# Assessment of Polygenic Architecture and Risk Prediction based on Common Variants Across Fourteen Cancers

Yan Zhang[1], Amber N. Wilcox[2,3], Haoyu Zhang[1,2], Parichoy Pal Choudhury[2], Douglas F. Easton[4,5], Roger L. Milne[6,7,8], Jacques Simard[9], Per Hall[10,11], Kyriaki Michailidou[5,12], Joe Dennis[5], Marjanka K. Schmidt[13,14], Jenny Chang-Claude[15,16], Puya Gharahkhani[17], David Whiteman[18], Peter T. Campbell[19], Michael Hoffmeister[20], Mark Jenkins[7], Ulrike Peters[21], Li Hsu[21], Stephen B. Gruber[22], Graham Casey[23], Stephanie L. Schmit[24], Tracy A. O'Mara[25], Amanda B. Spurdle[25], Deborah J. Thompson[5], Ian Tomlinson[26,27], Immaculata De Vivo[28,29], Maria Teresa Landi[2], Matthew H. Law[17], Mark M. Iles[30], Florence Demenais[31], Rajiv Kumar[32], Stuart MacGregor[17], D. Timothy. Bishop[33], Sarah V. Ward[34], Melissa L. Bondy[35], Richard Houlston[36], John K. Wiencke[37], Beatrice Melin[38], Jill Barnholtz-Sloan[39], Ben Kinnersley[36], Margaret R. Wrensch[37], Christopher I. Amos[40], Rayjean J. Hung[41], Paul Brennan[42], James McKay[42], Neil E. Caporaso[2], Sonja Berndt[2], Brenda M. Birmann[28], Nicola J. Camp[43], Peter Kraft[44], Nathaniel Rothman[2], Susan L. Slager[45], Andrew Berchuck[46], Paul DP. Pharoah[4,5], Thomas A. Sellers[24], Simon A. Gayther[47], Celeste L. Pearce[48,22], Ellen L. Goode[49], Joellen M. Schildkraut[50], Kirsten B. Moysich[51], Laufey T. Amundadottir[52], Eric J. Jacobs[19], Alison P. Klein[53], Gloria M. Petersen[54], Harvey A. Risch[55], Rachel Z. Stolzenberg-Solomon[2], Brian M. Wolpin[56], Donghui Li[57], Rosalind A. Eeles[58], Christopher A. Haiman[22], Zsofia Kote-Jarai[58], Fredrick R. Schumacher[59], Ali Amin Al Olama[60,61], Mark P. Purdue[2], Ghislaine Scelo[42], Marlene D. Dalgaard[62,63], Mark H. Greene[64], Tom Grotmol[65], Peter A. Kanetsky[24], Katherine A. McGlynn[2], Katherine L. Nathanson[66], Clare Turnbull[36], Fredrik Wiklund[67], BCAC, BEACON, CCFR, CORECT, ECAC, GECCO, GenoMEL, GICC, ILCCO, Integral, InterLymph, OCAC, Oral Cancer GWAS, PANC4, PanScan, PRACTICAL, Renal Cancer GWAS, TECAC, Stephen J. Chanock[2], Nilanjan Chatterjee[1,53*†], Montserrat Garcia-Closas[2†]

[1]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA, [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA, [3]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [4]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK, [5]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [6]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia, [7]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia, [8]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia, [9]Centre Hospitalier Universitaire de Québec–Université Laval Research Center, Québec City, QC, Canada, [10]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [11]Department of Oncology, Södersjukhuset, Stockholm, Sweden, [12]Department of Electron Microscopy/Molecular Pathology and The Cyprus School of Molecular Medicine, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus, [13]Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, [14]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands, [15]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [16]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [17]Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Australia, [18]Cancer Control, QIMR Berghofer Medical Research Institute, Brisbane, Australia, [19]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA, [20]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany, [21]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [22]Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [23]Department of Public Health Sciences, Center

