Exploring Various Polygenic Risk Scores for Skin Cancer in the Phenomes of the Michigan Genomics Initiative and the UK Biobank with a Visual Catalog: *PRSWeb*

Lars G. Fritsche[1,2,¶*], Lauren J. Beesley[1,¶], Peter VandeHaar[1,2], Robert B. Peng[1], Maxwell Salvatore[1], Matthew Zawistowski[1,2], Sarah A. Gagliano[1,2], Sayantan Das[1,2], Jonathon LeFaive[1,2], Erin O. Kaleba[3], Thomas T. Klumpner[3,4], Stephanie E. Moser[3], Victoria M. Blanc[5], Chad M. Brummett[3,4], Sachin Kheterpal[3,4], Gonçalo R. Abecasis[1,2], Stephen B. Gruber[6], Bhramar Mukherjee[1,2,7,8,9*]

[1] Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America
[2] Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America
[3] Division of Pain Medicine, Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America
[4] Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan, United States of America
[5] Central Biorepository, University of Michigan Medical School, Ann Arbor, Michigan, United States of America
[6] USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America
[7] Michigan Institute for Data Science, University of Michigan, Ann Arbor, Michigan, United States of America
[8] Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America
[9] University of Michigan Rogel Cancer Center, University of Michigan, Ann Arbor, Michigan, United States of America

* Corresponding authors
E-mail: bhramar@umich.edu (BM), larsf@umich.edu (LGF)

[¶] These authors contributed equally to this work

Abstract

Polygenic risk scores (PRS) are designed to serve as a single summary measure, condensing information from a large number of genetic variants associated with a disease. They have been used for stratification and prediction of disease risk. The construction of a PRS often depends on the purpose of the study, the available data/summary estimates, and the underlying genetic architecture of a disease. In this paper, we consider several choices for constructing a PRS using summary data obtained from various publicly-available sources including the UK Biobank and evaluate their abilities to predict outcomes derived from electronic health records (EHR). We examine the three most common skin cancer subtypes in the USA: basal cell carcinoma, cutaneous squamous cell carcinoma, and melanoma. The genetic risk profiles of subtypes may consist of both shared and unique elements and we construct PRS to understand the common versus distinct etiology. This study is conducted using data from 30,702 unrelated, genotyped patients of recent European descent from the Michigan Genomics Initiative (MGI), a longitudinal biorepository effort within Michigan Medicine. Using these PRS for various skin cancer subtypes, we conduct a phenome-wide association study (PheWAS) within the MGI data to evaluate their association with secondary traits. PheWAS results are then replicated using population-based UK Biobank data. We develop an accompanying visual catalog called *PRSweb* that provides detailed PheWAS results and allows users to directly compare different PRS construction methods. The results of this study can provide guidance regarding PRS construction in future PRS-PheWAS studies using EHR data involving disease subtypes.

Author summary

In the study of genetically complex diseases, polygenic risk scores synthesize information from multiple genetic risk factors to provide insight into a patient's risk of developing a disease based on his/her genetic profile. These risk scores can be explored in conjunction with health and disease information available in the electronic medical records. They may be associated with diseases that may be related to or precursors of the underlying disease of interest. Limited work is available guiding risk score construction when the goal is to identify associations across the medical phenome. In this paper, we compare different polygenic risk score construction methods in terms of their relationships with the medical phenome. We further propose methods for using these risk scores to decouple the shared and unique genetic profiles of related diseases and to explore related diseases' shared and unique secondary associations. Leveraging and harnessing the rich data resources of the Michigan Genomics Initiative, a biorepository effort at Michigan Medicine, and the larger population-based UK Biobank study, we investigated the performance of genetic risk profiling methods for the three most common types of skin cancer: melanoma, basal cell carcinoma and squamous cell carcinoma.

Introduction

The underlying risk factors of genetically complex diseases are numerous. Genome-wide association studies (GWAS) on thousands of diseases and traits have made great strides to uncover a vast array of genetic variants that contribute to genetic predispositions to a disease [1]. In order to harness the information from a large number of genetic variants, a popular approach is to summarize their contribution through polygenic risk scores (PRS). While the performance of PRS to predict disease outcomes at a population level has been modest for many diseases including most cancers, PRS have successfully been applied for risk stratification of cohorts [2, 3] and recently have been used to screen a multitude of clinical phenotypes (collectively called the medical phenome) for secondary trait associations [4, 5]. The goal of these phenome-wide screenings is to uncover phenotypes that share genetic components with the primary trait that, if pre-symptomatic, could shed biological insights into the disease pathway and inform early interventions or screening efforts for individuals at risk. However, limited prior work is available guiding the choice of PRS construction for testing associations across the medical phenome.

In the post-GWAS era and with the availability of large biobank data from multiple sources, general guidance for constructing a PRS for a phenotype of interest is needed. A PRS of the general form $\sum_{i=1}^{K} \widehat{\beta}_i G_i$ requires specification of three things: a list of markers $G_1, G_2, \cdots G_K$, the depth of the list or the number of markers (*K),* and the choice of the weights $\widehat{\beta}_i$. These choices can be based on information extracted from the latest GWAS or GWAS meta-analysis (when available), the NHGRI-EBI GWAS catalog of published results [1] (when available), or summary data for GWAS corresponding to

each phenotype, e.g., from efforts that comprehensively screened the UK Biobank (UKB) phenome [6, 7]. While various methods of constructing PRS have been widely studied for predicting the primary phenotype collected through population-based sampling [8, 9], it is unknown how the different PRS will be associated with a multitude of other diagnoses across the medical phenome. This study attempts to bridge this knowledge gap.

In this paper, we first explore strategies for constructing a PRS using markers and weights obtained from either the latest GWAS or the NHGRI-EBI GWAS catalog that have reached genome-wide significance. We compare the PRS in terms of their performance [10] for the three most common skin cancer subtypes in the USA: basal cell carcinoma (MIM: 614740) [11], cutaneous squamous cell carcinoma [12] and melanoma (MIM: 155601) [13] . We compare the two strategies using an independent biobank of genetic, demographic, and phenotype data collected by the Michigan Genomics Initiative (MGI), a longitudinal biorepository effort within Michigan Medicine (University of Michigan) [4, 14]. Based on these results, we choose a PRS construction strategy for each skin cancer subtype for further analysis.

For the chosen PRS corresponding to each skin cancer subtype, we perform a phenome-wide association study (PheWAS) relating the PRS to the electronic health record (EHR)-based phenome of MGI. We call such a study a PRS-PheWAS.[4] PRS-PheWAS results are then replicated using the population-based UK Biobank data. In order to identify secondary associations that are not driven by the primary phenotype, we perform an additional “exclusion” PRS-PheWAS for each skin cancer subtype in which we exclude subjects with any type of observed skin cancer.[4] These studies

118 demonstrate differences in PheWAS results for PRS constructed for particular disease
119 subtypes and the ability of such studies to reproduce known associations between
120 secondary phenotypes and particular disease subtypes.

121 We then describe an approach for using PRS to (1) understand the shared and
122 unique genetic architecture of disease subtypes and to (2) identify shared and unique
123 secondary phenotype associations related to this genetic architecture. We define a new
124 PRS for each skin cancer subtype using loci ***unique*** to that subtype's chosen PRS. We
125 further construct a composite PRS for general skin cancer consisting of loci ***common***
126 among all subtypes' PRS. While merging distinct clinical entities into a compound PRS
127 may seem counterintuitive in terms of specificity, such an approach may increase power
128 to identify dermatological features through PheWAS that are shared by all three
129 subtypes, which may in turn provide guidance for general skin cancer screening efforts
130 and sun protection behavior.

131 The NHGRI-EBI GWAS Catalog and Latest GWAS PRS construction methods
132 are based on published GWAS studies, which only report risk variants that reached
133 genome-wide significance (usually defined by a P-value threshold of $P < 5x10^{-8}$).
134 However, it is likely that there are additional risk variants below this threshold that could
135 be associated with the trait but have not reached statistical significance [15].
136 Incorporating non-significant variants may conceivably improve the predictive power of
137 a PRS but may also add additional random false positive signals, which in turn could
138 dilute the discriminatory power of the true risk variants and diminish any predictive gain
139 [8, 16]. To explore whether a PRS constructed using additional non-significant loci may
140 outperform a PRS using only loci reaching genome-wide significance, we evaluated a

PRS constructed using publicly available genome-wide summary statistics from the UK Biobank at six different p-value thresholds both in terms of associations with skin cancer phenotypes and in terms of secondary phenotype associations. There is an extensive literature on constructing genome-wide PRS using random effects, shrinkage methods, or thresholding (our focus) [17-19], but none of these methods have been evaluated in a PheWAS setting.

In this paper, we focus our attention on skin cancer, but the approaches used in this paper can be applied to study many other phenotypes. We chose to use skin cancer as a demonstrative example for a variety of reasons. First, our discovery dataset (MGI) is particularly enriched for skin cancer cases due to the strong skin cancer clinical program at Michigan Medicine and due to the high rate of surgery for skin cancer patients. MGI primarily recruits participants undergoing surgery and is therefore enriched for cancers and other medical comorbidities when compared to a general population [4]. Additionally, skin cancer has well-defined subtypes, which allows us to explore subtype-specific PRS constructed for several related but distinct diseases in terms of their performance for related skin cancer outcomes. Skin cancer also provides a setting in which there may be genetic factors uniquely related to particular subtypes as well as genetic factors that are shared risk factors for all skin cancer subtypes. The various PRS construction methods explored in this paper delivered tools to explore shared and subtype-specific phenotypes and may provide an enhanced understanding of the genome x phenome landscape.

We develop an online visual web catalog called *PRSweb* that provides PRS-PheWAS results for melanoma, basal cell carcinoma, and squamous cell carcinoma.

PheWAS results are available using three different PRS construction methods explored in this paper: Latest GWAS, NHGRI-EBI GWAS Catalog, and the UK Biobank GWAS summary statistics using different significance thresholds. The weights and the marker list for each PRS method can be downloaded. Furthermore, PheWAS summary statistics can be accessed from *PRSweb* (see **Web resources**), providing future investigators with readily available and useful tools to perform further analyses.

Comprehensive phenome-wide and genome-wide analyses of large biobank studies with publicly available summary statistics can be rich resources for PRS construction, especially if the trait-of-interest's prevalence is high in the biobank. Using PRS, we can synthesize complex genetic information that is then used to identify these shared genetic components across phenotypes. Compared to prior and existing literature, our contribution is new in four principal directions: (1) comparing various PRS construction methods in terms of their relationships with related EHR-derived phenotypes (2) comparing PRS associations with secondary phenotypes across the phenome of MGI (academic medical center) and UK Biobank (population-based), (3) developing PRS-based methods for understanding the shared and unique genetic contribution across disease sub-types both in terms of disease biology and in terms of secondary phenotype associations, and (4) introducing a publicly accessible online visual catalog to visually represent the genome x phenome landscape and access summary data from GWAS and PheWAS.

Material and methods

Discovery and replication cohorts

MGI cohort (discovery cohort). Participants were recruited through the Michigan Medicine health system while awaiting diagnostic or interventional procedures either during a preoperative visit prior to the procedure or on the day of the procedure that required anaesthesia. Opt-in written informed consent was obtained. In addition to coded biosamples and secure protected health information, participants understood that all EHR, claims, and national data sources – linkable to the participant – may be incorporated into the MGI databank. Each participant donated a blood sample for genetic analysis, underwent baseline vital signs and a comprehensive history and physical assessment. Data were collected according to Declaration of Helsinki principles. Study participants' consent forms and protocols were reviewed and approved by local ethics committees (IRB ID HUM00099605). In the current study, we report results obtained from 30,702 unrelated, genotyped samples of recent European ancestry with available integrated EHR data (~90 % of all MGI participants were inferred to be of recent European ancestry) [4].

