NG 2008	18264096
	8q24.21	Intergenic	rs16901979	A	1.40E-01	Gudmundsson et al. NG 2007	17401366
	8q24.21	Intergenic	rs4242382	A	1.47E-06	Thomas et al. NG 2008	18264096
	8q24.21	<i>POU5F1B</i>	rs6983267	G	7.50E-06	Thomas et al. NG 2008	18264096
	8q24.21	Intergenic	rs7841060	G	1.15E-01	Yeager et al. NG 2009	19767755
	9q33.2	<i>DAB2IP</i>	rs1571801	T	1.46E-03	Duggan et al. JNCI 2007	18073375
	10q11.23	<i>MSMB</i>	rs10993994	T	2.25E-03	Thomas et al. NG 2008	18264096
	10q26.13	<i>CTBP2</i>	rs4962416	C	2.67E-05	Thomas et al. NG 2008	18264096
	11p15.5	<i>IGF2, IGF2A, INS, TH</i>	rs7127900	A	3.06E-01	Eeles et al. NG 2009	19767753
	11q13.2	Intergenic	rs10896449	G	2.30E-03	Thomas et al. NG 2008	18264096

Locus	Reported Genes	SNP	RA	P-Value	Reference	PMID
11q13.2	Intergenic	rs12418451	A	1.38E-02	Zheng et al. CEBP 2009	19505914
17q12	<i>HNF1B</i>	rs4430796	A	1.47E-02	Gudmundsson et al. NG 2007	17603485
17q12	<i>HNF1B</i>	rs11649743	G	2.58E-02	Sun et al. NG 2008	18758462
17q24.3	Intergenic	rs11859962	G	3.80E-03	Gudmundsson et al. NG 2007	17603485
19q13.33	<i>KLK15, KLK3</i>	rs266849	A	8.04E-01	Eeles et al. NG 2008	18264097
22q13.2	<i>TTLL1, BIK</i>	rs5759167	T	3.01E-03	Eeles et al. NG 2009	19767753
Xp11	<i>NUDT10, NUDT11</i>	rs5945619	C	1.89E-04	Eeles et al. NG 2008	18264097