for Public Health Genomics, University of Virginia, Charlottesville, VA, USA, [24]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institution, Tampa, FL, USA, [25]Genetics and Computational Biology Division, QIMR Berghofer Medical Research Institute, Brisbane, Australia, [26]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK, [27]Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK, [28]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, [29]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [30]Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK, [31]Université de Paris, Team of Genetic Epidemiology and Functional Genomics of Multifactorial Diseases, UMR-1124, Institut National de la Santé et de la Recherche Médicale (INSERM), Paris, France, [32]Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [33]Division of Haematology and Immunology, Leeds Institute of Medical Research, University of Leeds, Leeds, UK, [34]Centre for Genetic Origins of Health and Disease, School of Biomedical Sciences, The University of Western Australia, Perth, Australia, [35]Department of Medicine, Section of Epidemiology and Population Sciences, Baylor College of Medicine, Houston, TX, USA, [36]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK, [37]Department of Neurological Surgery, School of Medicine, University of California, San Francisco, San Francisco, CA, USA, [38]Department of Radiation Sciences Oncology, Umeå University, Umeå, Sweden, [39]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA, [40]Institute for Clinical and Translational Research, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA, [41]Lunenfeld-Tanenbuaum Research Institute, Sinai Health System, Toronto, ON, Canada, [42]International Agency for Research on Cancer, World Health Organization, Lyon, France, [43]Division of Hematology and Hematological Malignancies, University of Utah School of Medicine, Salt Lake City, UT, USA, [44]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [45]Department of Health Sciences Research, Division of Biomedical Statistics & Informatics, Mayo Clinic, Rochester, MN, USA, [46]Department of Gynecologic Oncology, Duke University Medical Center, Durham, NC, USA, [47]Center for Bioinformatics and Functional Genomics and the Cedars Sinai Genomics Core, Cedars-Sinai Medical Center, Los Angeles, CA, USA, [48]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA, [49]Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA, [50]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA, [51]Division of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA, [52]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [53]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA, [54]Department of Health Sciences Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA, [55]Chronic Disease Epidemiology, Yale School of Medicine, New Haven, CT, USA, [56]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA, [57]GI Medical Oncology Department, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA, [58]Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey, UK, [59]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA, [60]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Cambridge, UK, [61]Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK, [62]Department of Growth and Reproduction, Copenhagen University Hospital (Rigshospitalet), Copenhagen, Denmark, [63]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark, [64]Clinical Genetics Branch, Division of Cancer Genetics and Epidemiology, National Cancer Institute, Rockville, MD, USA, [65]Cancer Registry of Norway, Oslo, Norway, [66]Department of Medicine, Division of Translational Health and Human Genetics, University of Pennsylvania, Philadelphia, PA, USA, [67]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[†]Authors contributed equally

***Corresponding author:**
Nilanjan Chatterjee (nchatte2@jhu.edu)

**Abstract**

We analyzed summary-level data from genome-wide association studies (GWAS) of European ancestry across fourteen cancer sites to estimate the number of common susceptibility variants (polygenicity) contributing to risk, as well as the distribution of their associated effect sizes. All cancers evaluated showed polygenicity, involving at a minimum thousands of independent susceptibility variants. For some malignancies, particularly chronic lymphoid leukemia (CLL) and testicular cancer, there are a larger proportion of variants with larger effect sizes than those for other cancers. In contrast, most variants for lung and breast cancers have very small associated effect sizes. For different cancer sites, we estimate a wide range of GWAS sample sizes, required to explain 80% of GWAS heritability, varying from 60,000 cases for CLL to over 1,000,000 cases for lung cancer. The maximum relative risk achievable for subjects at the 99th risk percentile of underlying polygenic risk scores, compared to average risk, ranges from 12 for testicular to 2.5 for ovarian cancer. We show that polygenic risk scores have substantial potential for risk stratification for relatively common cancers such as breast, prostate and colon, but limited potential for other cancer sites because of modest heritability and lower disease incidence.