UK Biobank cohort (replication cohort). The UK Biobank is a population-based cohort collected from multiple sites across the United Kingdom and includes over 500,000 participants aged between 40 and 69 years when recruited in 2006–2010 [20]. The open access UK Biobank data used in this study included genotypes, ICD9 and ICD10 codes, inferred sex, inferred white British-European ancestry, kinship estimates

down to third degree, birthyear, genotype array, and precomputed principal components of the genotypes.

Genotyping, sample quality control and imputation

MGI. DNA from 37,412 blood samples was genotyped on customized Illumina Infinium CoreExome-24 bead arrays and subjected to various quality control filters that resulted in a set of 392,323 polymorphic variants. Principal components and ancestry were estimated 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) [21, 22]. Pairwise kinship was assessed with the software KING [23], and the software fastindep was used to reduce the data to a maximal subset that contained no pairs of individuals with 3rd-or closer degree relationship [24]. We also removed patients not of recent European descent from the analysis, resulting in a final sample of 30,702 unrelated subjects. Additional genotypes were obtained using the Haplotype Reference Consortium using the Michigan Imputation Server [25] and included over 17 million imputed variants with R^2<0.3 and/or minor allele frequency (MAF) <0.1%. Genotyping, quality control and imputation are described in detail elsewhere [4]. **Table 1** provides some descriptive statistics of the MGI and UK Biobank samples.

Table 1. Demographics and Clinical Characteristics of the Analytic Datasets

Characteristic	MGI	UK Biobank*
n	30,702	408,961
Females, n (%)	16,297 (53.1%)	221,052 (54.1%)
Mean Age, years (S.D.)	54.2 (15.9)	57.7 (8.1)
Median number of visits per participant	27	n/a
Median days between first and last visit	1,469	n/a
Total number of ICD9 code days	3,459,331	49,085
Number of unique ICD9 codes	10,323	3,126
Median ICD9 code days per participant	58	2
Total number of ICD10 code days	1,311,264	2,764,868
Number of unique ICD10 codes	14,997	11,059
Median ICD10 code days per participant	27	6
Total number of PheWAS code days	6,367,117	3,679,624
Number of unique PheWAS codes	1,856	1,680
Median PheWAS code days per participant	94	8
n cases with Skin Cancer	4,503	13,782 (13,624***)
n cases with melanomas of skin	1,772	2,724 (2,718***)
n cases with epithelial skin cancer and others**	3,220	11,152 (11,030***)
n cases with basal cell carcinoma	1,303	Not available
n cases with squamous cell carcinoma	836	Not available

* The provided characteristics are based a subset of white British subjects of the UK Biobank Study for which phenotype data and imputed data was available. To retain as many unrelated cases as possible for each trait, a maximal set of unrelated cases was identified before choosing controls from the pool of subjects unrelated to these cases or to each other.

** Original PheWAS code “172.2” description "Other non-epithelial cancer of skin".

*** Unrelated cases

ICD9 and ICD10: International Statistical Classification of Diseases codes (9^{th} and 10^{th} revision)

UK Biobank. The UK Biobank is a population-based cohort collected from multiple sites across the United Kingdom [20]. After quality control, we phased and imputed the 487,409 UK Biobank genotyped samples against the Trans-Omics for Precision Medicine (TOPMed) reference panel (see **Web resources**), which is composed of 60,039 multi-ethnic samples and 239,756,147 SNP and indel variants sequenced at high depth (30x). The phasing step was carried out on 81 chromosomal chunks with around 10,000 genotyped variants in each chunk using the software Eagle (with the "kbpwt" parameter set at 80,000) [26]. The imputation was carried out in 137 chromosomal chunks of around 20 Mbp in length with Mbp of total overlap on either side using the imputation tool Minimac4 (see **Web resources**). To increase computational efficiency, we imputed each of the chunks in batches of 10,000 samples at a time and then merged them back using BCFtools. Since Minimac4 imputes each sample independently, analyzing our samples in batches did not change their imputation estimates. However, this sampling would result in different imputation quality estimates for each batch, and thus we collapsed the estimates to generate imputation quality estimates across all the study samples. After imputation, we filtered out variants with estimated imputation accuracy of $R^2 < 0.1$, which left us with 177,895,992 variants.

Phenome generation

MGI. The MGI phenome was used as the discovery dataset and was based on the Ninth and Tenth Revision of the International Statistical Classification of Diseases (ICD9 and ICD10) code data for 30,702 unrelated, genotyped individuals of recent European ancestry. These ICD9 and ICD10 codes were aggregated to form up to 1,857

PheWAS traits using the PheWAS R package (as described in detail elsewhere[4, 27]). For each trait, we identified case and control samples. To minimize differences in age and sex distributions or extreme case-control ratios as well as to reduce computational burden, we matched up to 10 controls to each case using the R package "MatchIt" [28]. Nearest neighbor matching was applied for age and PC1-4 (using Mahalanobis-metric matching; matching window caliper/width of 0.25 standard deviations) and exact matching was applied for sex and genotyping array. A total of 1,578 case control studies with >50 cases were used for our analyses of the MGI phenome.

UK Biobank. The UK Biobank phenome was used as a replication dataset and was based on ICD9 and ICD10 code data of 408,961 white British [14], genotyped individuals that were aggregated to PheWAS traits in a similar fashion (as described elsewhere [7]). To remove related individuals and to retain larger sample sizes, we first selected a maximal set of unrelated cases for each phenotype (defined as no pairwise relationship of 3rd degree or closer [24, 29]) before selecting a maximal set of unrelated controls unrelated to these cases. Similar to MGI, we matched up to 10 controls to each case using the R package "MatchIt" [28]. Nearest neighbor matching was applied for birthyear and PC1-4 (using Mahalanobis-metric matching; matching window caliper/width of 0.25 standard deviations) and exact matching was applied for sex and genotyping array. 1,366 case control studies with >50 cases each were used for our analyses of the UK Biobank phenome.

Additional phenotype information for MGI and UK Biobank is included in **S1 Text Fig B** and **S2 Text Tables F-H**.

Risk SNP selection

For each skin cancer subtype (melanoma, basal cell carcinoma, and squamous cell carcinoma), we generated three different sets of PRS: (1) based on merged summary statistics published in the NHGRI EBI GWAS catalog [1], (2) based on the latest available GWAS meta-analysis [30-32] and (3) based on publicly available GWAS summary statistics from the UK Biobank data [7].

GWAS Catalog SNP selection. We downloaded previously reported GWAS variants from the NHGRI-EBI GWAS Catalog (file date: February 28, 2018) [1, 33]. None of the currently available skin cancer discovery studies included in the catalog used any subset of the MGI cohort or data from the UK Biobank. Single nucleotide polymorphism (SNP) positions were converted to GRCh37 using variant IDs from dbSNP: build 150 (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. If a reported risk allele did not match any of the reported forward strand alleles of a non-ambiguous SNP (not A/T or C/G) in the imputed genotype data (which correspond to the alleles of the imputation reference panel), we assumed minus strand designation and corrected the effect allele to its complementary base of the forward strand. Entries with a reported risk allele that did not match any of the alleles of an ambiguous SNP (A/T and C/G) in our data were excluded at this step. We only included entries with broad European ancestry (as reported by the NHGRI-EBI GWAS Catalog). As a quality control check, we compared the reported risk allele frequencies (RAF) in controls with the RAF of 14,770 MGI individuals who had no cancer diagnosis (for chromosome X variants, we calculated RAF in females only). We then excluded

306 entries whose RAF deviated more than 15%. This chosen threshold is subjective and
307 was based on clear differentiation between correct and likely flipped alleles on the two
308 diagonals (see **S1 Text Fig A**) as noted frequently in GWAS meta-analyses quality
309 control procedures [34]. For each analyzed cancer type, we extracted risk variants that
310 were also present in our genotype data and estimated pairwise linkage disequilibrium
311 (LD; correlation r^2) using the allele dosages of the corresponding controls. For pairwise
312 correlated SNPs (r^2>0.1) or SNPs with multiple entries, we kept the SNP with the most
313 recent publication date (and smaller *P* value, if necessary) and excluded the other (**S2**
314 **File Table I**).

315 **Selection of risk SNPs from largest GWAS.** In a similar fashion, we extracted
316 and filtered reported association signals from large GWAS meta-analyses on basal cell
317 carcinoma [31], cutaneous squamous cell carcinoma [30] and melanoma [32] (**S2 File**
318 **Table I**).

319 **Genome-wide SNP selection of UK-Biobank-based GWAS.** We obtained
320 GWAS summary statistics for the ICD9- and ICD10-based PheWAS codes “172” (skin
321 cancer; 13,752 cases versus 395,071 controls), “172.11” (melanoma; 2,691 cases
322 versus 395,071 controls), and “172.2” (non-epithelial skin cancer; 11,149 cases versus
323 395,071 controls) from a public download [7] (see **Web resources**). These GWAS
324 analyzed up to 408,961 white British European-ancestry samples with generalized
325 mixed model association tests that used the saddlepoint approximation to calibrate the
326 distribution of score test statistics and thus could control for unbalanced case-control
327 ratios and sample relatedness [7]. For each trait, we reduced these summary statistics
328 to SNPs that were reported with minor allele frequencies > 0.5% and were also

329 available for the MGI data. Next, we performed linkage LD clumping of all variants with
330 p-values < $5x10^{-4}$ using the imputed allele dosages to obtain independent risk SNPs (LD
331 threshold of r^2 > 0.1 and a maximal SNP distance of 1 Mb). We limited the LD
332 calculations to 10,000 randomly selected, unrelated, white British individuals to reduce
333 the computational burden. Finally, we created subsets of these independent SNPs with
334 p-values <$5x10^{-9}$, <$5x10^{-8}$, <5 $x10^{-7}$, <$5x10^{-6}$, <$5x10^{-5}$, and <$5x10^{-4}$ (**S2 File Table J**).
335
336 **Construction of the polygenic risk scores**
337 For each of the obtained SNP sets for each trait, we constructed a PRS as the
338 sum of the allele dosages of risk increasing alleles of the SNPs weighted by their
339 reported log odds ratios. Restated, the PRS for subject j in MGI was of the form
340 $PRS_j=\sum_i \beta_i G_{ij}$ where *i* indexes the included loci for that trait, β_i is the log odds ratios
341 retrieved from the external GWAS summary statistics for locus *i,* and G_{ij} is a continuous
342 version of the measured dosage data for the risk allele on locus *i* in subject *j.* The PRS
343 variable was created for each MGI and UKB participant. For comparability of effect
344 sizes corresponding to the continuous PRS across cancer traits and PRS construction
345 methods, we transformed each PRS to the standard Normal distribution using
346 "ztransform" of the R package "GenABEL" [35].
347
348 **Statistical analysis**
349 In this study, we constructed PRS for three skin cancer subtypes using two
350 different PRS construction methods (using the Latest GWAS or the corresponding
351 entries of the GWAS Catalog). To compare the association between PRS and skin

cancer phenotypes across different PRS construction methods, we fit the following model for each PRS and skin cancer phenotype:

logit (P(Phenotype is present | PRS, Age, Sex, Array, PC)) $=\beta_0 + \beta_{PRS}\mathrm{PRS} + \beta_{Age}\mathrm{Age} + \beta_{Sex}\mathrm{Sex} + \beta_{Array}\mathrm{Array} + \boldsymbol{\beta}\,\mathrm{PC}$, 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 genotyping array. Our primary interest is in β_{PRS}, while the other factors (Age, Sex and PC) were included to address potential residual confounding and do not provide interpretable estimates due to the preceding application of case control matching. Firth's bias reduction method was used to resolve the problem of separation in logistic regression (Logistf in R package "EHR") [36-38], a common problem for binary or categorical outcome models when for a certain part of the covariate space there is only one observed value of the outcome, which often leads to very large parameter estimates and standard errors.

