bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative

Lars G. Fritsche,[1,2,3] Stephen B. Gruber,[4] Zhenke Wu,[1,5] Ellen M. Schmidt,[1] Matthew Zawistowski[1,2], Stephanie E. Moser,[6] Victoria M. Blanc,[7] Chad M. Brummett,[6,8] Sachin Kheterpal,[6,8] Gonçalo R. Abecasis,[1,2] Bhramar Mukherjee,[1,2,5,9,10,11*]

1. Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.
2. Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.
3. K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, 7491 Trondheim, Sør-Trøndelag, Norway.
4. USC Norris Comprehensive Cancer Center, Los Angeles, CA 90033, USA.
5. Michigan Institute of Data Science, University of Michigan, Ann Arbor, MI 48109, USA.
6. Division of Pain Medicine, Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI 48109, USA.
7. Central Biorepository, University of Michigan Medical School, Ann Arbor, MI 48109, USA.
8. Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI 48109, USA.
9. Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.
10. University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, MI 48109, USA.
11. Present Address: Department of Biostatistics and Epidemiology, University of Michigan School of Public Health, 1415 Washington Heights, SPH 1 Room 4619, Ann Arbor, MI 48109, USA.

* bhramar@umich.edu

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Health systems are stewards of patient electronic health record (EHR) data with extraordinarily rich depth and breadth, reflecting thousands of diagnoses and exposures. Measures of genomic variation integrated with EHRs offer a potential strategy to accurately stratify patients for risk profiling and discover new relationships between diagnoses and genomes. The objective of this study was to evaluate whether Polygenic Risk Scores (PRS) for common cancers are associated with multiple phenotypes in a Phenome-wide Association Study (PheWAS) conducted in 28,260 unrelated, genotyped patients of recent European ancestry who consented to participate in the Michigan Genomics Initiative, a longitudinal biorepository effort within Michigan Medicine. PRS for 12 cancer traits were calculated using summary statistics from the NHGRI-EBI GWAS catalog. A total of 1,711 synthetic case-control studies was used for PheWAS analyses. Patients with at least one cancer diagnosis constituted 13,490 patients. PRSs exhibited strong association for several cancer traits they were designed for including female breast cancer, prostate cancer and melanomas of skin. Phenome-wide significant associations were observed between PRS and many non-cancer diagnoses. To differentiate primary trait-mediated PRS associations from associations that arise through shared genetic risk profiles, the idea of “exclusion PRS PheWAS” was introduced and uncovered phenome-wide significant associations between a lower risk for hypothyroidism in patients with high thyroid cancer PRS and a higher risk for actinic keratosis in patients with high squamous cell carcinoma PRS after removing all cases of the primary cancer trait. This is the first PheWAS study using PRS instead of single variant.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

In the past decade, genome-wide association studies (GWAS) using single nucleotide polymorphisms (SNPs) led to discovery of many common disease susceptibility loci [1-3]. An alternative agnostic way of exploring gene-disease association is through phenome-wide association studies (PheWAS) [4-6]. PheWAS enable simultaneous exploration of the association between genetic variants and a broad spectrum of physiological/clinical phenotypes. To explore the joint genome x phenome landscape one needs access to both Electronic Health Records (EHRs) and GWAS data. The promise and potential of these studies have recently been illustrated by the electronic Medical Records and Genomics (eMERGE) network [7; 8]. Beyond genetic associations, EHR has enabled discovery of new associations between disease and secondary effects of drugs or blood biomarker levels [9-11].

PheWAS have been used to both replicate known genetic-phenotypic associations and to discover new consequences for disease associated variants. PheWAS use computable phenotypes derived from EHR databases. Traditional PheWAS have used International Classification of Disease (ICD) codes to define a set of computable phenotypes or “PheWAS codes” defined and validated by experts using a combination of ICD codes [12]. Standard PheWAS have primarily focused on correlating genetic variants, **one at a time**, to a spectrum of phenotypes. When each variant is associated with a small effect size, these studies can only provide limited insight. For this reason, many areas of genetics now use ensembles of variants that cumulatively explain substantial variation in disease risk [13-15]. For example, polygenic risk scores (PRS) constructed from multiple GWAS identified loci have been proposed for cancer screening, risk prediction and risk stratification [16-19].

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In this paper, we introduce the new concept of PheWAS based on a PRS instead of a single genetic variant. We construct PRS for multiple cancers including some of the most common groupings of cancers in the United States: prostate cancer (PCa, MIM: 176807), breast cancer (MIM: 114480), colorectal cancer (MIM: 114500), lung cancer (MIM: 211980), melanoma of skin (MIM: 155601) and basal cell carcinoma (MIM: 614740) and correlate them with PheWAS codes.

Our study is based on the Michigan Genomics Initiative (MGI) launched in 2012, a biorepository effort to create a longitudinal cohort of participants in Michigan Medicine. MGI enrolled participants undergoing anesthesia prior to a surgery or diagnostic procedure, creating a patient community with genome-wide data, electronic health information, and permission for follow-up and re-contact in future studies. Our current analysis of 28,260 patients in MGI indicates that 47.7% of these patients have at least one current or previous neoplasm diagnosis (excluding benign neoplasms). This presents a unique opportunity to study multiple cancer outcomes leveraging both EHR and genomic data in MGI.

At the same time, this enrichment of cancer patients in MGI highlights some of the special features of the sampling frame for the study and source population. Because of the self-selective, consent-based nature of MGI patient enrollment, the sample selection mechanism is non-probabilistic, that is, the probability of a sampling unit being included in the study is not pre-determined. The MGI sample is enriched for neoplasm diagnoses which could be related to the fact that many surgeries are diagnostic procedures specifically indicated for cancer treatment and screening (e.g., colonoscopy, skin biopsies). Cancer patients undergo surgery more frequently than the general population

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and frequently choose an academic medical center for diagnostic and/or interventional procedures. The analytic framework presented in this paper conducts careful sensitivity analysis for protecting our inference against such selection biases, unbalanced case-control ratios, and phenotypic enrichment.

There are several innovative aspects to our study. Our study represents the first comprehensive phenome-wide association analysis focused on using **PRS in a cancer-enriched cohort** accrued in an academic health center. Our study is also the first PheWAS focused **on cancer**. Our results demonstrate PRS, a summary score constructed based on results of large population-based GWAS, as a biomarker that enables useful cancer risk stratification among patients in an academic medical center. We also observe that when a PRS-based PheWAS leads to the association of a cancer-specific PRS (e.g., prostate cancer PRS) with other secondary related phenotypes (e.g., erectile dysfunction or urinary incontinence), these findings may require careful consideration. We observe that many of these secondary associations are often driven by the primary cancer diagnosis and introduce the notion of “exclusion” PheWAS to detect truly independent secondary associations that have shared genetic etiology.

Subjects and Methods

MGI cohort

Participants were recruited through Michigan Medicine health system while awaiting diagnostic or interventional procedures either during a preoperative visit prior to the procedure or on the day of procedure that required anesthesia. Opt-in written informed consent is obtained. In addition to coded biosamples and protected secure health information, participants understand that all EHR, claims, and national data sources

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

linkable to the participant may be incorporated into the MGI databank. Each participant donates a blood sample for genetic analysis, undergoes baseline vital signs and a comprehensive history and physical, and completes validated self-report measures of pain, mood and function, including NIH Patient Reported Outcomes Measurement Information System (PROMIS) measures. Data were collected according to Declaration of Helsinki principles. Study participants provided written informed consent, and protocols were reviewed and approved by local ethics committees (IRB ID HUM00099605). In the current study, we report results obtained from 28,260 genotyped samples of European ancestry with available integrated EHR data (see summary characteristics of the cohort in **Table 1**).

Genotyping and Sample QC

DNA from 37,412 blood samples was genotyped on two batches of customized Illumina HumanCoreExome v12.1 bead arrays ("UM_HUNT_Biobank_11788091_A1" [N = 21,207] and "UM_HUNT_Biobank_v1-1_20006200_A" [N = 16,205]) that in addition to standard genome-wide tagging SNPs (~N=240,000) and exomic variants (N=~280,000) contained about 70,000 additional custom content variants, e.g. candidate variants from GWAS experiments, nonsense and missense variants from sequencing studies, ancestry informative markers, and Neanderthal variants. Genotype analysis was performed with Illumina GenomeStudio (module 1.9.4, algorithm GenTrain 2.0). After initial clustering, variant cluster boundaries were re-defined in a second run using only individuals with call rate of at least 99% and genotyped the remaining samples afterwards.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We excluded samples with: (1) call rate <99%, (2) estimated contamination >2.5 % (BAF Regress)[20], (3) large chromosomal copy number variants (single chromosome with missingness >= five times larger than other chromosomes), (4) lower call rate then its technical duplicate or twin, (5) gonosomal constellations other than XX and XY, or (6) whose inferred sex did not match the reported gender. We excluded variants if: (1) their probes could not be perfectly mapped or mapped perfectly to multiple position in the human genome assembly (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; BLAT) [21], (2) they showed deviations from Hardy Weinberg equilibrium in reported European ancestry samples (P<0.0001), (3) had a call rate <99%, (4) another variant with higher call rate assayed the same variant or (5) if the allele frequency differences between the two array versions within unrelated, European ancestry samples had a P-value < 0.001 (PLINK v1.90)[22]. After quality control, 392,323 polymorphic variants remained.

Before preparing the final analytical data set, we reduced the data to 33,028 samples for which we had complete age and ICD9 data available. Next, we estimated pairwise kinship with the software KING [23] and limited further analysis to a subset that contained no pairs of individuals with a 1st- or 2nd-degree relationship. We inferred recent ancestry by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project reference panel using PLINK (938 unrelated individuals) [24; 25]. We limited the principal component analysis to variants that were shared between the HGDP reference and the MGI data, had a minor allele frequency >1%, and remained after LD pruning (r^2<0.5; PLINK). Samples of recent European

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ancestry (~90% of participants) were defined as samples that fell into a circle that spanned the European populations of the HGDP panel in the PC1 versus PC2 space. Principal components were stored and used for further association tests. After quality control, 28,260 unrelated, genotyped individuals of recent European ancestry with age and ICD9 data remained for further analysis.

Phasing and Genotype Imputation

We imputed genotypes of the Haplotype Reference Consortium using the Michigan Imputation Server [26] and filtered poorly imputed variants with R^2<0.3 and/or minor allele frequency (MAF) < 0.1% resulting in over 17 million well-imputed variants. The obtained accuracy for imputed variants, i.e. the average empirical R^2 values for different MAF frequency bins, was: 0.89 (0.1%<=MAF≤0.5%), 0.94 (0.5%<MAF≤5%), and 0.96 (MAF>5%).