Genome-wide association studies (GWAS) have led to the identification of hundreds of independent cancer susceptibility loci containing common, low-risk variants[1,2]. The number of discoveries varies widely across cancers, largely driven by available sample size, which reflects, in part, disease incidence in the general population. However, specific cancers, e.g., chronic lymphoid leukemia (CLL)[3] and testicular cancer[4], are notable for unexpectedly high numbers of genome-wide significant discoveries from GWAS of relatively small sample size. Previous studies have also reported that these two cancers have high heritability[5]. Across cancer types, polygenic risk scores (PRS) show varying levels of risk stratification depending on the heritability explained by the identified variants and the disease incidence rates in the population[6–12]. Their potential clinical utility would depend not only on the level of risk stratification, but also on other factors such as the availability of appropriate risk-reducing interventions for those identified as at high risk.

Estimation of heritability due to additive effects of all single nucleotide polymorphisms (SNPs) included in GWAS arrays[13], referred to as GWAS heritability in this article, have shown that common variants have substantial potential to identify individuals at different levels of risk for many cancer types[14]. It remains, however, unclear how large the sample sizes of GWAS need to be to reap the full potential of PRS-based risk prediction. Herein, we apply our recently published method[15] to estimate the degree of polygenicity and the effect-size distribution associated with common variants (MAF>0.05) across fourteen different cancer types, based on summary-level association statistics from available GWAS[16–28] from populations of European ancestry (**Supplementary Table 1**). From these inferred parameters, we then provide projections of the expected number of common variants to be discovered and predictive performance of associated PRS as a function of increasing sample size for future GWAS. Finally, by incorporating age-specific incidence [29] from population-based cancer registries, we explore the magnitude of absolute risk stratification potentially achievable by PRS.

We found that cancers are highly polygenic, like other complex traits[15,30,31]. Estimates of the number of susceptibility variants with independent risk associations vary from ~1,000 to 7,500 between the fourteen cancer sites (**Table 1**). For comparability, effect-size distributions are shown in groups of similarly-sized GWAS with similar power for detecting associations (**Figure 1**). For GWAS with <10,000 cancer cases (group 1), CLL and testicular cancer are each associated with 2,000-2,500 variants and characterized by a much larger proportion of variants with larger estimated effect sizes than for the other group 1 cancers, as reflected by wider effect size distribution with heavier tails (**Figure 1, Table 1**). GWAS heritability estimates indicate that, in aggregate, common variants explain a high degree of variation of risk for these two cancers. In contrast, in group 1, esophageal and oropharyngeal cancers are associated with a larger proportion of variants with substantially smaller effect sizes, compared with CLL, testicular and pancreatic cancers in group 1.

For GWAS with 10,000-25,000 cases (group 2), melanoma is noteworthy because it is associated with a wider effect size distribution than other group 2 cancers. The estimated number of susceptibility variants in this group ranges from 1,000 to 2,000. GWAS heritability estimates indicate that aggregated common variants make a relatively small contribution to ovarian and

endometrial cancer susceptibility. Finally, for the three GWAS with >25,000 cases each (group 3), prostate cancer is remarkable for having more variants with large effect sizes, namely, the underlying effect-size distribution has a heavier tail, compared with cancers of the breast and lung (**Figure 1**). In this group, all three cancer types tend to have large numbers of associated variants (>4,500) compared with cancer sites in other groups, but this pattern could partially be due to the very large sample sizes of group 3 GWAS[15].

For a large majority of the fourteen cancer sites, a two-component normal-mixture model for non-null effects provides a substantially better fit to observed summary-statistics than a single-normal distribution; this indicates the presence of a fraction of variants with distinctly larger effect sizes than the remaining (**Supplementary Figures 1-2**). In contrast, a single normal distribution appears to be adequate for esophageal and oropharyngeal cancer, indicating the presence of a large number of variants with a continuum of small effects, similar to our previous findings for traits related to mental health and abilities[15]. Across all fourteen cancers, the predicted number of discoveries and their associated genetic variance explained for current GWAS sample sizes match well to those observed empirically (**Supplementary Table 2**), indicating good fit of our model to the observed data.