We then evaluated each PRS's (1) ability to discriminate between cases and controls by determining the area under the receiver-operator characteristics (ROC) curve (AUC) using R package "pROC" [39]; (2) calibration using Hosmer-Lemeshow Goodness Of Fit test of the R package "ResourceSelection" [40, 41]; and (3) accuracy with the Brier Score of R package "DescTools" [42]. These evaluations did not adjust for additional covariates. We used these metrics and the logistic regression results to choose a PRS construction method to use for each skin cancer subtype moving forward. To explore the impact of incorporating non-significant loci into the PRS construction, we further performed the above analyses with PRS constructed using UK Biobank GWAS summary statistics with different p-value thresholds.

Using the chosen PRS for each subtype, we conducted two PheWAS to identify other phenotypes associated with the PRS first for the 1,578 phenotypes in MGI and then for the 1,366 phenotypes from UK Biobank. To evaluate PRS-phenotype associations, we conducted Firth bias-corrected logistic regression by fitting a model of the above form for each phenotype and data source. Age represents the birth year in UK Biobank. To adjust for multiple testing, we applied the conservative phenome-wide Bonferroni correction according to the analyzed PheWAS codes (n_{MGI} = 1,578 or $n_{UK\ Biobank}$ = 1,366). In Manhattan plots, we present –log10 (*p*-value) corresponding to tests of $H_0: \beta_{PRS} = 0$. Directional triangles on the PheWAS plot indicate whether a phenome-wide significant trait was positively (pointing up) or negatively (pointing down) associated with the PRS.

To investigate the possibility of the secondary trait associations with PRS being completely driven by the primary trait association, we performed a second set of PheWAS after excluding individuals affected with the primary or related cancer traits for which the PRS was constructed, referred to as “exclusion PRS PheWAS” as described previously [4]. We then constructed new PRS scores representing shared and subsite-unique genetic components and performed a PheWAS for each.

To evaluate how well prior presence of an identified secondary non-skin-cancer diagnosis can identify subjects with increased risk of developing skin cancer, we created a binary variable taking the value 1 if a given subject (1) was diagnosed with the non-skin-cancer diagnosis and then diagnosed with skin cancer at least 365 days after or (2) was diagnosed with the non-skin-cancer diagnosis and never diagnosed with skin cancer. We then fit a Firth bias-corrected logistic regression of the following form:

398 logit (P(Primary phenotype is present | Predictor, Age, Sex, Array, PC))

399 $=\beta_0 + \beta_{PRS} I(\text{Secondary non skin cancer trait}) + \beta_{Age}\text{Age} + \beta_{Sex}\text{Sex} + \beta_{Array}\text{Array} + \boldsymbol{\beta}\ \text{PC}$

400 where Array and PC were defined as before. Unless otherwise stated, analyses were

401 performed using R 3.4.4 [43].

402

403 **Development of an online visual catalog: *PRSweb***

404 The online open access online visual catalog *PRSweb* available at

405 https://statgen.github.io/PRSweb was implemented using “Pandas”, a Data Analysis

406 Library, which offers high level of performance for large data structures and data

407 analysis in the Python3 environment [44]. In combination with “Jinja2”, a templating

408 language for Python, and “Bootstrap”, a Cascading Style Sheets (CSS) framework (see

409 Web resources), static HTML files were compiled and allow easy and fast hosting of all

410 PRS-PheWAS results. The interactive plots are drawn with the JavaScript library

411 “LocusZoom.js” (see Web resources) offers dynamic plotting, automatic plot sizing and

412 label positioning.

Results

Assessing various PRS construction methods

We first explored the comparative performance of two PRS construction strategies in terms of the resulting PRS associations with related phenotypes in the skin cancer setting. **Table 2** provides the results.

Table 2. Associations of constructed PRS with skin cancer traits in MGI

PRS (Number of SNPs)		Skin Cancer n = 4,503	Melanoma n = 1,896	Basal Cell Carcinoma n = 1,303	Squamous Cell Carcinoma n = 836
PRS based on GWAS Catalog					
Melanoma (29)	PRS OR[a] P-value[a] AUC[b] HL χ^2, P-value [c] Brier Score	1.41 (1.35,1.47) 2.7×10^{-53} 0.57 (0.56,0.58) 10,0.24 0.14	**1.68 (1.57,1.79)** 1.3×10^{-53} **0.61 (0.60,0.62)** 5.3,0.72 0.09	1.42 (1.31,1.53) 7.3×10^{-19} 0.57 (0.56,0.59) 12,0.16 0.091	1.3 (1.19,1.44) 4.3×10^{-08} 0.55 (0.53,0.57) 3.7,0.89 0.09
Basal cell carcinoma (32)	PRS OR[a] P-value[a] AUC[b] HL χ^2, P-value [c] Brier Score	1.39 (1.33,1.44) 8×10^{-60} 0.57 (0.56,0.58) 13,0.12 0.14	1.37 (1.29,1.45) 4.8×10^{-25} 0.57 (0.56,0.58) 8.6,0.38 0.091	**1.82 (1.70,1.95)** 3.6×10^{-65} **0.64 (0.62,0.65)** 9.5,0.3 0.09	1.4 (1.28,1.52) 1.4×10^{-14} 0.57 (0.55,0.59) 13,0.11 0.09
Squamous cell carcinoma (18)	PRS OR[a] P-value[a] AUC[b] HL χ^2, P-value [c] Brier Score	1.28 (1.24,1.33) 4.8×10^{-42} 0.56 (0.56,0.57) 4.7,0.79 0.14	1.35 (1.28,1.43) 2×10^{-28} 0.58 (0.56,0.59) 5.3,0.72 0.091	**1.39 (1.31,1.48)** 7.9×10^{-26} **0.59 (0.57,0.60)** 5.1,0.75 0.091	1.29 (1.19,1.39) 1.8×10^{-10} 0.56 (0.54,0.59) 7.8,0.46 0.09
PRS based on Latest GWAS					
Melanoma (20)	PRS OR[a] P-value[a] AUC[b] HL χ^2, P-value [c] Brier Score	1.48 (1.41,1.55) 3.5×10^{-55} 0.57 (0.56,0.58) 3,0.93 0.14	**1.78 (1.65,1.92)** 7×10^{-53} **0.61 (0.59,0.62)** 6.7,0.56 0.09	1.60 (1.47,1.75) 7.9×10^{-27} 0.59 (0.57,0.60) 2.5,0.96 0.091	1.38 (1.24,1.53) 4×10^{-09} 0.56 (0.54,0.58) 4.3,0.83 0.09
Basal cell carcinoma (28)	PRS OR[a] P-value[a] AUC[b] HL χ^2, P-value [c] Brier Score	1.42 (1.36,1.48) 5.8×10^{-61} 0.58 (0.57,0.58) 4.3,0.83 0.14	1.43 (1.34,1.52) 7×10^{-29} 0.58 (0.56,0.59) 16,0.051 0.091	**1.84 (1.71,1.97)** 2.8×10^{-60} **0.63 (0.62,0.65)** 4,0.86 0.09	1.45 (1.32,1.58) 1.2×10^{-15} 0.57 (0.55,0.60) 17,0.035 0.09
Squamous cell carcinoma (10)	PRS OR[a] P-value[a] AUC[b] HL χ^2, P-value [c] Brier Score	1.44 (1.38,1.5) 1.1×10^{-70} 0.58 (0.57,0.59) 17,0.027 0.14	1.54 (1.45,1.64) 2.9×10^{-46} 0.60 (0.58,0.61) 13,0.13 0.09	**1.62 (1.52,1.73)** 1.8×10^{-43} **0.61 (0.60,0.63)** 6,0.64 0.09	1.52 (1.39,1.65) 2.1×10^{-21} 0.59 (0.57,0.61) 4.9,0.76 0.09

421 [a] Association of each cancer with continuous PRS that were transformed to standard normal
422 distribution. Point estimates, 95% confidence intervals and P- values are obtained by fitting Firth's
423 Bias-Corrected Logistic Regression.
424 [b] Area under the curve of the receiver operating characteristic (ROC) curve with 95% confidence
425 intervals.
426 [c] Hosmer-Lemeshow Goodness-of-Fit Test
427

428 **Comparisons within methods.** Using the GWAS Catalog construction method,
429 the melanoma PRS was more strongly associated with and had better discrimination for
430 the melanoma phenotype than the other skin cancer phenotypes. For the PRS based on
431 the GWAS Catalog, the odds ratio (OR) of the melanoma PRS was 1.68 (95% CI, [1.57,
432 1.79]). By "discrimination," we refer to the ability of the PRS to distinguish melanoma
433 cases and controls, which is measured by AUC. The melanoma PRS AUC for the
434 melanoma phenotype is 0.61 (95 % CI, [0.60, 0.62]). Similarly, the basal cell carcinoma
435 PRS was most strongly associated with and had the best discrimination for the basal
436 cell carcinoma phenotype, with an OR of 1.82 (95% CI, [1.70, 1.95]) and an AUC of
437 0.64 (95% CI, [0.62, 0.65]). Unlike the other cancer subtypes, the squamous cell
438 carcinoma PRS did not appear to be most strongly associated with the squamous cell
439 carcinoma phenotype. Instead, it was most strongly associated with and most
440 discriminative for basal cell carcinoma. For all three skin cancer subtypes, the PRS
441 produced higher Brier scores for overall skin cancer, suggesting that the subtype-
442 defined PRS were less accurate for predicting skin cancer as a whole. We obtain similar
443 conclusions for the Latest GWAS method.

444 **Comparisons across methods.** For each cancer subtype, we compared the
445 PRS-subtype associations for the two PRS construction methods. **Melanoma**: For the
446 melanoma PRS, the GWAS Catalog method and the Latest GWAS method produced
447 similar performance in terms of AUC, OR, Hosmer-Lemeshow goodness of fit, and Brier
448 score. For example, the AUC for melanoma for the GWAS Catalog melanoma PRS was
449 0.61 (95% CI, [0.60, 0.62]). The corresponding AUC for the Latest GWAS method was
450 0.61 (95% CI, [0.59, 0.62]). **S1 Text Fig J** compares PRS weights to corresponding
451 SNP-melanoma associations in MGI and UK Biobank. **Basal Cell Carcinoma**: As with
452 melanoma, the basal cell carcinoma PRS produced similar results under the GWAS
453 Catalog and Latest GWAS construction methods. The basal cell carcinoma AUC under
454 the GWAS catalog method was 0.64 (95% CI, [0.62, 0.65]) and the AUC under the
455 Latest GWAS method was 0.63 (95% CI, [0.62, 0.65]). The OR values and Brier score
456 values were nearly identical, and neither approach produced evidence of lack of fit
457 based on the Hosmer-Lemeshow statistic. **Squamous Cell Carcinoma:** The squamous
458 cell carcinoma PRS was not more strongly associated with the squamous cell
459 carcinoma phenotype than the other phenotypes. However, we do observe that the
460 squamous cell carcinoma phenotype using the GWAS Catalog method (0.56, 95% CI
461 [0.54, 0.59]) produced a lower AUC compared to the Latest GWAS method (0.59, 95%
462 CI [0.57, 0.61]). While a difference of 0.03 may not seem like a large difference in AUC
463 in other applications, any improvement in AUC for PRS associations with observed
464 phenotypes may be considered appreciable [45]. These two methods produced identical
465 Brier scores, and the Latest GWAS method resulted in a stronger association between

the PRS and the squamous cell carcinoma phenotype (OR of 1.29, 95% CI [1.19, 1.39] vs OR of 1.52, 95% CI [1.39, 1.65]).

Using the above comparisons between the two PRS construction methods, we chose a single PRS construction method for each skin cancer subtype to use in subsequent analyses. For melanoma and basal cell carcinoma, we chose the GWAS Catalog method. While the GWAS Catalog and Latest GWAS methods were very similar for these two subtypes, we chose to pursue the GWAS Catalog PRS for future analysis due to the larger number of loci for these PRS (29 vs 20 for melanoma and 32 vs 28 for basal cell carcinoma). We choose the Latest GWAS method for squamous cell carcinoma due to its improved AUC over the GWAS Catalog method. We will denote the chosen PRS for melanoma, basal cell carcinoma, and squamous cell carcinoma as mPRS, bPRS, and sPRS respectively.