Phenome Generation

We extracted the ICD9 data for 28,260 unrelated, genotyped individuals of recent European ancestry and mapped 3,5 million ICD9 codes to PheWAS codes (PheWAS translation table version 1.2) [12]. The ICD9 codes (10,322 unique ICD9 codes) were aggregated to PheWAS traits using the PheWAS R package. Cases for a given PheWAS code were defined if an individual had at least one assignment of that PheWAS code in their record. The remaining individuals that did not have overlapping PheWAS codes of exclusion criteria were considered controls. A total of 1,857 case control studies were generated of which 1,711 with ≥20 cases were used for further analyses (see **Table S1**; phenotypes with < 50 cases were coded as “<50”).

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GWAS Catalog SNP Extraction and Construction of PRS

We downloaded previously reported GWAS catalog variants from the NHGRI-EBI Catalog (file date: June 31, 2017) [27]. Variant positions were converted to GRCh37 using variant IDs from dbSNP version 144 (UCSC genome browser) after updating outdated dbSNP IDs to their merged dbSNP IDs. Entries with missing risk alleles, risk allele frequencies, or odds ratios were excluded. We corrected strand flips so that the reported risk allele matched one of the alleles found at the corresponding position in the 1000 Genomes Project data. We only included entries with broad European ancestry (as reported by the NHGRI-EBI Catalog) to allow an additional quality control check by comparing the reported risk allele frequencies (RAF) in controls with the frequencies of the 503 European samples of the 1000 Genomes Project reference data (Phase 3, release 5) [28], and then excluded entries whose RAF deviated more than 15% from the reference. For each analyzed cancer type, we extracted overlapping GWAS hits in our genotype data and estimated pairwise LD (r^2) using the available allele dosages of the matched controls. For pairwise correlated SNPs (r^2>0.5) or SNPs with multiple entries, we kept the SNP with the younger publication date (and smaller *P* value, if necessary) and excluded the other (**Figure S1 & Table S2**). Finally, we calculated risk allele counts as the sum of allele dosages of risk SNPs of the risk increasing alleles and weighted the allele dosages of risk SNPs of the risk increasing alleles with the logs of their reported odds ratios and calculated PRS as their sum. Namely, for subject *j* in MGI the PRS was of the form $PRS_j=\sum_i \beta_i G_{ij}$ where the sum extends over all included loci, β_i or the log odds ratios were retrieved from the GWAS catalog for locus *i* and G_{ij} was the measured

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dosage data for the risk allele on locus *i* in subject *j.* This variable was created for each MGI participant.

Statistical Analysis

For the current study, we initially explored 30 cancer traits that had matching entries in the GWAS catalog (**Table S3**), and restricted our analysis to 12 cancer traits with at least 5 risk SNPs detected in the GWAS catalog after filtering that had relatively larger samples sizes in MGI (namely N≥250 cases) (**Table 2**).

To estimate the association of PRS with the primary cancer phenotype, we first determined the PRS quartiles using all control samples, categorized all samples according to these PRS quartiles and fitted Firth bias corrected logistic regression adjusting for age, sex, genotyping array, and the first four principal components to obtain odds ratio corresponding to the top versus the bottom quartile PRS (reference), referred to as **PRS OR**.

To compare reported associations of GWAS catalog SNPs with association observed in the MGI data set, we tested the association between reported GWAS hits and its corresponding trait using Firth bias-corrected logistic likelihood implemented in EPACTS (version 3.3, see **Web Resources**). Age, sex, genotyping array, and principal components 1–4 were included as covariates (see ***Kinship and Ancestry Inference***).

To determine the agreement of estimated effect sizes [estimated log(odds ratios)] between the MGI case- control studies and the published GWAS catalog hits, we estimated Pearson's correlation coefficient [$\hat{\rho}$] between the two sets of coefficients.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Towards more standard discovery type genome-wide association analysis with MGI data, we performed genome-wide logistic regression analysis for the 9 cancer traits where we observed strong correlation between GWAS catalog and MGI effect estimates (estimated Pearson's correlation coefficient [$\hat{\rho}$] > 0.6; **Table 2**). For scalable GWAS analysis we used fast association tests by using the saddle point approximation method [32] adjusting for age, sex, genotyping array, the first four principal components and only included SNPs with MAF>0.1%.

For our primary PRS-PheWAS, for the six PRS in **Table 2** that showed strong and significant association, we conducted Firth bias corrected logistic regression by fitting a model of the following form and repeated them for each of the 1,711 phenotypes.

logit (P(Disease=1|PRS, Age, Sex, Array, PC)

$=\beta_0 + \beta_{PRS}PRS + \beta_{Age}Age + \beta_{Sex}Sex + \beta_{Array}Array + \boldsymbol{\beta}\, PC$

Where the PCs were the first four principal components obtained from the principal component analysis of the genotyped GWAS markers and where "Array" represents the two genotyping array versions accounting for potential batch effects. To adjust for multiple testing, we applied the conservative phenome-wide Bonferroni correction according to the 1,711 analyzed PheWAS codes (**Table S1**). Through a PheWAS plot, we present –log10 (*P*-values) corresponding to each of the 1,711 association tests for $H_0: \beta_{PRS} = 0$. Firth's bias correction method was applied to all single SNP and PRS models to resolve the problem of separation in logistic regression (R package "logistf") [29; 30]. Directional arrows on the PheWAS plot indicate whether a phenome-wide significant trait was positively or negatively associated with the PRS.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Furthermore, our sensitivity analyses included similar models using the unweighted risk allele counts as predictor to consider effect orientation regardless of effect size as well as conditional logistic regression on matched case control studies to check if differences in age or sex distributions or extreme case - control ratios influenced the main analysis. We performed matching using the R package "MatchIt" and applied nearest neighbor matching for age, PC1-4 (Mahalanobis-metric matching; matching window caliper/width of 0.25 standard deviations) and exact matching for sex, limiting the maximal number of controls per case to 10. [31]

To follow-up on the observed phenome-wide significant associations and investigate the possibility of the secondary trait associations with PRS being completely mediated through the primary trait association, we performed a second set of PheWAS after excluding individuals affected with the primary cancer trait underlying the PRS, namely the "exclusion PRS PheWAS". We applied "exclusion PRS PheWAS" instead of a PRS PheWAS conditioned on the primary cancer trait as the control exclusion criteria implemented in the PheWAS phenotype construction pipeline will often eliminate these primary cancer cases from being eligible controls for some selected secondary related phenotypes and thus a logistic analysis will lead to complete separation [12]. We also stratified the MGI data set (or the corresponding gender subset depending on cancer type) into ten equally sized PRS deciles and determined the percentage of observed cases in each risk decile and conducted a test of significance in difference in proportions across the deciles before and after removing individuals affected with cancer traits related to the primary cancer trait.

If not stated otherwise, analyses were performed using R 3.4.1 [33]

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

In the current study, we report results obtained from 28,260 genotyped and unrelated samples of inferred European ancestry with available integrated ICD9-based EHR data (**Table 1**). We conducted our initial analysis on 12 cancer traits that after quality control had at least five independent risk variants in the NHGRI EBI GWAS Catalog and more than 250 cases in our cohort (**Table 2, Table S2**). **Table 2** summarizes data on 8,423 distinct individuals that were affected by at least one of the 12 cancers. Of these, 6,398 had one cancer, 1,574 had two cancers, and 451 had more than two cancer sites involved.

Correspondence of MGI effect estimates with those reported in GWAS: To assess the calibration properties of the 12 ICD-9 based cancer case-control studies, we first compared the concordance of observed effect estimates (log odds ratios) from MGI with published effect estimates reported in the NHGRI EBI GWAS catalog.

We found strong positive correlation (estimated Pearson's correlation coefficient [$\hat{\rho}$] > 0.6; P < 0.05) between the MGI and GWAS reported estimates for 9 of the 12 cancers: female breast cancer (78 SNPs; [$\hat{\rho}$]=0.67), prostate cancer (PCa; 93 SNPs; [$\hat{\rho}$]=0.81), melanoma (16 SNPs; [$\hat{\rho}$]=0.92), basal cell carcinoma (19 SNPs; [$\hat{\rho}$]=0.88), bladder cancer (MIM: 109800; 16 SNPs; [$\hat{\rho}$]=0.65), squamous cell carcinoma (5 SNPs; [$\hat{\rho}$]=0.95), lung cancer (MIM: 211980; 9 SNPs; [$\hat{\rho}$]=0.90), thyroid cancer (9 SNPs; [$\hat{\rho}$]=0.79), and cancer of brain and nervous system (9 SNPs; [$\hat{\rho}$]=0.79) (**Table 2; Table S5; Figure S2**).

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cancer GWAS in MGI: After having established strong positive correlation for 9 of the 12 cancer traits (see above) and thus a phenotype quality that might be in line with their corresponding published GWAS, we performed for each of these 9 cancers a GWAS to explore our capability to replicate and/or uncover cancer risk variants in a genome-wide setting. We found genome-wide significant signals ($P<5x10^{-8}$) for female breast cancer, melanoma of skin, basal cell carcinoma, squamous cell carcinoma and thyroid cancer. All but one of the genome-wide signals were discovered in loci already reported in the GWAS catalog for the corresponding cancer trait or related phenotypes. For instance, the four melanoma of skin loci with risk variants near *SLC45A2* (MIM 606202), *IRF4* (MIM 601900), *MC1R* (MIM 155555) and *ASIP/RALY* (MIM: 600201) were previously reported to be associated with melanoma, non-melanoma skin cancer, squamous cell carcinoma, or basal cell carcinoma [34-38]. Also, the two breast cancer risk loci near *FGFR2* (MIM: 176943) and *FGF3/FGF4* (MIM: 164950 / 164980) as well as the thyroid cancer risk loci near *NRG1* (MIM: 142445) and *FOXE1* (MIM: 602617) were previously described [39-42]. The only novel finding was the SNP rs77909434 on chromosome 13 associated with melanoma (MAF in cases = 5.3%; MAF in controls = 3.4%; $P = 1.5x10^{-8}$) located 53 kb downstream of the Fibroblast Growth Factor 9 gene (*FGF9*, MIM: 600921) on chromosome 13. We present GWAS Manhattan and QQ plots for all nine cancer traits in **Figure S3**.