PheWAS using the chosen PRS in MGI

Using each of the chosen PRS described above (mPRS, bPRS, and sPRS), we tested the association between each PRS and each of the 1,578 constructed phenotypes in MGI. For each PRS, the strongest associations were observed with dermatologic neoplasms that included overall skin cancer, melanoma, “other non-epithelial cancer of skin” (the PheWAS over-category of basal and squamous cell carcinoma), and carcinoma in situ of skin. In addition, secondary dermatologic traits such as actinic keratosis (with over-category “degenerative skin conditions and other dermatoses”), chronic dermatitis due to solar radiation (with over-category “dermatitis due to solar radiation”), and seborrheic keratosis were found to be associated with all

three PRS (**Fig 1** and **S2 File Table K**). mPRS was most strongly associated with the melanoma phenotype (OR 1.67, 95% CI [1.56, 1.79]), while bPRS was most strongly associated with carcinoma in situ of the skin (OR 1.51, 95% CI [1.39, 1.64]) followed closely by "non-epithelial cancer of the skin" (OR 1.47, 95% CI [1.41, 1.54]). sPRS was most strongly associated with carcinoma in situ of the skin (OR 1.79, 95% CI [1.65, 1.94]). The OR of all these phenotypes indicated an increased risk for primary and secondary traits with increasing PRS.

Validation of PRS-PheWAS in UK Biobank

To substantiate the detected dermatologic associations, we reiterated the association screen of the three PRS in the matched phenome of the population-based UK Biobank data set (**Fig 1**). In general, stronger evidence for association was found in UKB compared to MGI. This may be driven by the larger sample sizes, e.g. a total of 13,623 skin cancer cases versus 4,503 in MGI. In the UK Biobank phenome, the large majority of the previous associations with dermatologic neoplasms were validated with the exception of the trait "dermatitis due to solar radiation", which had substantially fewer cases in UKB compared to MGI (390 versus 2,959 cases). Unlike MGI, all three PRS were significantly associated (at the phenome-wide level) with "cancer, suspected or other" and "malignant neoplasm, other."

Exclusion PheWAS using the chosen PRS in MGI

In order to explore whether the identified PRS-phenotype associations were driven by the primary trait used to define the PRS (for example, as a side effect of

512 treatment given after diagnosis with the primary trait), we performed a PheWAS for
513 each PRS in which we excluded subjects who were cases for the primary trait or other
514 skin cancer subtypes [4]. Results are shown in **S2 File Table K** and **S1 Text Fig C**.
515 Actinic keratosis, a skin condition believed to be a precursor to non-melanoma skin
516 cancers, remained significantly associated with the squamous cell carcinoma PRS in
517 MGI and all three PRS in UK Biobank [46][47, 48]. No other phenotypes were significant
518 for MGI. “Sebaceous cyst” and its over-category “diseases of the sebaceous gland”
519 were significant in the main UK Biobank PheWAS and remained significantly associated
520 with basal cell carcinoma PRS and squamous cell carcinoma PRS in UK Biobank in the
521 Exclusion PheWAS.
522
523 **Sub-analysis of actinic keratosis as a predictor of future skin cancer**
524 Actinic keratosis (AK) is a rough, scaly patch of skin that usually develops after
525 years of cumulative skin exposure [49]. Previous research has identified actinic
526 keratosis as a common pre-malignant condition for squamous cell carcinoma (SCC)
527 [46]. Actinic keratosis has also been identified as a potential precursor to basal cell
528 carcinoma (BCC) [47, 48]. The availability of temporal information of diagnoses in the
529 MGI cohort offered the opportunity to explore actinic keratosis as a potential precursor
530 for development of skin cancer in MGI.
531 **Fig 2** shows the ROC curves and AUC values for diagnosis of actinic keratosis at
532 least one year before any skin cancer diagnosis and its association with future BCC or
533 SCC diagnosis. AK diagnosis alone has little discrimination abilities, with AUC values of
534 0.52 (95% CI [0.51, 0.53]) for BCC and 0.51 (95% CI [0.50, 0.61]) for SCC. The bPRS

and sPRS provide comparatively good discrimination SCC (AUC 0.63 [0.62, 0.65] for BCC and 0.59 [0.57, 0.61] for SCC). The combination of prior AK diagnosis and bPRS provided further improvement in discrimination, with an AUC of 0.65 (95% CI [0.64, 0.67]).

S1 Text Tables A and B provides odds ratio estimates relating AK and the PRS to future BCC and SCC diagnosis. In unadjusted models, the odds of BCC diagnosis were significantly higher in subjects with a prior actinic keratosis diagnosis (OR 1.46, 95% CI [1.18, 1.80]). Notably, when we adjust for both bPRS and AK diagnosis, the unadjusted and adjusted effects of both variables are similar, suggesting that AK diagnosis may be an independent predictor of future BCC diagnosis. In contrast, AK diagnosis was **not** an independent predictor of SCC diagnosis. **S1 File Fig G** shows the timing of an AK diagnosis relative to a skin cancer diagnosis for patients with both diagnoses. For subjects with basal cell carcinoma or squamous cell carcinoma, AK diagnoses tended to occur prior to the skin cancer diagnosis (often within 8 years).

PRS-PheWAS for shared and unique loci across skin cancer subtypes

In the PRS-PheWAS analyses, we note a striking overlap in the secondary dermatological traits significantly associated with each of the three PRS (mPRS, bPRS, sPRS). One potential explanation for this is that subjects may have more screening after an initial skin cancer diagnosis. Indeed, many subjects have multiple skin cancer diagnoses (**S1 Text Fig D**). **Fig 3** shows the number of risk loci shared by different PRS. Six risk loci are shared between the mPRS, bPRS, and sPRS.

This observation inspired follow-up exploration in which we defined a PRS for each cancer subtype using the loci unique to that subtype's chosen PRS. We call these new PRS scores mPRS-u, bPRS-u, and sPRS-u, which reflect the unique loci in the PRS for melanoma, basal cell carcinoma, and squamous cell carcinoma respectively. We also define a PRS consisting of all loci shared across the three skin cancer subtypes, which we call the shared PRS.

S1 Text Table C shows the association between the various constructed PRS and the skin cancer phenotypes. As with mPRS, mPRS-u was most strongly associated with the melanoma phenotype and is not significantly associated with the other skin cancer subtypes. The bPRS-u score was similarly most strongly associated with basal cell carcinoma and not significantly associated with the other subtypes. We note that the melanoma AUC for the mPRS score was 0.61 (95% CI, [0.60, 0.62]) and is only 0.55 (95% CI, [0.53, 0.55]) for the mPRS-u score. Similarly, the basal cell carcinoma AUC for the bPRS score was 0.64 (95% CI, [0.62, 0.65]) and is only 0.57 (95% CI, [0.55, 0.58]) for the bPRS-u score. The sPRS-u score is not more strongly associated with the squamous cell carcinoma phenotype than the other skin cancer subtypes. For this reason, we do not include this PRS in further analyses. The shared PRS constructed as the unweighted sum of risk alleles of loci present in all three PRS scores (mPRS, bPRS, and sPRS) is more strongly associated with all three subtype phenotypes than the overall skin cancer phenotype, and the overall skin cancer phenotype also has the lowest AUC and highest Brier score.

S1 Text Fig E shows PRS-PheWAS results using mPRS-u and bPRS-u. The scores again reveal their subtype specificity in both phenomes, while no secondary

associations were observed. Although not shown here, additional exploration into the loci identified uniquely for each subtype may provide some insight into subtype-specific biological mechanisms. **S1 Text Fig F** shows PRS-PheWAS results for the shared PRS. Most strikingly, the shared skin cancer PRS was associated with the top skin cancer and dermatologic traits that were previously found to be associated with the three partially overlapping PRS constructs, suggesting that a shared genetic risk may be driving many of these secondary associations. These six underlying loci (*HERC2* [MIM 605837] /*OCA2* [MIM 611409], *IRF4* [MIM 601900], *MC1R* [MIM 155555], *RALY* [MIM 614663], *SLC45A2* [MIM 606202] and *TYR* [MIM 606933]) were previously found to be associated not only with skin cancer traits, but also with pigmentation traits of skin, eyes and hair (**Fig 3;** MIM 266300) [31, 50-69].

One of these pigmentation traits, skin tanning ability, the tendency of skin to sunburn rather than to suntan, is a well-known risk factor for all skin cancer traits [69, 70]. A PRS based on the independent risk variants of a recent GWAS meta-analysis on skin tanning ability [70] was strongly associated with overall skin cancer, melanoma, basal cell carcinoma, and squamous cell carcinoma and even outperformed the constructed PRS of the former two traits (**S1 Text Table C**). Furthermore, the skin tanning ability PRS PheWAS identified a very similar set of traits as the shared skin cancer PRS but in general revealed stronger associations (**S1 Text Fig F**).

PRS construction based on UK Biobank summary statistics

The NHGRI-EBI GWAS Catalog and Latest GWAS PRS construction methods are based on published GWAS studies, which often only report risk variants that

603 reached genome-wide significance, but we may believe that incorporating additional risk
604 variance below this threshold may improve predictive power of a PRS. To explore
605 whether a PRS incorporating non-significant loci will outperform a PRS incorporating
606 only significant loci, we constructed PRS using loci related to the phenotype at six
607 different p-value thresholds based on publicly available GWAS summary statistics from
608 the UK Biobank. Larger p-values indicate greater SNP depth (with more SNPs being
609 incorporated into the PRS).

610 The collection of UK Biobank GWAS results did not include basal cell carcinoma
611 or squamous cell carcinoma subtypes; rather, it included only the merged trait 'non-
612 epithelial cancer of skin' (**S1 Text Fig B**). Thus, we limited our assessment of the
613 summary statistics to the overall skin cancer GWAS (UKB PheWAS code "172": 13,752
614 skin cancer cases versus 395,071 controls) and the melanoma GWAS (UKB PheWAS
615 code "172.11": 2,691 melanoma cases versus 395,071 controls) (**S2 File Table J**).

616 **S1 Text Table D** provides the results. As with the other PRS construction
617 methods, the melanoma PRS was most strongly associated with and discriminative for
618 the melanoma phenotype for all p-value cutoffs except $5x10^{-4}$. For this p-value cutoff,
619 the melanoma PRS had similar AUC and OR for the melanoma and basal cell
620 carcinoma phenotypes. This p-value cutoff represents the least conservative inclusion
621 cutoff with 1,193 included loci, and its results indicated that inclusion of too many
622 suggestive SNPs at lower thresholds may reduce PRS performance. However, we also
623 note that the most conservative cutoff ($5x10^{-9}$) produced a PRS with only six loci and a
624 weaker OR and AUC compare to other PRS created with less stringent cutoffs. Like the
625 other PRS construction methods, the melanoma PRS was less accurate for predicting

626 overall skin cancer compared to the individual skin cancer subtypes. The best
627 performance in terms of AUC and OR relating to the melanoma phenotype were
628 observed for p-value thresholds $5x10^{-7}$ and $5x10^{-8}$, which included 13 and 9 loci
629 respectively. The small number of loci identified by this method at more conservative p-
630 value cutoffs may be driven by the lower sample size for melanoma in the UK Biobank
631 compared to the published melanoma GWAS meta-analyses (n cases = 2,691 and n
632 cases = 6,628, respectively). We note that the melanoma PRS constructed using the
633 UK Biobank summary statistics produced lower AUC across all p-value thresholds than
634 was seen for the Latest GWAS and GWAS Catalog PRS construction methods.