Owing to the smaller sample sizes compared to the studies included in the NHGRI-EBI GWAS catalog, only 8 of 320 catalog SNPs exceeded the genome-wide significance (**Table S5**). However, we found catalogued risk SNPs in **Table 2** were markedly enriched in the top 1% of GWAS associations, especially for the larger case/control

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

studies. For example, 27 out of the 93 GWAS catalog PCa risk SNPs fall in Top 1% of associated SNPs in the MGI GWAS (with *P*<0.0083) (**Table S6**).

Reproducibility of PRS Primary Cancer association: Since PRS integrate multiple SNPs, weighted by prior effect estimates and substantially improve the power to detect an association compared to an analysis with individual variants, we constructed effect size weighted PRS for each cancer trait. To evaluate the PRSs' properties to distinguish their corresponding cancer patients from controls, we estimated the OR for patients in the top risk quartile compared to the bottom quartile (PRS OR). Six of the 12 cancer PRS revealed an at least twofold enrichment of cases with $P_{Q1vsQ4} < 4.4x10^{-9}$; all of which also showed strong positive correlation between the MGI and GWAS reported estimates (see above): female breast cancer (PRS OR = 2.3 [95% CI: 2.0;2.7], $P_{Q1vsQ4} = 6.8x10^{-28}$), prostate cancer (PRS OR = 3.2 [95% CI: 2.7;3.0], $P_{Q1vsQ4} = 1.3x10^{-38}$), melanoma (PRS OR = 2.4 [95% CI: 2.0;2.8], $P_{Q1vsQ4} = 2.4x10^{-25}$), basal cell carcinoma (PRS OR = 2.7 [95% CI: 2.2;3.2], $P_{Q1vsQ4} = 6.2x10^{-25}$), squamous cell carcinoma (PRS OR = 2.0 [95% CI: 1.6;2.5], $P_{Q1vsQ4} = 4.4x10^{-9}$), and thyroid cancer (PRS OR = 3.2 [95% CI: 2.5;4.5], $P_{Q1vsQ4} = 6.4x10^{-16}$) (**Figure 1A-F, Table 2**).

The corresponding P values obtained from Firth's Bias-Reduced Logistic Regression using continuous PRS were even stronger as expected and indicated that these six cancer traits would withstand a Bonferroni multiple testing correction in phenome setting (1,711 traits; $P_{PRS} < 2.9 \times 10^{-5}$): female breast cancer ($P_{PRS}=8.4x10^{-37}$), prostate cancer ($P_{PRS}=2.4x10^{-65}$), melanoma ($P_{PRS}=2.6x10^{-36}$), basal cell carcinoma ($P_{PRS}=4.8x10^{-44}$), cutaneous squamous cell carcinoma (SCC, $P_{PRS}=1.8x10^{-18}$), thyroid cancer (MIM: 188550) $P_{PRS}=1.6x10^{-18}$) (**Table 2, Table S3).** We excluded the remaining six cancer

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

traits from further investigation, because in this initial analysis they showed only little or moderate enrichment (PRS OR < 1.5) and consequently only modest power for subsequent exploration of phenome-wide associations (**Table 2**).

***PRS PheWAS*:** Next, we evaluated each of the six remaining PRS that were strongly associated with the primary trait across a collection of 1,711 EHR-derived phenotypes (not limited to cancer traits) with at least 20 cases each (**Table S1**). For each of the six cancer PRS, we found strongest associations with their primary traits, except for squamous cell carcinoma PRS which revealed its strongest association with the more general skin cancer trait definition ($P_{PRS} = 5.8x10^{-62}$) (**Figure 2**). We conducted sensitivity analyses of the reported results where we created 1:10 matched (for age, sex and first four PCs) case control studies. We also applied unweighted sum of risk allele counts instead of weighted PRS (**Table S7)**. The results remained robust with respect to sampling strategy and effect size weighting.

In addition, we identified for each PRS novel associations with secondary traits besides their primary traits (**Figure 2 A-F, Table S7**). For example, we observed associations of the three skin cancer PRSs (PRS for melanoma, basal cell carcinoma, and squamous cell carcinoma) with skin cancer and skin cancer sub categories – expected due to their overlapping SNP sets – but also significantly associated with dermatologic phenotypes, e.g. actinic keratosis ($P_{PRS} < 4.5x10^{-6}$) and other degenerative skin conditions or disorders, all potential pre-cancer stages (**Figure 2C,D,E**).

Similarly, the female breast cancer PRS was associated not only with breast cancer ($P_{PRS} = 8.4x10^{-37}$) but also with acquired absence of breast ($P_{PRS}=6.0x10^{-13}$), abnormal mammogram ($P_{PRS} = 1.2x10^{-8}$), benign neoplasms of the breast ($P_{PRS} = 1.5x10^{-7}$) and

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

benign mammary dysplasias (P_{PRS} =1.1x10^{-5}) (**Figure 2A**). The PRS originally constructed for prostate cancer associated with prostate cancer (P_{PRS} =2.4x10^{-65}), as expected, but also with four additional traits: elevated prostate specific antigen (P_{PRS} =3.5x10^{-26}), erectile dysfunction (P_{PRS} =8.4x10^{-15}), urinary incontinence (P_{PRS} =9.5x10^{-11}), frequency of urination and polyuria (P_{PRS} =3.6x10^{-6}), and hyperplasia of prostate (P_{PRS} =4.8x10^{-6}) (**Figure 2B, Table S7**).

While all of the above mentioned secondary trait associations were in the same effect orientation as their primary traits, i.e. increasing PRSs were associated with increased risk for the secondary trait, we observed an association of increasing thyroid cancer PRS with decreased risk for hypothyroidism (P_{PRS} =1.2x10^{-9}) **(Figure 2F)**.

Exploring Secondary PRS PheWAS Associations via Exclusion PRS PheWAS: To further explore the secondary associations in PRS PheWAS (**Figure 2)**, we performed exclusion PRS PheWAS by removing samples affected with the cancer or related cancer traits of its corresponding PRS. After removing all breast cancer cases (N = 1,894) no association with breast cancer PRS remained significant, e.g. acquired absence of breast (P_{PRS} = 0.98), abnormal mammogram (P_{PRS} = 0.51) or benign neoplasms of the breast (P_{PRS} = 0.98), indicating complete mediation of the secondary trait associations through the primary trait (**Figure S4A**). We made similar observations for prostate cancer PRS were none of the previously detected secondary trait associations remained phenome-wide significant after removing all 1,425 prostate cancer cases (**Figure S4B**).

In contrast, we found a markedly stronger association between hypothyroidism and thyroid cancer PRS after removing 472 thyroid cancer cases (P_{PRS} = 1.4x10^{-18})

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compared to the full analysis ($P_{PRS} = 1.2x10^{-9}$) which corresponded to the reversed effect orientations between thyroid cancer PRS and hypothyroidism (**Figure 2F**).

To account for the substantial overlap between skin cancer sub types, e.g. 253 of the 1,404 individuals affected with melanoma are also affected by basal and/or squamous cell carcinoma (**Figure S5**) and to account for the likely intensified skin cancer screening of individuals that were diagnosed with skin cancer once in their life time, we excluded any type of skin cancer (N = 3,910) and repeated the PheWAS for melanoma, basal cell carcinoma, and squamous cell carcinoma PRS. After doing so, only actinic keratosis remained statistically associated with squamous cell carcinoma PRS while all of the previously observed associations appeared mainly mediated through skin cancer diagnoses (**Figure 2C-E** and **Figure S4C-E**). The association between squamous cell carcinoma PRS and actinic keratosis was less pronounced after excluding skin cancer cases but still remained phenome-wide significant ($P_{PRS} = 1.3x10^{-37}$ versus $P_{PRS} = 2.8x10^{-13}$).

To further understand the discovered associations in the PRS PheWAS analyses (**Figure 2 and Figure S4)**, we conducted a simple follow-up analysis by stratifying the data into PRS deciles. We only discuss selected secondary trait associations for the prostate cancer (PCa), squamous cell carcinoma (SCC) and thyroid example in the main text and relegate their comprehensive analysis and a similar analysis of breast cancer, melanoma and basal cell carcinoma PRS to the supplemental material (**Table S8**). For prostate cancer, we stratified a total of 12,026 male individuals in MGI with age ≥30 years into deciles of PCa PRS. The observed PCa PRS associations in the

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PheWAS analysis are further supported by their respective increasing trait prevalences that are observed across 10 PCa PRS decile-stratified strata (**Figure 3A; Table S8**). These strata were not adjusted for confounders, but it is less likely that PRS is strongly associated with other covariates. A striking observation is that the proportion of PCa cases in lowest versus highest decile of PCa PRS is 5.4% versus 23.4% (Δ=18.0% [95% CI, 15.2 to 20.7%]; P=8.3x10^{-36}) emphasizing that the PRS can distinguish well between high and low risk individuals in a realistic academic medical center population. Focusing on the secondary traits that reached phenome-wide significance with the PCa PRS, all these traits are known to be associated with PCa: erectile dysfunction (ED), and urinary incontinence (UI) – which commonly follows invasive surgical removal of the prostate – and elevated prostate specific antigen (ePSA) – which is a known biomarker for an increased PCa risk being closely monitored after prostatectomy. For example, when comparing the lowest versus the highest PRS risk decile, we found significant differences for ePSA (3.9% versus 11.2%; Δ=7.3% [95% CI, 5.1 to 9.5%]; P=2.0x10^{-36}), ED (9.7% versus 17.1%; Δ=7.4% [95% CI, 4.6 to 10.2%]; P=1.4x10^{-7}), and UI (4.7% versus 11.9%; Δ=7.2% [95% CI, 5.0 to 9.5%]; P=2.0x10^{-10}). To test whether these associations are early indicator for PCa or whether they are driven by the fact that the subjects with ED are also PCa cases (perhaps as a side effect of PCa treatment), we removed PCa cases and evaluated secondary disease prevalence across PCa PRS deciles. By doing so, prevalence of all secondary traits became roughly constant across PRS strata (**Figure 3A; Table S8**) and can be illustrated by the comparison of the proportions of the lowest versus the highest PRS risk decile: ePSA (2.8% versus 2.2%, Δ=-0.6% [95% CI, -1.9 to 0.8%]; P=0.44), ED (7.6% versus 6.5%, Δ=-1.1% [95% CI, -

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2 to 1.0%]; *P*=0.34), and UI (3.0% versus 2.8%, Δ=0.2% [95% CI, -1.6 to 1.3%]; *P*=0.90). Based on these observations, we hypothesize that PCa likely mediates the association of PCa PRS on the secondary traits ePSA, ED, and UI, through either prior symptoms of PCa or prescribed medication, chemotherapy or surgical procedures for prostate removal (**Table S8**).

For SCC PRS stratification, there was a gradual increase of individuals affected with SCC with increasing PRS risk deciles, a trend that was also noted for actinic keratosis, dermatitis due to solar radiation, and seborrheic keratosis (**Figure 3B**). However, when excluding cases with skin cancer, the increasing trend for the latter two phenotypes disappeared. The previously observed difference between the top and bottom PRS risk decile of individuals affected with actinic keratosis (5.5% versus 13.2% (Δ=7.7% [95% CI, 6.1 to 9.3%]; $P=4.0\times10^{-21}$) was markedly reduced after excluding skin cancer cases but still remained significant (2.8% versus 5.1% (Δ=2.4% [95% CI, 1.3 to 3.5%]; $P=1.6\times10^{-5}$) (**Table S8**) suggesting the potential for common genetic risk profiles between SCC and actinic keratosis. Since actinic keratosis is a known precursor for squamous cell carcinoma[43], our approach indicated that it is possible to identify phenotypic risk factors through phenome-wide association scans and careful follow-up investigation of primary and secondary diagnoses.

Finally, we found an attenuated association between increasing thyroid cancer PRS and reduced risk for hypothyroidism: within all 25,681 samples >= 30 years of age the difference between bottom and top decile was Δ=-3.5% ([95% CI, -5.4 to -1.6%]; 15.1% versus 11.5%; $P=2.5\times10^{-4}$) and after excluding 452 thyroid cancer cases it increased to Δ=-5.3% ([95% CI, -7.1 to -3.5%]; 14.4% versus 9.1%; $P=4.5\times10^{-9}$) (**Table S8**). Several

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

studies previously reported genetic overlap of a subset of thyroid cancer risk variants and variants associated with serum levels of thyroid stimulating hormone (TSH) which matches the current observed association between thyroid cancer risk and risk for hypothyroidism [41; 42].

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Integration of large-scale biorepositories such as genetic data with EHRs are becoming increasingly common and indispensable for next-generation etiology studies. In this paper, we proposed, demonstrated and tested trait-specific polygenetic risk scores (PRS) that summarize the results of large population-based GWAS studies towards cancer risk prediction in an actual academic medical center population managed by Michigan Medicine. Data repositories like MGI, allow us to explore many traits simultaneously whereas population-based case-control studies focus on one specific trait. It is indeed encouraging that the results of population-based studies corroborate with the phenotypes computed from EHR data. We found improved trait prediction power of the composite PRS compared to single-SNP analyses. We also reproduced catalogued associations of SNPs for some cancer traits, observed excellent correspondence of effect estimates and discovered novel secondary trait associations that were not mediated through the primary cancer diagnoses. To our knowledge this is the first PRS PheWAS study and the first PheWAS study focused on cancer.

We have introduced several novel analytic strategies in this paper. We presented a principled framework and quality-control pipeline to create a PRS from a large curated, public database and to perform PRS PheWAS in a potentially biased sample. We introduced a primary PheWAS using Firth's bias-reduced logistic regression which has the advantage of resolving the problem of separation in logistic regression and providing well-controlled type I error rates for unbalanced case control studies with relatively small sample counts [29; 30; 44], that are often present in large EHR-based phenomes where

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

controls are frequently hundredfold more abundant than cases. In addition, we conducted thorough sensitivity analyses to check the robustness of our findings by using PheWAS with unweighted risk allele counts and PRS PheWAS based on matched controls. All our reported results remained robust under sensitivity analysis.

To differentiate between PRS association that truly derive from a shared genetic risk profile from secondary associations that are potentially mediated through the primary trait, (for example urinary incontinence or erectile dysfunction following prostate cancer treatment) we further introduced a follow-up PRS PheWAS approach that excludes the PRS's underlying cancer traits. While reducing overall sample size, this "exclusion PheWAS" approach is statistically preferable in contrast to a PRS PheWAS that conditions on the primary cancer trait. A conditional PheWAS approach is often affected by unilaterally applied exclusion criteria of controls that occur during defining a PheWAS code, e.g. PCa cases were excluded from being eligible controls for elevated PSA levels and vice versa. Our approach also could directly discard mediated trait associations and lead to the re-discovery of traits that may be useful for pre-symptomatic cancer screening: thyroid hormone levels for thyroid cancer and actinic keratosis for squamous cell carcinoma. Further exploration of our findings in larger biobank studies, like the UK Biobank study, is warranted and will empower a deeper understanding of relevant pre-cancer traits [45].

There are several limitations to the current study. We decided to rely on the associations reported in the NHGRI-EBI GWAS catalog instead of focusing on the latest and largest GWAS study specific for each cancer trait. We acknowledge that the implemented curation of GWAS findings delays inclusion of the latest publications which

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

would offer a more complete sets of genetic risk factors and more precise effect estimates. However, the NHGRI-EBI GWAS catalog's database is well organized and harmonized allowing systematic extraction of phenotype genotype associations [27]. We restricted our analysis to GWAS results from studies of broad European ancestry to match them to our cohort of predominantly European ancestry and to allow an extra filtering of potential swaps in directionality of risk allele in published GWAS studies that otherwise could have negatively affected the correlative properties of our constructed PRS. One could modify or extend construction of PRS based on global ancestry, functionality of the variant and use other weighting schemes. Stratifying the present analysis by young onset cancers, metastatic/aggressive cancers or tumor subtype will shed further insight into cancer biology, cancer genetics and specificity of the PRS-cancer association. We have completely ignored the temporal ordering in the diagnoses codes by defining dichotomous phenotypes of interest. Exploring the time-stamped data may be instrumental in understanding the secondary associations like the association between hypothyroidism and thyroid cancer PRS.

Though we note some very encouraging and promising results for the cancer traits with modest number of cases and controls and with a larger number of variants reported in the NHGRI-EBI GWAS catalog, we also note that the correlation of effect estimates or the PRS-cancer association was not very strong for some cancers with six out of twelve having a PRS OR<1.5 (**Table 2**). This could be due to several causes like limited sample size/power, heterogeneity in the definition of the cancer phenotype, incomplete specification of PRS, differences in allele frequencies in the MGI population,

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

misclassification of ICD-9 codes. To address concerns with misclassification we conducted detailed chart review of 50 randomly sampled cases with at least one neoplasm diagnosis and verified their primary and secondary cancer diagnosis. We could verify 149 of the 151 diagnoses and 49/50 patients had accurate record of their cancer diagnosis. Based on this we conclude that the rate of misclassification will likely be low for ICD 9 codes associated with cancer.

In this paper, we have focused on cancer traits. The low misclassification rate of cancer traits, typical within academic health and cancer centers, along with effective sensitivity analyses partly protect the results against imprecise case definitions and confounding. For non-cancer disease traits, more stringent ICD-9 defined cases, e.g., by repeated ICD-9 diagnoses, balanced with adequate sample sizes might alleviate the biases from case misclassification. Future analysis will need to control for potentially different levels of misclassification error across phenotypes.

Our phenome comprised a total of 1,711 ICD9-based phenotypes and by its implemented design of hierarchical phenotypes with different levels of specificity include a certain degree of redundancy. While we applied the multiple testing correction for 1,711 performed tests, we acknowledge that this threshold might be too conservative. For examples, we estimated a maximal set of 1,452 phenotypes with all pair-wise correlations $r^2 < 0.5$ before applying any exclusion criteria to the controls. In addition, the PheWAS approach often applies similar exclusion criteria to related phenotypes and thereby further reduces the actual independence of case control studies. Future studies,

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which determine the effective number of independent tests in such a phenome-wide analysis, will need to be undertaken

Besides the ICD-9 codes used for case and control definitions, EHR databases generally contain vast amount of additional patient information including ICD-10 codes, temporal laboratory tests, drug prescriptions, inpatient and outpatient records, etc. Future analyses that leverage these heterogeneous data sources that might be predictive of disease outcomes could further improve disease risk predictions. It will be interesting to study whether PRS for cancer risk behaviors like smoking, alcohol and obesity predict cancer phenotypes. Tailored and validated models capable of integrating multiple sources of data for predicting risks of disease will be crucial. This initial paper provides a framework for the exploration of genetic risk profiles of specific cancer types by using structured and unstructured content of their electronic health records that in the future can be extended by informative molecular and environmental data.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Data

Supplemental Data include five figures and eight tables.

Conflicts of Interest

The authors declare no competing financial interest.

Acknowledgments

The authors acknowledge the University of Michigan Medical School Central Biorepository for providing biospecimen storage, management, and distribution services in support of the research reported in this publication.

Web Resources

Michigan Genomics Initiative, https://www.michigangenomics.org

OMIM, http://www.omim.org/

BAF Regress, http://genome.sph.umich.edu/wiki/BAFRegress

PLINK v1.90, https://www.cog-genomics.org/plink2

KING v2.0, http://people.virginia.edu/~wc9c/KING/

Human Genome Diversity Project reference panel, http://csg.sph.umich.edu/chaolong/LASER/

PheWAS R package, https://github.com/PheWAS/PheWAS

NHGRI-EBI GWAS Catalog, https://www.ebi.ac.uk/gwas/

UCSC genome browser, http://genome.ucsc.edu

EPACTS, http://genome.sph.umich.edu/wiki/

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Titles and Legends

Figure 1. Calibration of association parameters

Calibration of association parameters between the MGI-GWAS and NHGRI-EBI GWAS derived effect estimates [log(OR)] for (**A**) breast cancer (females only), (**B**) cancer of prostate, (**C**) melanoma, (**D**) basal cell carcinoma, (**E**) squamous cell carcinoma, and (**F**) thyroid cancer. The agreement of two sets of SNP-specific beta coefficients (non-reference allele is the effect allele) and their Pearson Correlation (Coefficient ρ and *P*) are shown; dashed line: perfect concordance; solid line: fitted line.

Figure 2. PRS PheWAS plots

PRS PheWAS plots for (**A**) breast cancer (females only), (**B**) cancer of prostate, (**C**) melanoma, (**D**) basal cell carcinoma, (**E**) squamous cell carcinoma, and (**F**) thyroid cancer. 1,711 traits are grouped into 16 color-coded categories as shown on the horizontal axis; the p-values for testing the associations of PRS with the traits are minus log-base-10-transformed and shown on the vertical axis. Triangles indicate phenome-wide significant associations with their effect orientation (up-pointing = risk increasing; down-pointing = risk decreasing). PRS upon multiplicity adjustment (see ***Methods***). The solid horizontal line for P=2.9 x10^{-5} cut-off.