635 The PRS constructed for overall skin cancer was most strongly associated with
636 and discriminative for basal cell carcinoma across all p-value thresholds, with AUCs
637 ranging from 0.59 (95% CI [0.57, 0.60]) to 0.64 (95% CI [0.62, 0.66]) and odds ratios
638 ranging from 1.42 (95% CI [1.33, 1.51]) to 1.73 (95% CI [1.63, 1.84]). The overall skin
639 cancer PRS had the highest Brier score for overall skin cancer, indicating that the
640 overall skin cancer PRS was more accurate at predicting the skin cancer subtypes
641 compared to overall skin cancer. The overall skin cancer PRS had very similar
642 association with and discrimination abilities for the overall skin cancer phenotype across
643 all p-value thresholds except the least conservative ($p = 5x10^{-4}$), for which the AUC and
644 odds ratio were smaller. Overall, the highest AUCs and strongest OR signals for both
645 PRS and all skin cancer phenotypes were found at depths of $5x10^{-7}$ and $5x10^{-8}$.

646 In addition to associations with the primary phenotype, we explored associations
647 between PRS constructed at various UK Biobank summary statistic depths and
648 secondary phenotypes. **Fig 3** (melanoma) and **S1 Text Fig H** (overall skin cancer) show

PRS-PheWAS results in MGI using PRS constructed at different depths. As shown in **S1 Text Table E and Fig I**, depths of $5x10^{-7}$ and $5x10^{-8}$ produced very similar results, and other depths identified fewer phenotypes associated with the corresponding PRS. Phenotypes that were associated with the PRS at other depths had weaker associations than those observed at $5x10^{-7}$ and $5x10^{-8}$.

Online visual catalog: *PRSweb*

For comparison of the aforementioned PRS-PheWAS results and to provide researchers with resources for future PRS-based analyses, we developed an open access, online visual catalog *PRSweb* available at https://statgen.github.io/PRSweb that enables interactive exploration of the PheWAS results for each of the skin cancer subtypes under each of three different PRS construction methods explored in this paper, for both the MGI and UK Biobank phenomes. *PRSweb* shows PRS-PheWAS plots with various choices of PRS in the drop-down menu (example screenshot in **Fig 5**) and offers downloadable PRS constructs (list of independent risk variants with corresponding weights). Mouse-over boxes offer detailed information about top results if needed, without impeding the overall user experience (grey box in **Fig 5**). Enrichment of cases in the tail of the PRS distribution are presented in interactive forest plots.

Discussion

PRS combine information from a large number of genetic variants to stratify subjects in terms of their risk for developing a particular disease. However, there are currently no general guidelines for how to construct a PRS for a given *EHR-derived* phenotype. In this paper, we explore strategies for constructing a PRS using markers

673 and weights obtained from various publicly-available sources. First, we consider PRS
674 constructed using markers and weights identified in either (1) the latest GWAS or
675 GWAS meta-analysis or (2) the NHGRI-EBI GWAS Catalog. We compare these two
676 PRS construction methods in terms of their associations with EHR-derived phenotypes
677 for the three most common skin cancer subtypes in the USA: basal cell carcinoma,
678 cutaneous squamous cell carcinoma, and melanoma.

679 A priori, we may have some belief that the latest (and often the largest) GWAS
680 may provide a better source of evidence to use for PRS construction due to larger
681 sample sizes and (potentially) more carefully curated data. The Latest GWAS and
682 GWAS Catalog methods produced PRS with similar performance in terms of their
683 associations with and discrimination for the primary phenotype used to construct the
684 PRS for both basal cell carcinoma and melanoma. Generally, PRS constructed for
685 melanoma and basal cell carcinoma were most strongly associated with and
686 discriminative for their target phenotypes, indicating that both PRS construction
687 methods were able to provide a higher degree of specificity for the intended skin cancer
688 subtype. In contrast, the PRS for squamous cell carcinoma were not more strongly
689 associated with the squamous cell carcinoma phenotype compared to other skin cancer
690 phenotypes. This may suggest a need for further exploration into genetic factors
691 uniquely related to the squamous cell carcinoma subtype.

692 For each skin cancer subtype, we performed a PRS-PheWAS to identify
693 secondary phenotypes that are associated with the corresponding PRS. We generally
694 identified many dermatological features in addition to the primary phenotype, indicating
695 the ability of PRS to reproduce associations with the primary phenotype even after

multiple testing corrections and covariate adjustment. The majority of these associations were replicated in a PRS-PheWAS performed for the UK Biobank phenome. Our analyses identified actinic keratosis, which is believed to be a precursor to squamous cell and basal cell carcinoma, as an independent predictor of basal cell and squamous cell carcinoma, and we demonstrated that incorporating the PRS in addition to clinical information improved discrimination for future skin cancer diagnoses [46-48].

In an additional analysis, we identified loci that were shared among all three skin cancer subtypes' PRS. Loci overlap between the PRS for the three subtypes may indicate factors related to common biology between the subtypes. We noted that all shared loci (*HERC2*/*OCA2*, *IRF4*, *MC1R*, *RALY*, *SLC45A2* and *TYR*) were also loci that had been associated with human pigmentation traits and/or harbor key genes of the biochemical pathway of melanogenesis [50, 54-62, 64, 67-71]. We constructed PRS using SNPs shared by all three skin cancer subtypes and a PRS for skin tanning ability using results from a recent GWAS meta-analysis.[70] The skin tanning ability PRS PheWAS identified a very similar set of traits to the shared PRS PheWAS, suggesting that the shared genetic component may in part represent genetic factors influencing the skin pigmentation and the skin reaction to sun exposure. However, the PRS that are unique to subtypes did not show such a common pathway or mechanism.

The Latest GWAS and the GWAS Catalog methods for constructing the PRS involve incorporating only loci that reached genome-wide significance for at least one study, as non-significant loci are usually not reported. However, incorporating non-significant loci that are associated with the primary phenotype may help improve the predictive ability of the PRS [8, 16]. We found that incorporating additional loci that

would not reach genome-wide significance did improve the PRS' ability to discriminate cases from controls for the primary phenotype up to a point. In particular, PRS constructed using SNPs with p-values less than $5x10^{-8}$ or $5x10^{-7}$ resulted in the best performance, but further increasing the p-value threshold resulted in reduced performance. Crucially, we also observed stronger associations between the PRS and secondary phenotypes for PRS constructed using depths of $5x10^{-8}$ and $5x10^{-7}$. These results suggest that some benefit may be seen by incorporating loci that do not reach significance into the PRS construction but incorporating too many loci with larger p-values may not improve the predictive ability of the PRS (for both primary and secondary phenotypes). However, this gain or reduction in PRS performance may depend on the phenotype of interest and on the prevalence of the phenotype in the analytical sample.

As a product of this study, we provide an online visual catalog *PRSweb* that provides PRS-PheWAS results for the various skin cancer phenotypes for PRS constructed using the different methods explored in this paper. *PRSweb* will provide a routine way to compare different PRS construction methods and to explore PRS-PheWAS results in detail. Additionally, *PRSweb* provides the PRS construction details, which researchers can download and use in their own analyses. In the future, we plan to extend this online platform to include PheWAS for many other cancer phenotypes, which will make this online platform a general tool for identifying phenotypes related to particular types of cancer.

One limitation of the generalizability of this study comes from the homogeneous race profile of MGI and UK Biobank. UK Biobank consists of subjects of primarily

742 European descent, and we restricted our analyses to subjects of European descent in
743 MGI (excluding about 10% of the subjects in MGI) in order to ensure greater
744 comparability between the two datasets. Additionally, many of the existing GWAS were
745 conducted on European populations, and we wanted to consider similar samples when
746 comparing the performance of PRS constructed using summary statistics from
747 European populations. Unlike UK Biobank, MGI is not a population-based sample;
748 rather, it is a sample of patients recruited from a large academic medical center.
749 Patients were recruited prior to surgery through the anesthesiology department, and
750 therefore they may present a potential for selection bias. Additionally, the comparative
751 performance of the PRS across construction methods will depend on the phenotype of
752 interest. In spite of these limitations, a principled comparison of the various methods
753 explored in this paper may provide researchers with a sense of the robustness of their
754 PheWAS inference to the PRS construction method and an analytical framework for
755 exploration of shared genetic architecture of related traits.
756

Acknowledgement

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. The presented research was funded by the National Cancer Institute at the National Institutes of Health (P30 CA046592 [LGF, LJB, MS, BM] and T32 CA83654 [RBP]). Part of this research has been conducted using the UK Biobank Resource under application number 24460. This material is based upon work supported by the National Science Foundation under Grant No. (NSF DMS 1712933). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Web resources

University of Michigan Medical School Central Biorepository; https://research.medicine.umich.edu/our-units/central-biorepository

UK Biobank; http://www.ukbiobank.ac.uk/

UK Biobank GWAS summary statistics; https://tinyurl.com/UKB-SAIGE

TOPMed variant browser, https://bravo.sph.umich.edu/freeze5/hg38/

TOPMed program, https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program

Minimac4; https://genome.sph.umich.edu/wiki/Minimac4

BCFtools; https://samtools.github.io/bcftools/bcftools.html

KING; http://people.virginia.edu/~wc9c/KING/

781 FASTINDEP; https://github.com/endrebak/fastindep

782 PLINK; https://www.cog-genomics.org/plink2/

783 Eagle; https://data.broadinstitute.org/alkesgroup/Eagle/

784 UCSC Genome Browser; http://genome.ucsc.edu/

785 R; https://cran.r-project.org/

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

787 dbSNP; https://www.ncbi.nlm.nih.gov/projects/SNP/

788 Imputation server; https://imputationserver.sph.umich.edu/

789 Jinja, https://github.com/pallets/jinja

790 Locuszoom, https://github.com/statgen/locuszoom

791 PRSweb; https://statgen.github.io/PRSweb

References

1. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901.
2. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17(10):1520-8.
3. Khera AV, Chaffin M, Aragam K, Emdin CA, Klarin D, Haas M, et al. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. bioRxiv. 2017.
4. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, et al. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet. 2018;102(6):1048-61.
5. Docherty AR, Moscati A, Dick D, Savage JE, Salvatore JE, Cooke M, et al. Polygenic prediction of the phenome, across ancestry, in emerging adulthood. Psychol Med. 2017:1-10.
6. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol. 2017.
7. Zhou W, Nielsen JB, Fritsche LG, Dey R, Elvestad MB, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. bioRxiv. 2017.
8. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9(3):e1003348.
9. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018.
10. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
11. Kauvar AN, Cronin T, Jr., Roenigk R, Hruza G, Bennett R, American Society for Dermatologic S. Consensus for nonmelanoma skin cancer treatment: basal cell carcinoma, including a cost analysis of treatment methods. Dermatol Surg. 2015;41(5):550-71.
12. Kallini JR, Hamed N, Khachemoune A. Squamous cell carcinoma of the skin: epidemiology, classification, management, and novel trends. Int J Dermatol. 2015;54(2):130-40.
13. Berwick M, Buller DB, Cust A, Gallagher R, Lee TK, Meyskens F, et al. Melanoma Epidemiology and Prevention. Cancer Treat Res. 2016;167:17-49.
14. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv. 2017.
15. Reisberg S, Iljasenko T, Lall K, Fischer K, Vilo J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. PLoS One. 2017;12(7):e0179238.

16. 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(18):3525-31.
17. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2015;31(9):1466-8.
18. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97(4):576-92.
19. So HC, Sham PC. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci Rep. 2017;7:41262.
20. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
21. Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, Chew EY, et al. Ancestry estimation and control of population stratification for sequence-based association studies. Nat Genet. 2014;46(4):409-15.
22. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-4.
23. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
24. Abraham KJ, Diaz C. Identifying large sets of unrelated individuals and unrelated markers. Source Code Biol Med. 2014;9(1):6.
25. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
26. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443-8.
27. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 2014;30(16):2375-6.
28. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software. 2011;42(8):1-28.
29. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.
30. Chahal HS, Lin Y, Ransohoff KJ, Hinds DA, Wu W, Dai HJ, et al. Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nat Commun. 2016;7:12048.
31. Chahal HS, Wu W, Ransohoff KJ, Yang L, Hedlin H, Desai M, et al. Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. Nat Commun. 2016;7:12510.
32. Ransohoff KJ, Wu W, Cho HG, Chahal HC, Lin Y, Dai HJ, et al. Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. Oncotarget. 2017;8(11):17586-92.