Figure 3. Primary and secondary traits stratified by PRS deciles

Percentage of primary and selected secondary traits in each cancer PRS category for (**A**) prostate cancer, (**B**) squamous cell carcinomas, and (**C**) thyroid cancer. Observed percentages in the MGI database as represented by the height of bars for each of 10

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

increasing decile-stratified PRS strata from left to right. The PRS's underlying trait is shown on top and secondary traits below with (blue) and without (green) overlapping relevant cancer cases. Only individuals with age≥30 years were included in each analysis, with the prostate cancer PRS example only includes male individuals (see **Table S8** for detailed sample sizes and percentages).

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

Table 1 Demographics and clinical characteristics of the final analytic data set

Characteristic	**Analytic Data Set**
N	28,260
Females, N (%)	15,113 (53.5%)
Age, years (S.D.)	54.1 (15.9)
Total number of ICD9 code days	3.5 million
Number of unique ICD9 codes	10,322
Median number of visits per participant	23
Median days between first and last visit	1,265
Median ICD9 code days per participant	28

Table 2. Association analysis of cancer traits with at least five NHGRI EBI GWAS Catalog risk SNPs and more than 250 cases in MGI.

Cancer Trait [a]	N Cases	N Controls	N Risk SNPs used for PRS [b]	Effect Size Correlation [$\hat{\rho}$] between GWAS Catalog and MGI [95% CI]	Estimated PRS-Cancer Association	
					PRS Odds Ratio [95% CI]; *P* [c]	Continous PRS *P* [d]
Breast cancer [female][e]	1,827	11,073	78	0.67 [0.53;0.78]	2.31 [1.99;2.69], 6.80×10^{-28}	8.39×10^{-37}
Cancer of prostate[e]	1,425	9,793	93	0.81 [0.73;0.87]	3.24 [2.72;3.88], 1.29×10^{-38}	2.41×10^{-65}
Melanomas of skin[e]	1,404	23,798	16	0.92 [0.77;0.97]	2.36 [2.03;2.82], 2.41×10^{-25}	2.57×10^{-36}
Basal cell carcinoma[e]	1,124	23,798	19	0.88 [0.71;0.95]	2.66 [2.21;3.21], 6.21×10^{-25}	4.80×10^{-44}
Cancer of bladder	978	26,748	16	0.65 [0.22;0.86]	1.42 [1.18;1.72], 1.99×10^{-4}	4.51×10^{-6}
Non-Hodgkins lymphoma	878	26,794	18	0.51 [0.05;0.79]	1.30 [1.08;1.57], 0.00662	0.00284
Colorectal cancer	718	22,183	42	0.48 [0.21;0.68]	1.31 [1.06;1.62], 0.0114	0.000729
Squamous cell carcinoma[e]	703	23,798	5	0.95 [0.39;1.00]	1.96 [1.57;2.47], 4.37×10^{-9}	1.83×10^{-18}
Malignant neoplasm of kidney, except pelvis	613	26,748	7	0.33 [-0.57;0.87]	0.99 [0.78;1.26], 0.934	0.902
Cancer of bronchus; lung	570	27,596	9	0.90 [0.60;0.98]	1.19 [0.907;1.6], 0.152	0.0847
Thyroid cancer[e]	472	26,692	8	0.97 [0.82;0.99]	3.23 [2.50;4.52], 6.41×10^{-16}	1.63×10^{-18}
Cancer of brain and nervous system	321	27,069	9	0.79 [0.26;0.95]	1.25 [0.92;1.74], 0.143	0.0406

[a] underlying ICD9 codes are listed in **Table S4**

[b] GWAS Catalog SNPs after quality control; corresponding summary statistics are listed in **Table S2, and Table S5**

[c] Odds ratio for each cancer with top PRS quartile compared to bottom PRS quartile. Point estimates, confidence intervals and P- values are obtained by fitting Firth's Bias-Corrected Logistic Regression.

[d] Association of each cancer with continuous PRS. P-values are obtained by fitting Firth's Bias-Corrected Logistic Regression.

[e] OR>1.5, $P_{continous\ PRS} < 2.9 \times 10^{-5}$, selected for PRS-PheWAS analysis

Notes: PRS = polygenic risk score, CI = confidence interval.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Witte, J.S. (2010). Genome-wide association studies and beyond. Annu Rev Public Health 31, 9-20 24 p following 20.
2. Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med 363, 166-176.
3. Raychaudhuri, S. (2011). Mapping rare and common causal alleles for complex human diseases. Cell 147, 57-69.
4. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26, 1205-1210.
5. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 31, 1102-1110.
6. Bush, W.S., Oetjens, M.T., and Crawford, D.C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat Rev Genet 17, 129-145.
7. Verma, A., Verma, S.S., Pendergrass, S.A., Crawford, D.C., Crosslin, D.R., Kuivaniemi, H., Bush, W.S., Bradford, Y., Kullo, I., Bielinski, S.J., et al. (2016). eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. BMC Med Genomics 9 Suppl 1, 32.
8. Rasmussen, L.V., Overby, C.L., Connolly, J., Chute, C.G., Denny, J.C., Freimuth, R., Hartzler, A.L., Holm, I.A., Manzi, S., Pathak, J., et al. (2016). Practical considerations for implementing genomic information resources. Experiences from eMERGE and CSER. Appl Clin Inform 7, 870-882.
9. Ananthakrishnan, A.N., Cagan, A., Cai, T., Gainer, V.S., Shaw, S.Y., Churchill, S., Karlson, E.W., Murphy, S.N., Liao, K.P., and Kohane, I. (2016). Statin Use Is Associated With Reduced Risk of Colorectal Cancer in Patients With Inflammatory Bowel Diseases. Clin Gastroenterol Hepatol 14, 973-979.
10. Ananthakrishnan, A.N., Cagan, A., Cai, T., Gainer, V.S., Shaw, S.Y., Churchill, S., Karlson, E.W., Murphy, S.N., Kohane, I., Liao, K.P., et al. (2015). Common Genetic Variants Influence Circulating Vitamin D Levels in Inflammatory Bowel Diseases. Inflamm Bowel Dis 21, 2507-2514.
11. Ananthakrishnan, A.N., Cagan, A., Cai, T., Gainer, V.S., Shaw, S.Y., Savova, G., Churchill, S., Karlson, E.W., Murphy, S.N., Liao, K.P., et al. (2016). Identification of Nonresponse to Treatment Using Narrative Data in an Electronic Health Record Inflammatory Bowel Disease Cohort. Inflamm Bowel Dis 22, 151-158.
12. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 30, 2375-2376.
13. Mavaddat, N., Pharoah, P.D., Michailidou, K., Tyrer, J., Brook, M.N., Bolla, M.K., Wang, Q., Dennis, J., Dunning, A.M., Shah, M., et al. (2015). Prediction of breast

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cancer risk based on profiling with common genetic variants. J Natl Cancer Inst 107.
14. Frampton, M., and Houlston, R.S. (2017). Modeling the prevention of colorectal cancer from the combined impact of host and behavioral risk factors. Genet Med 19, 314-321.
15. Szulkin, R., Whitington, T., Eklund, M., Aly, M., Eeles, R.A., Easton, D., Kote-Jarai, Z.S., Amin Al Olama, A., Benlloch, S., Muir, K., et al. (2015). Prediction of individual genetic risk to prostate cancer using a polygenic score. Prostate 75, 1467-1474.
16. Maas, P., Barrdahl, M., Joshi, A.D., Auer, P.L., Gaudet, M.M., Milne, R.L., Schumacher, F.R., Anderson, W.F., Check, D., Chattopadhyay, S., et al. (2016). Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol 2, 1295-1302.
17. Chatterjee, N., Shi, J., and Garcia-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet 17, 392-406.
18. Garcia-Closas, M., Rothman, N., Figueroa, J.D., Prokunina-Olsson, L., Han, S.S., Baris, D., Jacobs, E.J., Malats, N., De Vivo, I., Albanes, D., et al. (2013). Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. Cancer Res 73, 2211-2220.
19. Hsu, L., Jeon, J., Brenner, H., Gruber, S.B., Schoen, R.E., Berndt, S.I., Chan, A.T., Chang-Claude, J., Du, M., Gong, J., et al. (2015). A model to determine colorectal cancer risk using common genetic susceptibility loci. Gastroenterology 148, 1330-1339 e1314.
20. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91, 839-848.
21. Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.
22. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.
23. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867-2873.
24. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., Study, F., Fulton, R., et al. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. Nat Genet 46, 409-415.
25. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.
26. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48, 1279-1283.

bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

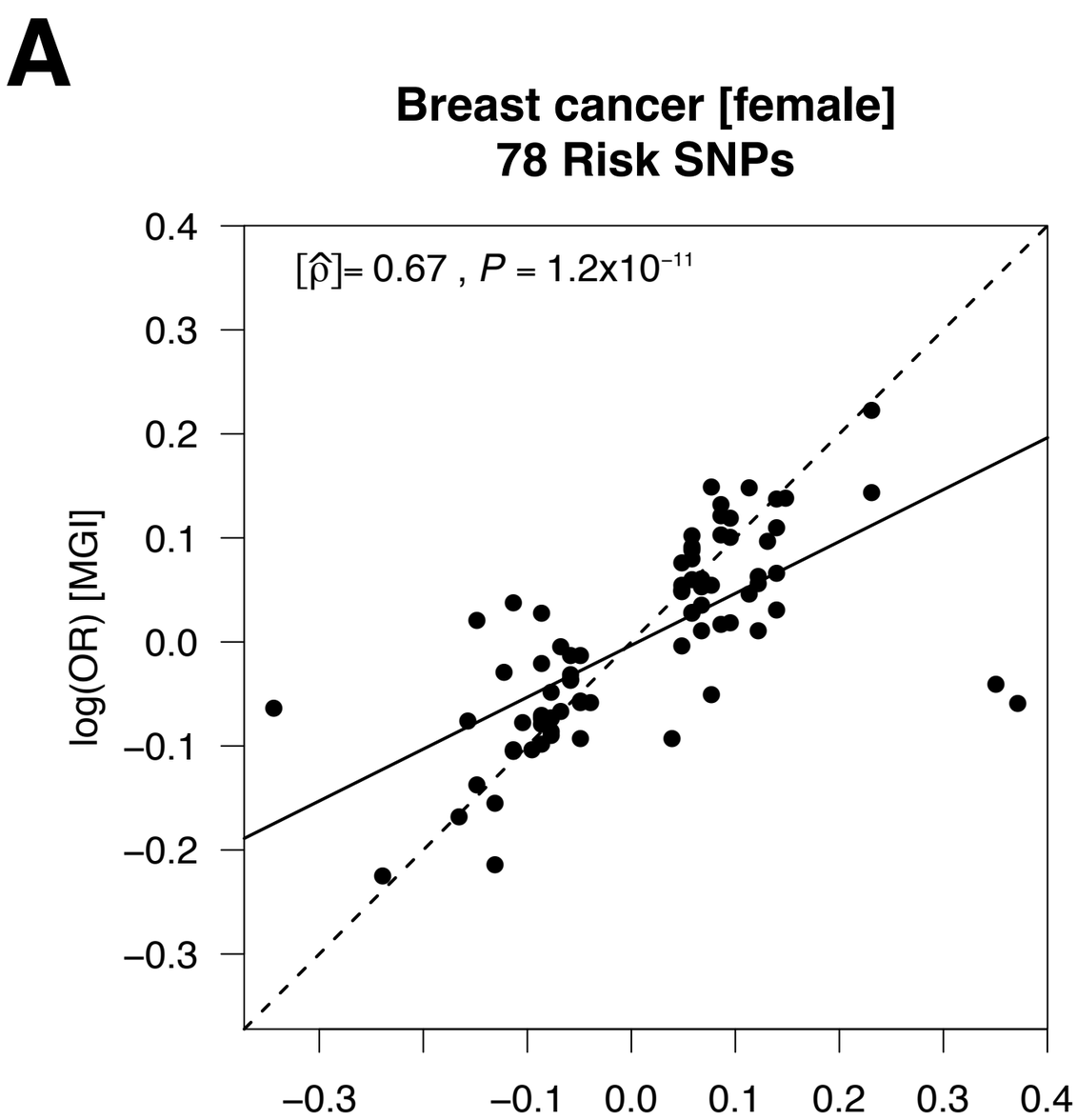
27. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001-1006.
28. The 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
29. Heinze, G., Ploner, M., Dunkler, D., and Southworth, H. (2013). logistf: Firth's bias reduced logistic regression. In. (
30. Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat Med 25, 4216-4226.
31. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software 42, 1-28.
32. Dey, R., Schmidt, E.M., Abecasis, G.R., and Lee, S.A. (2016). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. In. (Vancouver, BC, Presented at the American Society of Human Genetics Annual Meeting.
33. R Core Team. (2016). R: A Language and Environment for Statistical Computing. In. (R Foundation for Statistical Computing, Vienna, Austria.
34. Nan, H., Xu, M., Kraft, P., Qureshi, A.A., Chen, C., Guo, Q., Hu, F.B., Curhan, G., Amos, C.I., Wang, L.E., et al. (2011). Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma. Hum Mol Genet 20, 3718-3724.
35. Barrett, J.H., Iles, M.M., Harland, M., Taylor, J.C., Aitken, J.F., Andresen, P.A., Akslen, L.A., Armstrong, B.K., Avril, M.F., Azizi, E., et al. (2011). Genome-wide association study identifies three new melanoma susceptibility loci. Nat Genet 43, 1108-1113.
36. Bishop, D.T., Demenais, F., Iles, M.M., Harland, M., Taylor, J.C., Corda, E., Randerson-Moor, J., Aitken, J.F., Avril, M.F., Azizi, E., et al. (2009). Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet 41, 920-925.
37. Brown, K.M., Macgregor, S., Montgomery, G.W., Craig, D.W., Zhao, Z.Z., Iyadurai, K., Henders, A.K., Homer, N., Campbell, M.J., Stark, M., et al. (2008). Common sequence variants on 20q11.22 confer melanoma susceptibility. Nat Genet 40, 838-840.
38. Asgari, M.M., Wang, W., Ioannidis, N.M., Itnyre, J., Hoffmann, T., Jorgenson, E., and Whittemore, A.S. (2016). Identification of Susceptibility Loci for Cutaneous Squamous Cell Carcinoma. J Invest Dermatol 136, 930-937.
39. Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39, 870-874.
40. Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C.S., et al. (2010). Genome-wide

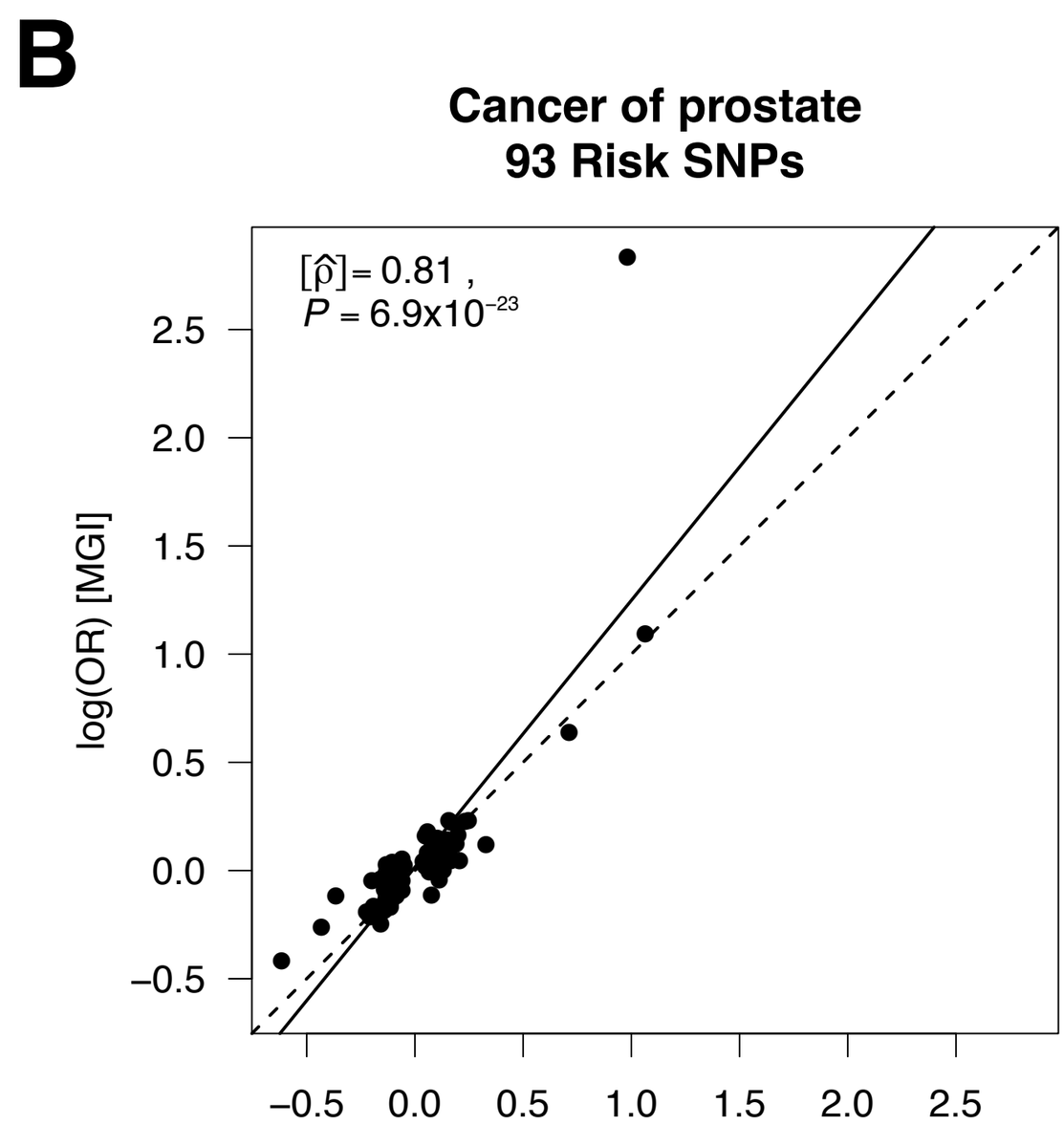
bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

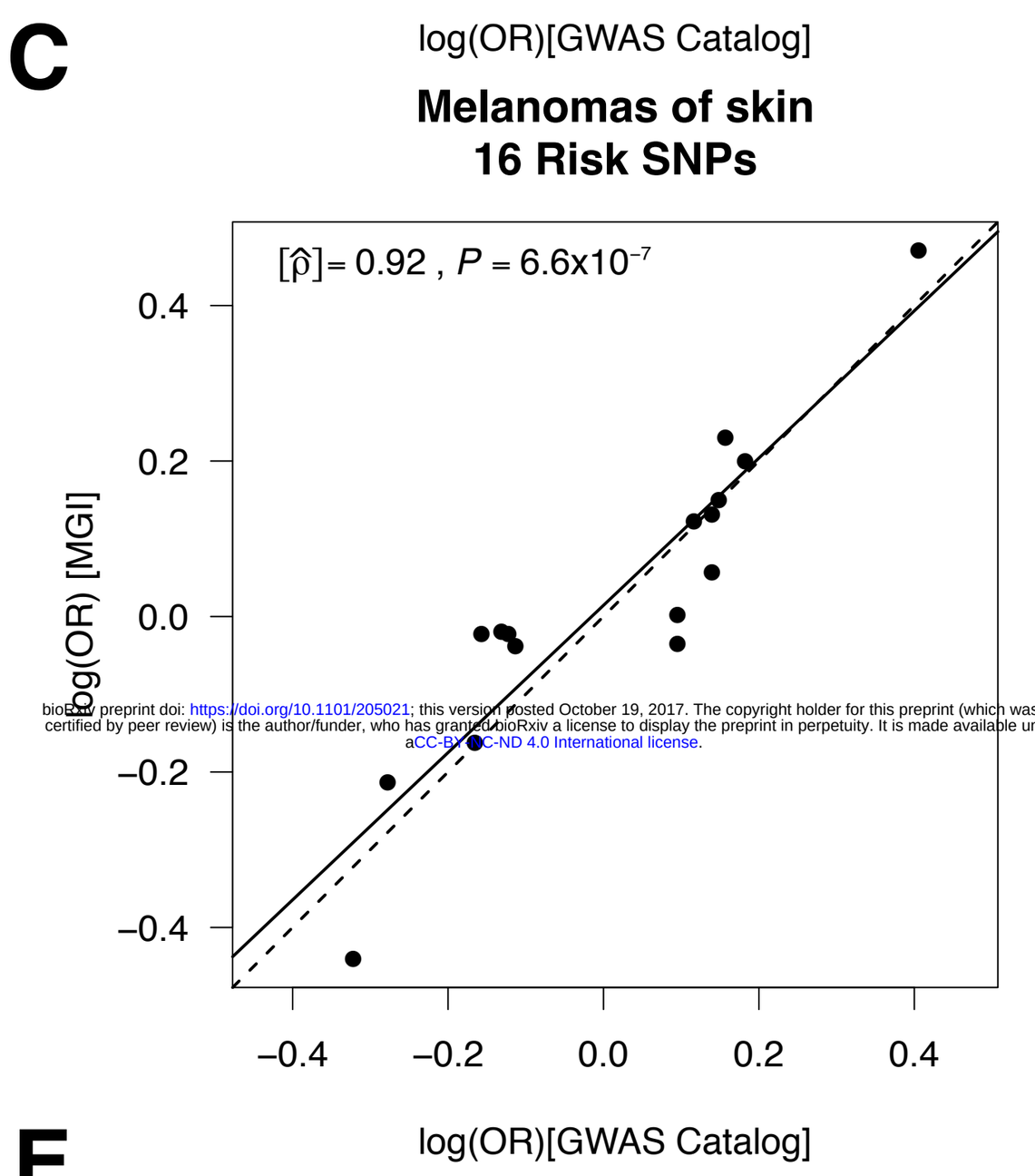
association study identifies five new breast cancer susceptibility loci. Nat Genet 42, 504-507.