882 33. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI
883 GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res.
884 2014;42(Database issue):D1001-6.
885 34. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Magi R, et al.
886 Quality control and conduct of genome-wide association meta-analyses. Nat Protoc.
887 2014;9(5):1192-212.
888 35. GenABEL project developers. GenABEL: genome-wide SNP association
889 analysis. 2013.
890 36. Heinze G. A comparative investigation of methods for logistic regression with
891 separated or nearly separated data. Stat Med. 2006;25(24):4216-26.
892 37. Heinze G, Ploner M, Dunkler D, Southworth H. logistf: Firth's bias reduced
893 logistic regression. 2013.
894 38. Choi L, Beck C. EHR: Electronic Health Record (EHR) Data Processing and
895 Analysis Tool. 2017.
896 39. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an
897 open-source package for R and S+ to analyze and compare ROC curves. BMC
898 Bioinformatics. 2011;12:77.
899 40. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York, USA: John
900 Wiley and Sons; 2010.
901 41. Lele S, R., Keim JL, Solymos P. ResourceSelection: Resource Selection
902 (Probability) Functions for Use-Availability Data. 2017.
903 42. Signorell A. DescTools: Tools for Descriptive Statistics. 2018.
904 43. R Core Team. R: A Language and Environment for Statistical Computing. R
905 Foundation for Statistical Computing, Vienna, Austria; 2016.
906 44. McKinney W. Data Structures for Statistical Computing in Python2010.
907 45. Baker SG, Schuit E, Steyerberg EW, Pencina MJ, Vickers A, Moons KG, et al.
908 How to interpret a small increase in AUC with an additional risk prediction marker:
909 decision analysis comes through. Stat Med. 2014;33(22):3946-59.
910 46. Fuchs A, Marmur E. The kinetics of skin cancer: progression of actinic keratosis
911 to squamous cell carcinoma. Dermatol Surg. 2007;33(9):1099-101.
912 47. Cohen JL. Actinic keratosis treatment as a key component of preventive
913 strategies for nonmelanoma skin cancer. J Clin Aesthet Dermatol. 2010;3(6):39-44.
914 48. Jacobs RJ, Phillips G. Basal cell carcinoma mistaken for actinic keratosis. Clin
915 Exp Optom. 2006;89(3):171-5.
916 49. Ko CJ. Actinic keratosis: facts and controversies. Clin Dermatol. 2010;28(3):249-
917 53.
918 50. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacon-
919 Duque JC, et al. A genome-wide association scan in admixed Latin Americans identifies
920 loci influencing facial and scalp hair features. Nat Commun. 2016;7:10815.
921 51. Asgari MM, Wang W, Ioannidis NM, Itnyre J, Hoffmann T, Jorgenson E, et al.
922 Identification of Susceptibility Loci for Cutaneous Squamous Cell Carcinoma. J Invest
923 Dermatol. 2016;136(5):930-7.
924 52. Barrett JH, Iles MM, Harland M, Taylor JC, Aitken JF, Andresen PA, et al.
925 Genome-wide association study identifies three new melanoma susceptibility loci. Nat
926 Genet. 2011;43(11):1108-13.

927 53. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, et al.
928 Genome-wide association study identifies three loci associated with melanoma risk. Nat
929 Genet. 2009;41(8):920-5.
930 54. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, et al.
931 Web-based, participant-driven studies yield novel genetic associations for common
932 traits. PLoS Genet. 2010;6(6):e1000993.
933 55. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, et al. A genome-wide
934 association study identifies novel alleles associated with hair color and skin
935 pigmentation. PLoS Genet. 2008;4(5):e1000074.
936 56. Hernandez-Pacheco N, Flores C, Alonso S, Eng C, Mak AC, Hunstman S, et al.
937 Identification of a novel locus associated with skin colour in African-admixed
938 populations. Sci Rep. 2017;7:44548.
939 57. Jacobs LC, Hamer MA, Gunn DA, Deelen J, Lall JS, van Heemst D, et al. A
940 Genome-Wide Association Study Identifies the Skin Color Genes IRF4, MC1R, ASIP,
941 and BNC2 Influencing Facial Pigmented Spots. J Invest Dermatol. 2015;135(7):1735-
942 42.
943 58. Law MH, Medland SE, Zhu G, Yazar S, Vinuela A, Wallace L, et al. Genome-
944 Wide Association Shows that Pigmentation Genes Play a Role in Skin Aging. J Invest
945 Dermatol. 2017;137(9):1887-94.
946 59. Lin BD, Mbarek H, Willemsen G, Dolan CV, Fedko IO, Abdellaoui A, et al.
947 Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family
948 Based Sample. Genes (Basel). 2015;6(3):559-76.
949 60. Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, et al. Digital
950 quantification of human eye color highlights genetic association of three new loci. PLoS
951 Genet. 2010;6(5):e1000934.
952 61. Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, et al. Genetics of skin
953 color variation in Europeans: genome-wide association studies with functional follow-up.
954 Hum Genet. 2015;134(8):823-35.
955 62. Nan H, Kraft P, Qureshi AA, Guo Q, Chen C, Hankinson SE, et al. Genome-wide
956 association study of tanning phenotype in a population of European ancestry. J Invest
957 Dermatol. 2009;129(9):2250-7.
958 63. Nan H, Xu M, Kraft P, Qureshi AA, Chen C, Guo Q, et al. Genome-wide
959 association study identifies novel alleles associated with risk of cutaneous basal cell
960 carcinoma and squamous cell carcinoma. Hum Mol Genet. 2011;20(18):3718-24.
961 64. Rawofi L, Edwards M, Krithika S, Le P, Cha D, Yang Z, et al. Genome-wide
962 association study of pigmentary traits (skin and iris color) in individuals of East Asian
963 ancestry. PeerJ. 2017;5:e3951.
964 65. Siiskonen SJ, Zhang M, Li WQ, Liang L, Kraft P, Nijsten T, et al. A Genome-Wide
965 Association Study of Cutaneous Squamous Cell Carcinoma among European
966 Descendants. Cancer Epidemiol Biomarkers Prev. 2016;25(4):714-20.
967 66. Song F, Amos CI, Lee JE, Lian CG, Fang S, Liu H, et al. Identification of a
968 melanoma susceptibility locus and somatic mutation in TET2. Carcinogenesis.
969 2014;35(9):2097-101.
970 67. Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, et al. A
971 genomewide association study of skin pigmentation in a South Asian population. Am J
972 Hum Genet. 2007;81(6):1119-32.

973 68. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP,
974 et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet.
975 2007;39(12):1443-52.
976 69. Zhang M, Song F, Liang L, Nan H, Zhang J, Liu H, et al. Genome-wide
977 association studies identify several new loci associated with pigmentation traits and skin
978 cancer risk in European Americans. Hum Mol Genet. 2013;22(14):2948-59.
979 70. Visconti A, Duffy DL, Liu F, Zhu G, Wu W, Chen Y, et al. Genome-wide
980 association study in 176,678 Europeans reveals genetic loci for tanning response to sun
981 exposure. Nat Commun. 2018;9(1):1684.
982 71. Sturm RA, Duffy DL. Human pigmentation genes under environmental selection.
983 Genome Biol. 2012;13(9):248.
984

Supporting information

S1 File. Supporting Material. This file contains supporting Figures A-J and Tables A-E.

S2 File. Supporting Tables. This Excel file contains the following tables

- Sheet 1: Table F, ICD9 codes to PheWAS code translations
- Sheet 2: Table G, ICD10 codes to PheWAS code translations
- Sheet 3: Table H, MGI and UK Biobank Phenome
- Sheet 4: Table I, Risk SNP Selection
- Sheet 5: Table J, Risk SNP Selection Depth
- Sheet 6: Table K, Omnibus Table Significant Results

Figure Titles and Legends

Fig 1. PRS-PheWAS in MGI and UKB phenomes.

The horizontal line indicates phenome-wide significance.

Fig 2 Comparison of predictors.

Actinic keratosis (AK), at least 365 prior to any skin cancer diagnosis as predictor for basal cell carcinoma (BCC) (A and B) and squamous cell carcinoma (SCC) (C and D). The PRS for BCC and SCC as well as the combined predictors are shown for comparison.

Fig 3 Overlap between the three skin cancer trait loci.

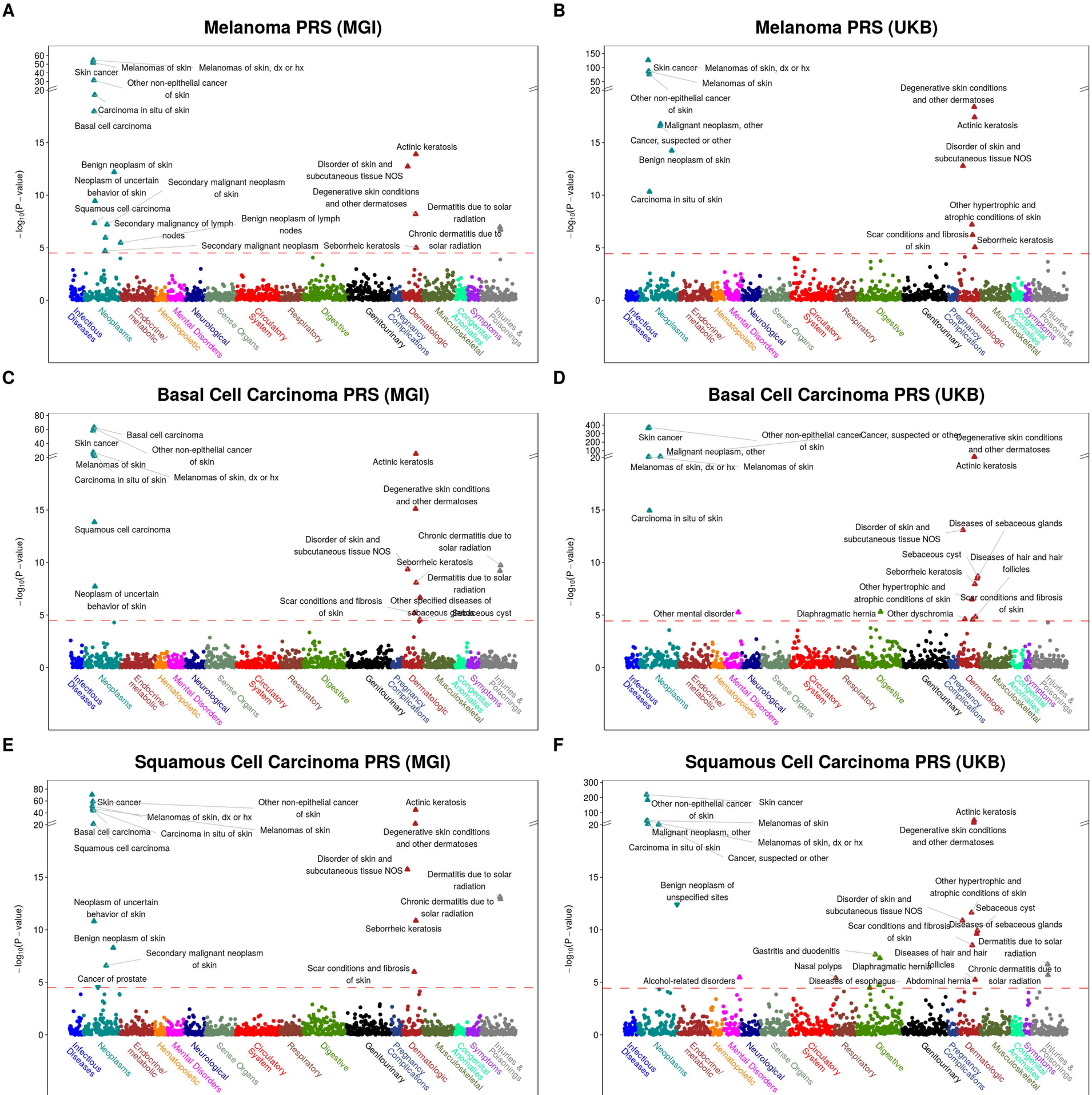
Reported risk SNPs within 1 Mb were merged into the same locus. Loci that were also reported to be associated with skin tanning ability are highlighted in bold. Loci were named according to the closest RefSeq genes (except *M1CR* a 385 kb locus with 16 RefSeq genes and *HV745896* named after a nearby, uncurated mRNA sequence).

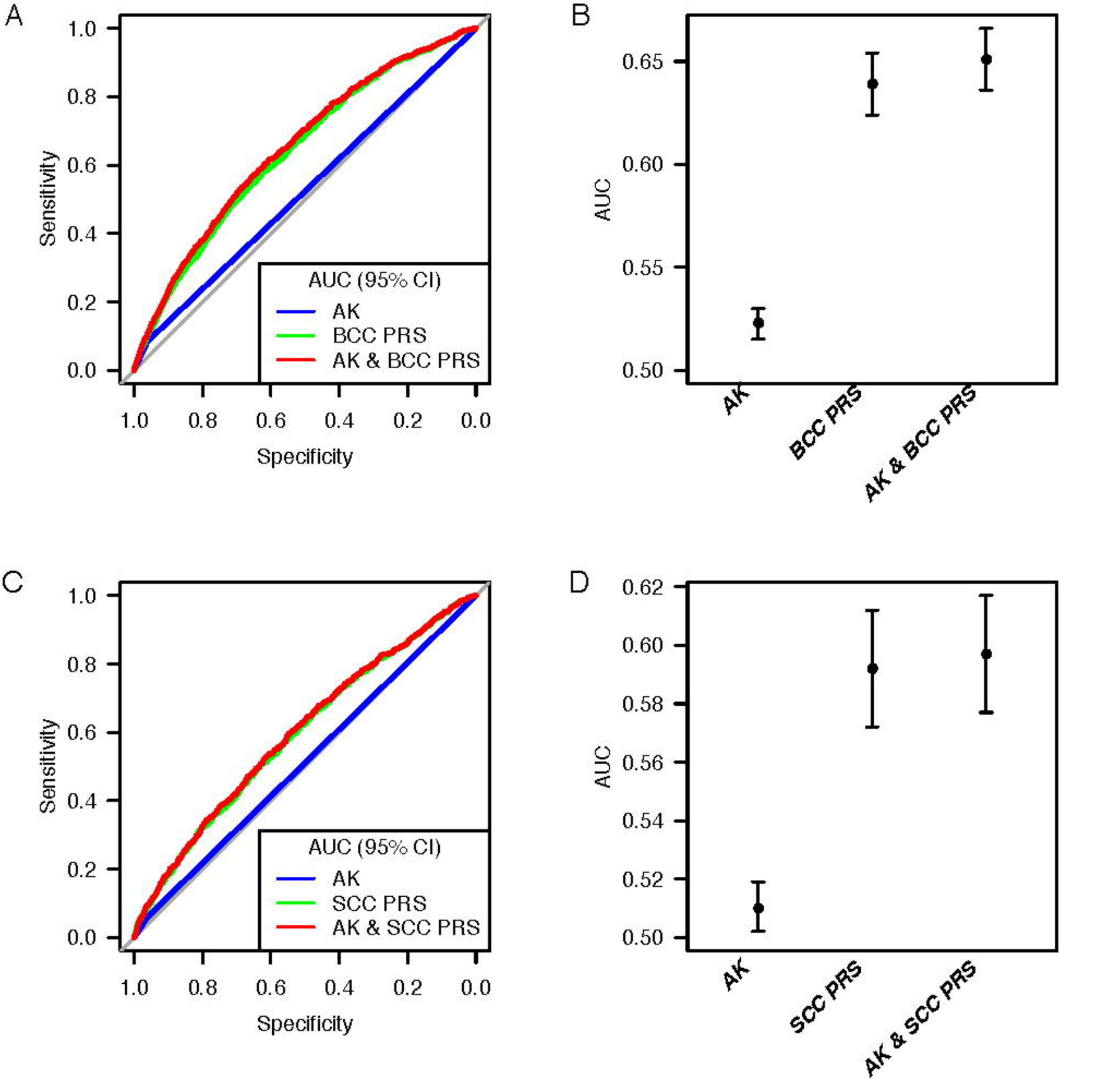
Fig 4 PheWAS on melanoma PRS constructed using UK Biobank statistics at different depths.

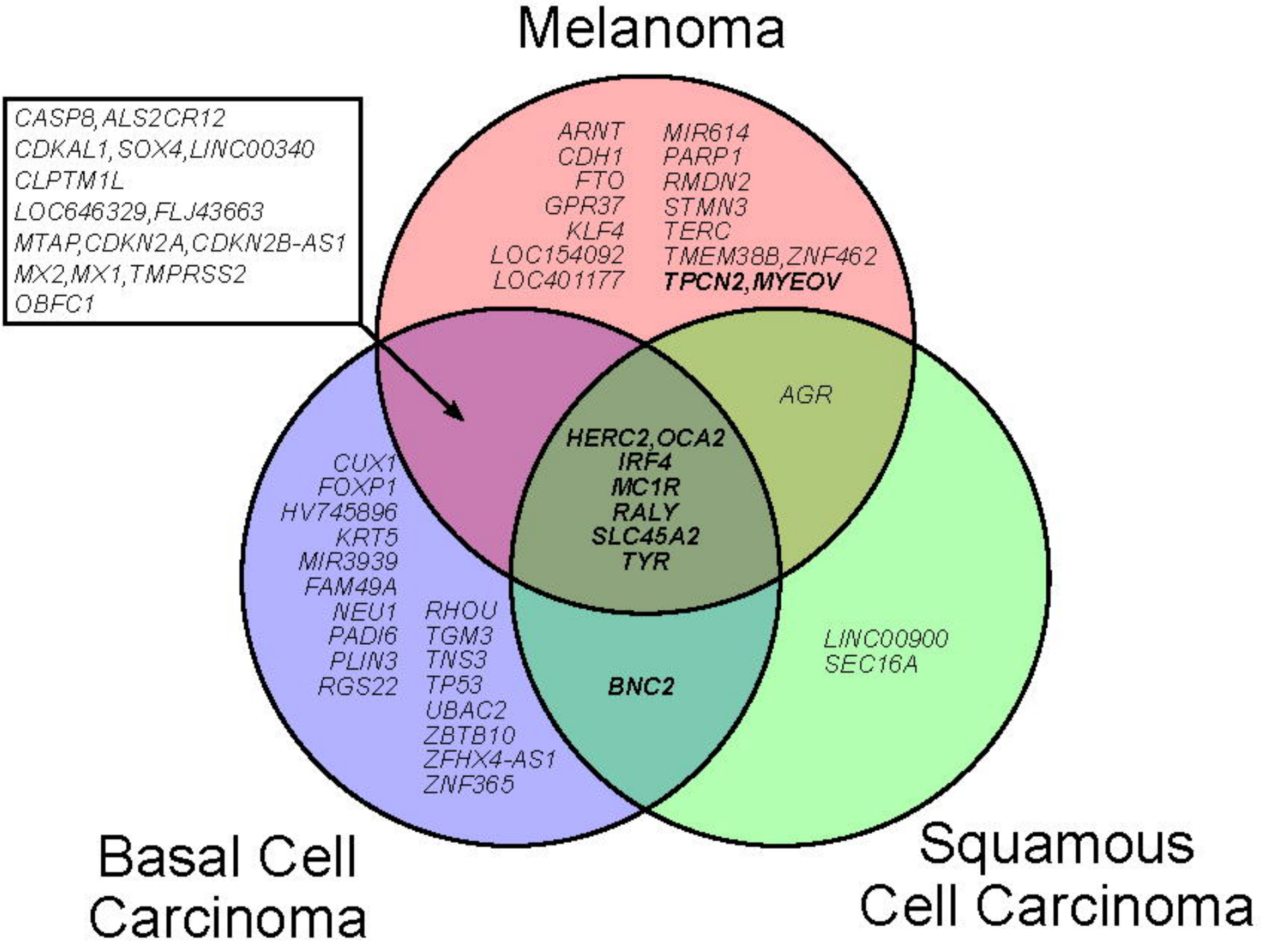
Results are shown with increasing depth from (A – F): P <= $5x10^{-9}$, $5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, $5x10^{-4}$.

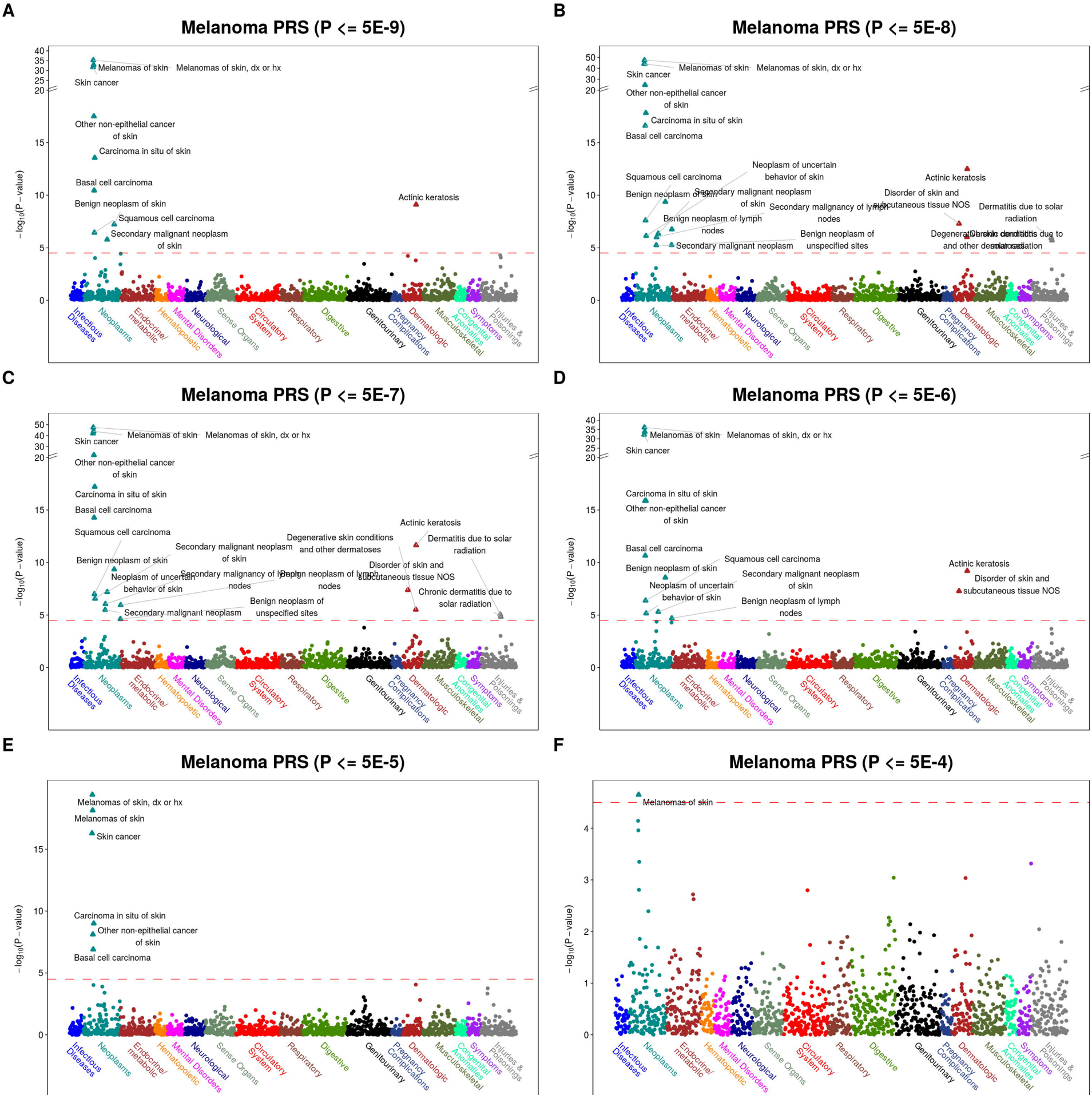
Fig 5. Example view from PRSweb. A selection menu on top allows selection of PRS constructs and phenome while interactive plots with “PheWAS results”, “Exclusion

PheWAS results", and "Associations between PRS and Selected Phenotype" are generated after selection.