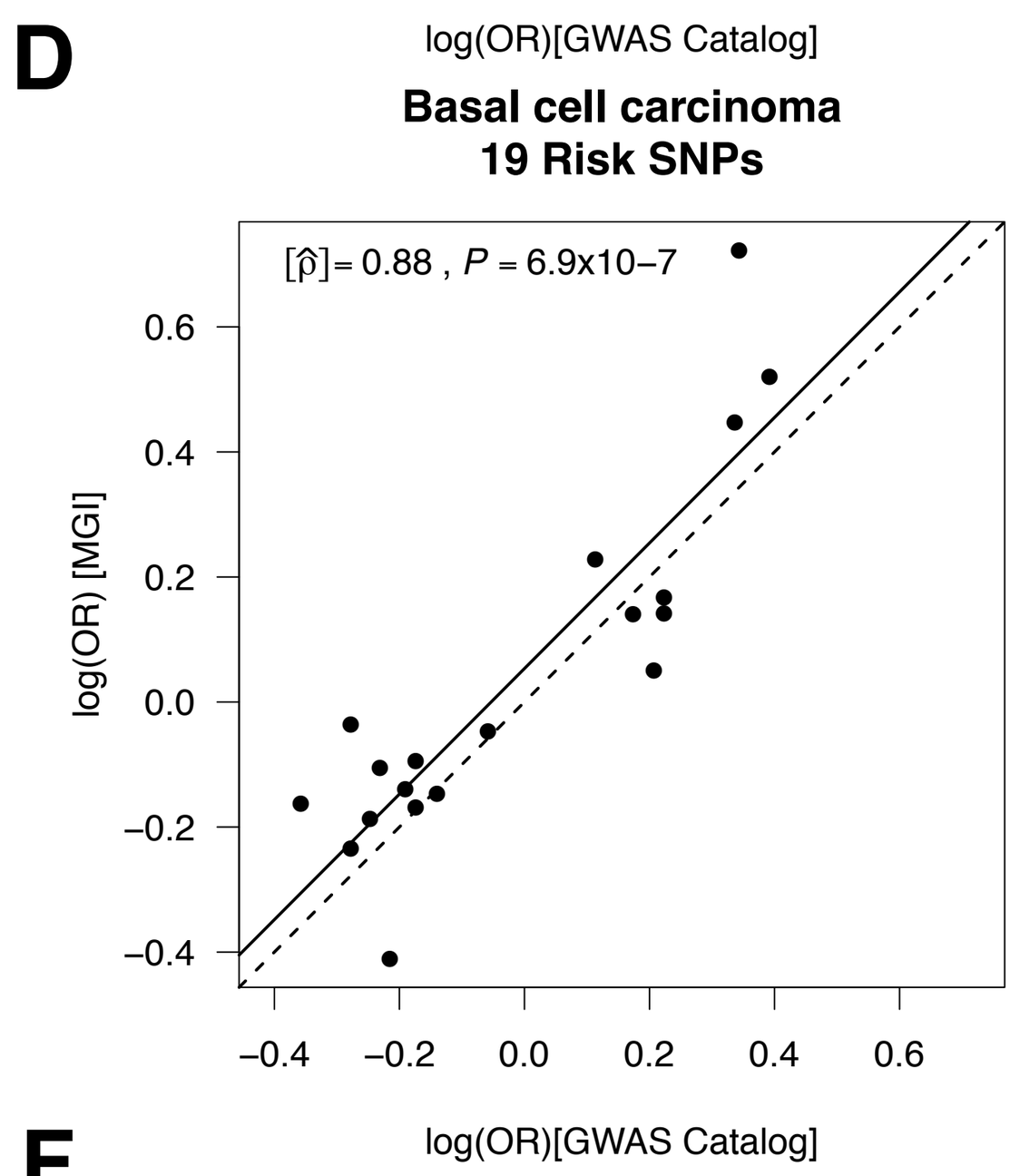
41. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Jonasson, J.G., Masson, G., He, H., Jonasdottir, A., Sigurdsson, A., Stacey, S.N., Johannsdottir, H., et al. (2012). Discovery of common variants associated with low TSH levels and thyroid cancer risk. Nat Genet 44, 319-322.
42. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Jonasson, J.G., Sigurdsson, A., Bergthorsson, J.T., He, H., Blondal, T., Geller, F., Jakobsdottir, M., et al. (2009). Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nat Genet 41, 460-464.
43. Werner, R.N., Sammain, A., Erdmann, R., Hartmann, V., Stockfleth, E., and Nast, A. (2013). The natural history of actinic keratosis: a systematic review. Br J Dermatol 169, 502-518.
44. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J., and Go, T.D.i. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37, 539-550.
45. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12, e1001779.