A
Melanoma PRS (MGI)
B
Melanoma PRS (UKB)
C
Basal Cell Carcinoma PRS (MGI)
D
Basal Cell Carcinoma PRS (UKB)
E
Squamous Cell Carcinoma PRS (MGI)
F
Squamous Cell Carcinoma PRS (UKB)
−log10(P−value)
Infectious Diseases
Neoplasms
Endocrine/ metabolic
Hematopoietic
Mental Disorders
Neurological
Sense Organs
Circulatory System
Respiratory
Digestive
Genitourinary
Pregnancy Complications
Dermatologic
Musculoskeletal
Congenital Anomalies
Symptoms
Injuries & Poisonings


A
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
AUC (95% CI)
AK
BCC PRS
AK & BCC PRS
B
AUC
0.65
0.60
0.55
0.50
AK
BCC PRS
AK & BCC PRS
C
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
AUC (95% CI)
AK
SCC PRS
AK & SCC PRS
D
AUC
0.62
0.60
0.58
0.56
0.54
0.52
0.50
AK
SCC PRS
AK & SCC PRS


Melanoma
CASP8,ALS2CR12
CDKAL1,SOX4,LINC00340
CLPTM1L
LOC646329,FLJ43663
MTAP,CDKN2A,CDKN2B-AS1
MX2,MX1,TMPRSS2
OBFC1
ARNT
CDH1
FTO
GPR37
KLF4
LOC154092
LOC401177
MIR614
PARP1
RMDN2
STMN3
TERC
TMEM38B,ZNF462
TPCN2,MYEOV
AGR
HERC2,OCA2
IRF4
MC1R
RALY
SLC45A2
TYR
CUX1
FOXP1
HV745896
KRT5
MIR3939
FAM49A
NEU1
PADI6
PLIN3
RGS22
RHOU
TGM3
TNS3
TP53
UBAC2
ZBTB10
ZFHX4-AS1
ZNF365
BNC2
LINC00900
SEC16A
Basal Cell Carcinoma
Squamous Cell Carcinoma


A
Melanoma PRS (P <= 5E-9)
Melanomas of skin
Melanomas of skin, dx or hx
Skin cancer
Other non-epithelial cancer of skin
Carcinoma in situ of skin
Basal cell carcinoma
Benign neoplasm of skin
Squamous cell carcinoma
Secondary malignant neoplasm of skin
Actinic keratosis
B
Melanoma PRS (P <= 5E-8)
Skin cancer
Melanomas of skin
Melanomas of skin, dx or hx
Other non-epithelial cancer of skin
Carcinoma in situ of skin
Basal cell carcinoma
Squamous cell carcinoma
Benign neoplasm of skin
Neoplasm of uncertain behavior of skin
Secondary malignant neoplasm of skin
Secondary malignancy of lymph nodes
Benign neoplasm of lymph nodes
Secondary malignant neoplasm
Benign neoplasm of unspecified sites
Actinic keratosis
Disorder of skin and subcutaneous tissue NOS
Dermatitis due to solar radiation
C
Melanoma PRS (P <= 5E-7)
Skin cancer
Melanomas of skin
Melanomas of skin, dx or hx
Other non-epithelial cancer of skin
Carcinoma in situ of skin
Basal cell carcinoma
Squamous cell carcinoma
Benign neoplasm of skin
Neoplasm of uncertain behavior of skin
Secondary malignant neoplasm of skin
Secondary malignancy of lymph nodes
Benign neoplasm of lymph nodes
Secondary malignant neoplasm
Benign neoplasm of unspecified sites
Actinic keratosis
Degenerative skin conditions and other dermatoses
Dermatitis due to solar radiation
Disorder of skin and subcutaneous tissue NOS
Chronic dermatitis due to solar radiation
D
Melanoma PRS (P <= 5E-6)
Melanomas of skin
Melanomas of skin, dx or hx
Skin cancer
Carcinoma in situ of skin
Other non-epithelial cancer of skin
Basal cell carcinoma
Benign neoplasm of skin
Squamous cell carcinoma
Neoplasm of uncertain behavior of skin
Secondary malignant neoplasm of skin
Benign neoplasm of lymph nodes
Actinic keratosis
Disorder of skin and subcutaneous tissue NOS
E
Melanoma PRS (P <= 5E-5)
Melanomas of skin, dx or hx
Melanomas of skin
Skin cancer
Carcinoma in situ of skin
Other non-epithelial cancer of skin
Basal cell carcinoma
F
Melanoma PRS (P <= 5E-4)
Melanomas of skin
$-\log_{10}(P-value)$
Infectious Diseases
Neoplasms
Endocrine/ metabolic
Hematopoietic
Mental Disorders
Neurological
Sense Organs
Circulatory System
Respiratory
Digestive
Genitourinary
Pregnancy Complications
Dermatologic
Musculoskeletal
Congenital Anomalies
Symptoms
Injuries & Poisonings

PRS trait: Squamous Cell Carcinoma (172.22) > PRS weights: GWAS-Catalog > Study: MGI

Figure 1: Squamous Cell Carcinoma PRS (172.22)

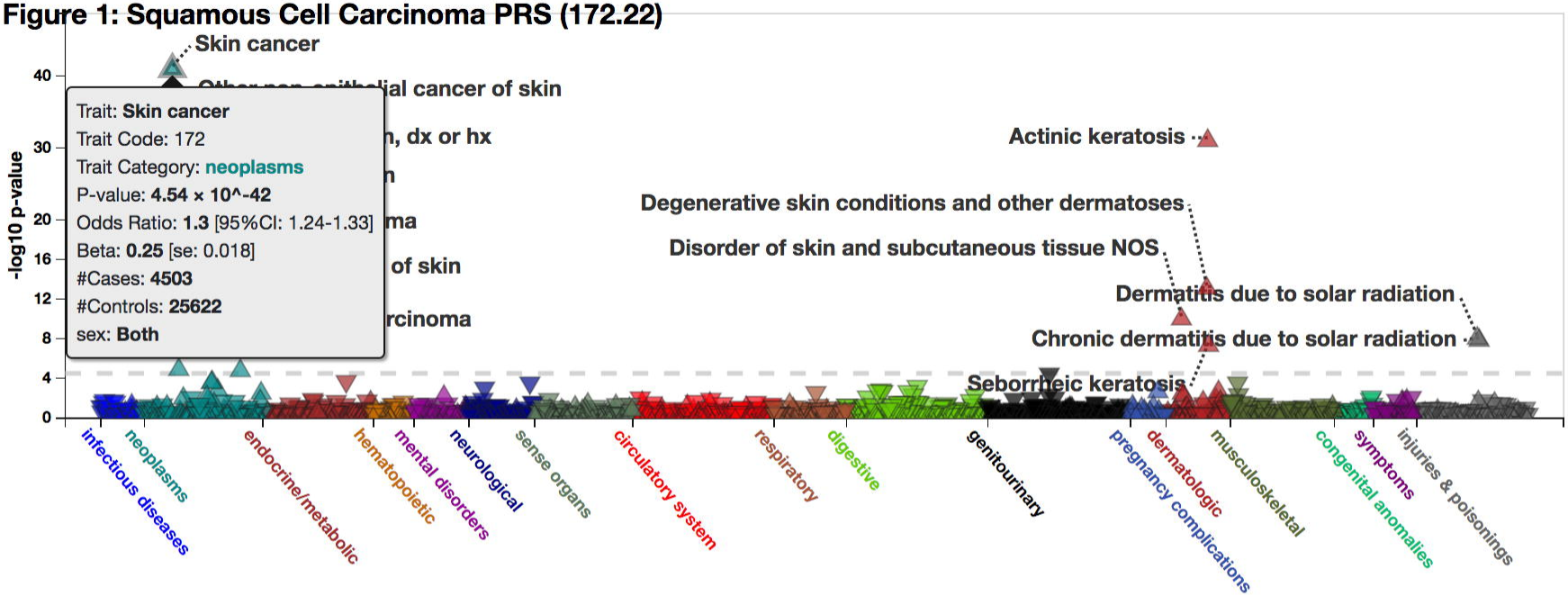


Figure 2: Squamous Cell Carcinoma PRS (172.22) (exclusion)

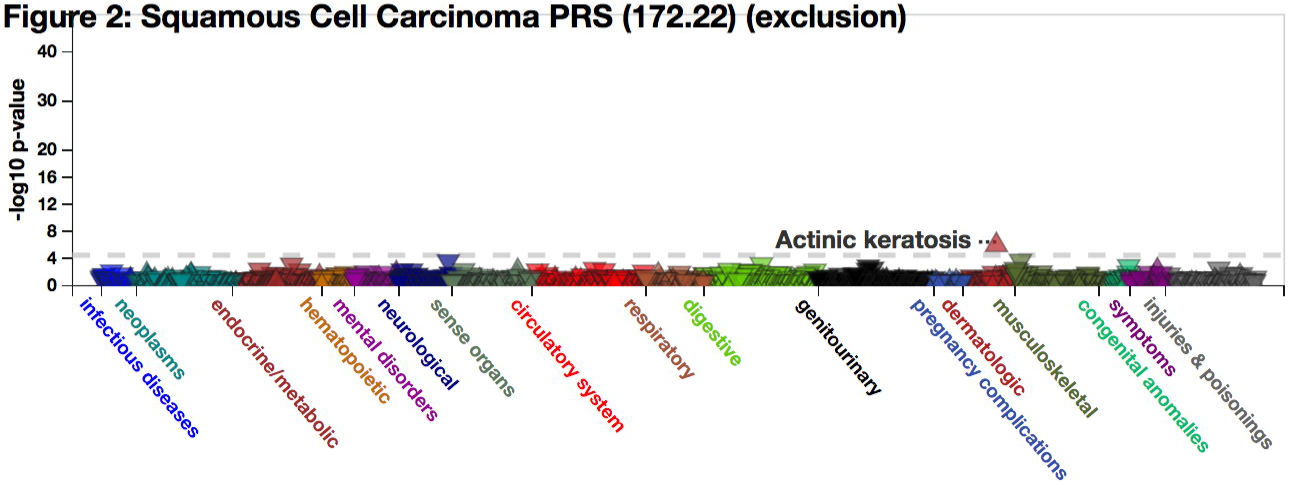
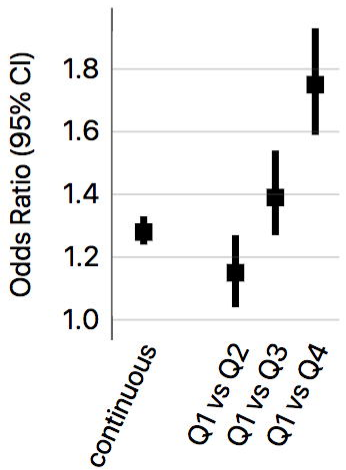


Figure 3: Skin cancer



Download weights for GRCh37 | Download weights for GRCh38