Figure 1

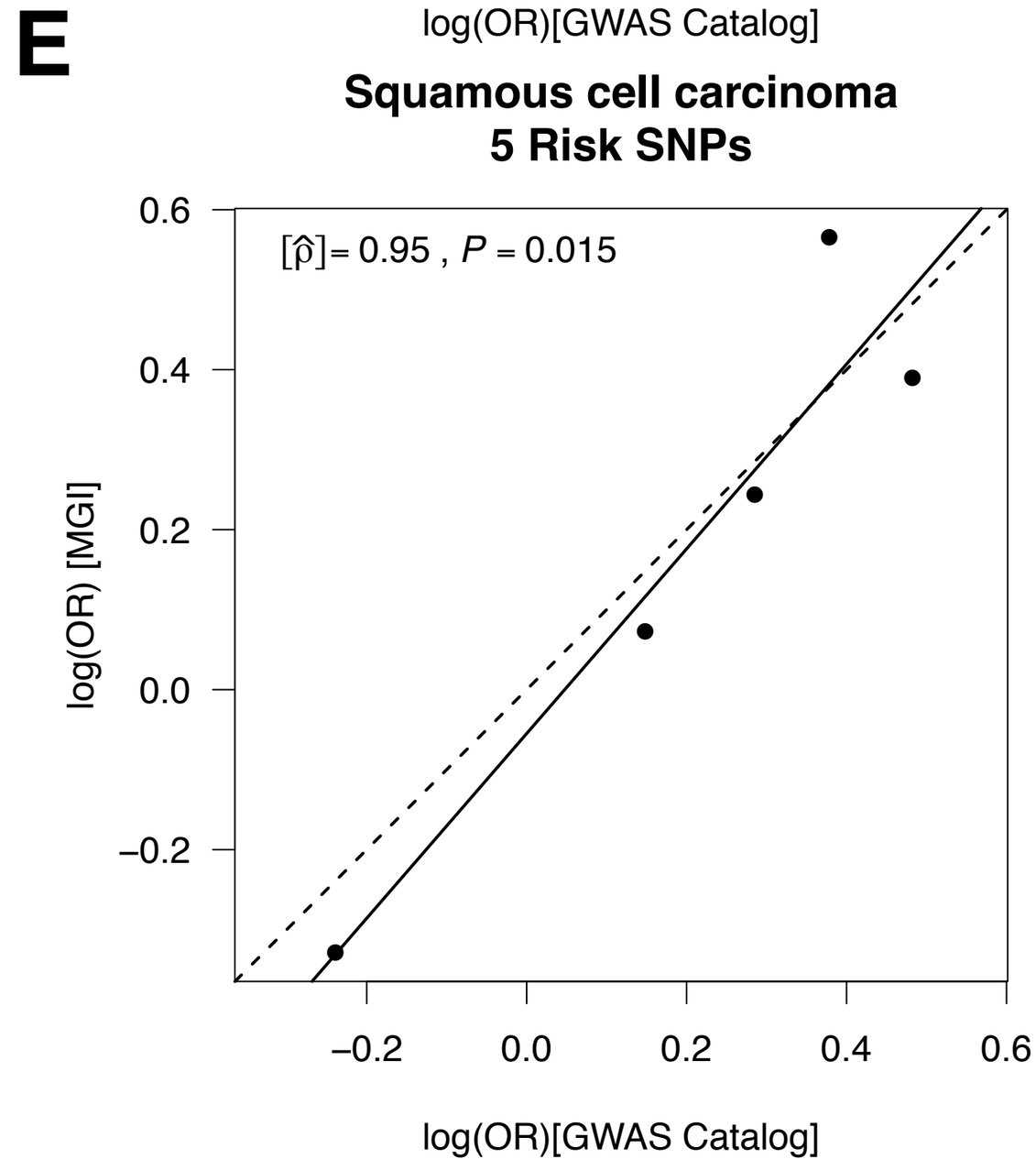








bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



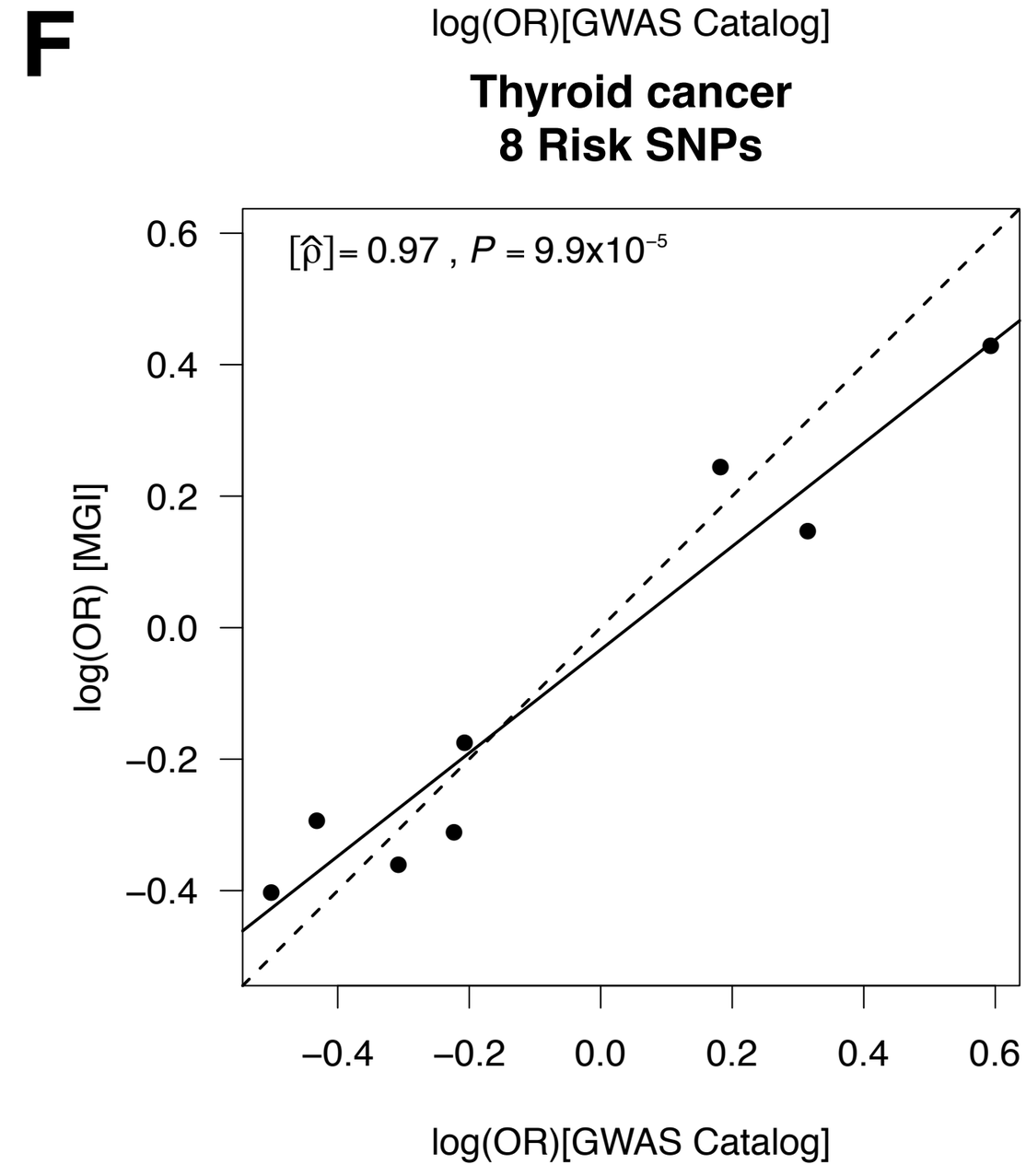
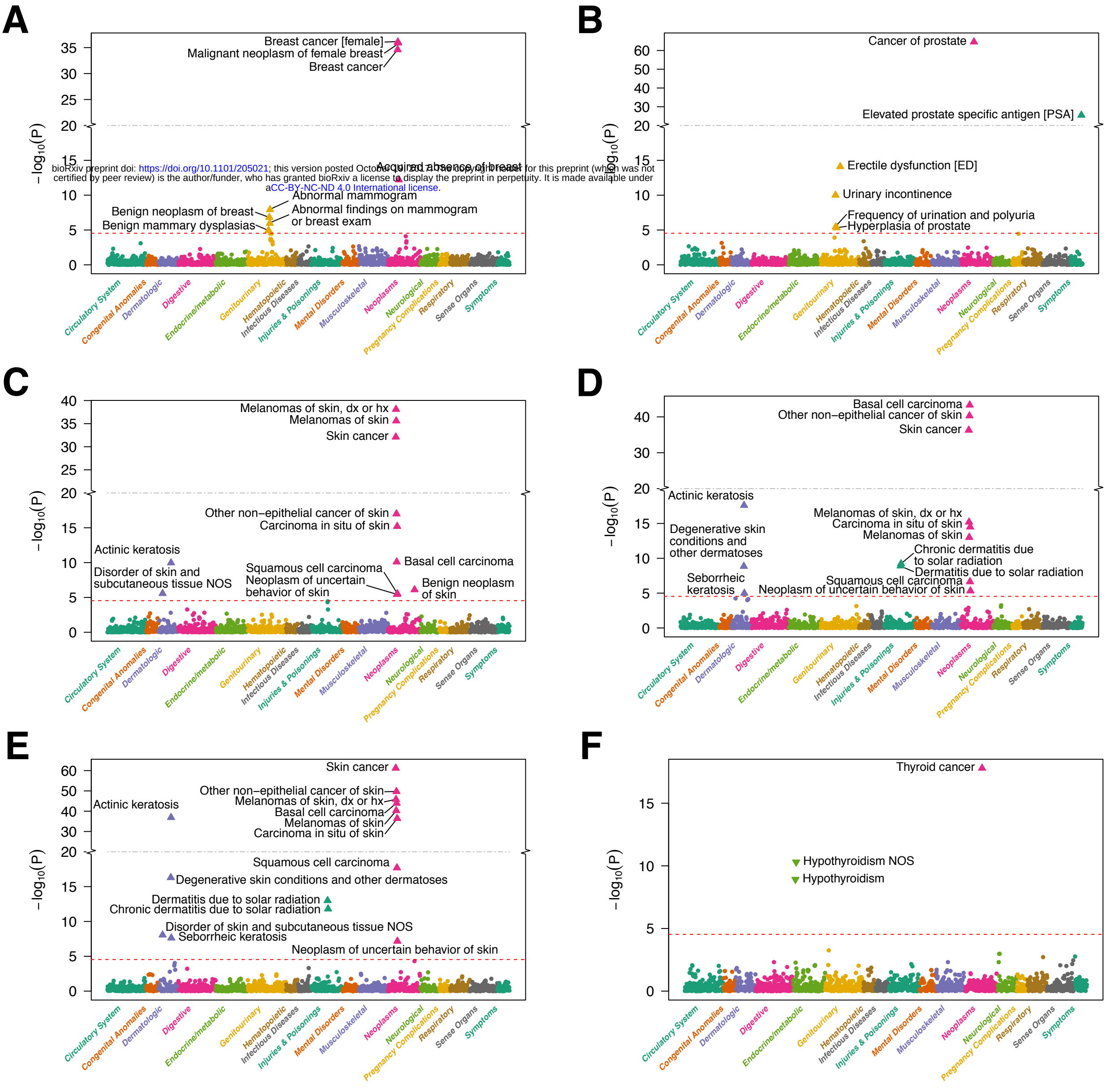
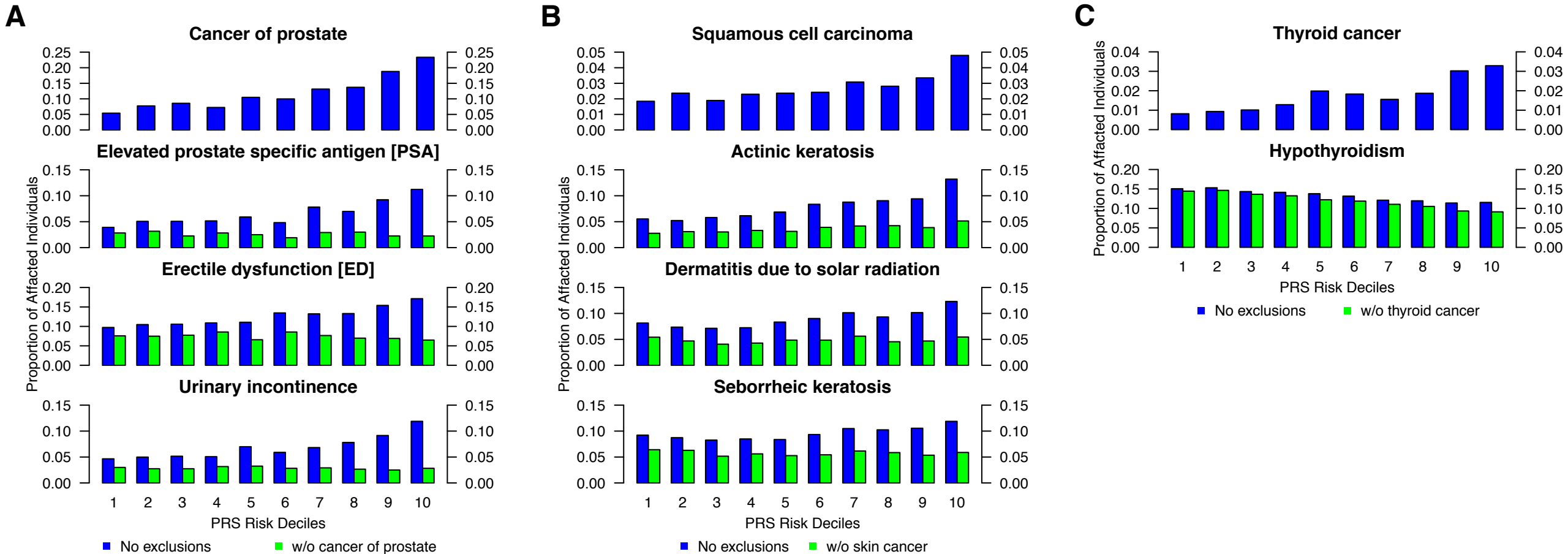


Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/205021; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3


A
Cancer of prostate
Elevated prostate specific antigen [PSA]
Erectile dysfunction [ED]
Urinary incontinence
Proportion of Affected Individuals
PRS Risk Deciles
No exclusions
w/o cancer of prostate
B
Squamous cell carcinoma
Actinic keratosis
Dermatitis due to solar radiation
Seborrheic keratosis
Proportion of Affected Individuals
PRS Risk Deciles
No exclusions
w/o skin cancer
C
Thyroid cancer
Hypothyroidism
Proportion of Affected Individuals
PRS Risk Deciles
No exclusions
w/o thyroid cancer