GWAS heritability estimates indicate that the potential of PRS for risk discrimination in the population varies widely among cancer types (**Table 1**). The area under the curve (AUC) statistics associated with the best achievable PRS varies from 64% (endometrial and ovarian cancer) to 88% (testicular cancer), and in the range of 70 to 80% for most cancers. The percentage of GWAS heritability explained by known variants varies widely, depending on study sample size and the underlying trait genetic architecture (**Figure 2**). Known variants explain more than a quarter of heritability for cancer sites based on very large sample sizes (e.g., breast and prostate cancer) or for cancer sites that have susceptibility variants with relatively large effect sizes (e.g., CLL, melanoma and testicular cancer). Oropharyngeal cancer, in contrast, has both a small sample size and small effect sizes; its percentage heritability currently explained is almost zero.

The sample size needed to identify common variants that could explain approximately 80% of the total GWAS heritability for the cancers evaluated is generally very large, requiring 200,000 to 1,000,000 cancer cases, with a comparable number of controls (**Figure 2**). However, for three sites, namely, testicular cancer, CLL, and melanoma, the required sample size is smaller, 60,000, 80,000 and 110,000 cases, respectively, due to the large effect sizes of their associated variants. By quadrupling the sample sizes of currently published GWAS, the percentage of GWAS heritability explained would rise to more than 40% across all cancers, except for oropharyngeal cancer. Such sample size increases would also lead to appreciable improvements in PRS discriminatory power across all these sites (**Figures 3-4**). For cancers that were found to be the most polygenic and that had small effect sizes (e.g., cancers of breast, lung and oropharynx), improvement would occur at a slower rates as sample sizes increase, and these sites would require the largest sample sizes to generate PRSs with discriminatory power close to theoretical limits. Of note, for a number of cancers, the achievable relative risks for subjects at the 99$^{th}$ percentile of PRS distribution compared with those at average risk, are comparable to those for

monogenic disorders[32] (e.g., relative-risk more than 3-4 fold) (**Figure 4).** Across all fourteen cancer types, inclusion of SNPs using more liberal but optimized p-value thresholds (**see Methods**) would improve performance of PRS-based risk prediction versus using the stringent genome-wide significance level, but the anticipated gains would be generally modest (**Supplementary Figures 3-4).**

Projections of residual lifetime cancer risks for the US non-Hispanic white population show that the discriminatory power of PRS built from current or foreseeable studies will depend heavily on the underlying cancer incidence in the population (**Figure 5, Supplementary Figures 5-7)**. The potential clinical utility of PRS depends on the degree of risk stratification and specific prevention or early detection strategies for a given cancer, should they exist. For common cancers, such as breast, colorectal and prostate, a PRS with even modest discriminatory power (maximum AUC of approximately 70%, **Figure 3**) can provide substantial stratification of absolute risk in the population. In contrast, for CLL and testicular cancer, even though its PRS could achieve a higher AUC (e.g. in the range 80-90%, **Figure 3**), the degree of absolute risk stratification will be modest because of the infrequency of these cancers. Thus, a PRS by itself has the least impact on risk stratification for cancer sites that are infrequent or/and that have low heritability. However, it is possible that PRS could have clinical utility for some of these cancers in the presence or in combination with other risk factors and biomarkers. For example, a PRS for lung cancer may provide larger stratification for absolute risk among smokers than never smokers because of the higher baseline risk in smokers.

Our study is subject to several limitations. We may have underestimated the number of underlying common susceptibility loci, especially for those cancers for which current GWAS have small sample sizes[15]. Thus, the interpretation of comparisons of the underlying genetic architecture across cancer types with very different sample sizes requires caution. Nevertheless, the major patterns are unlikely to be due to differences in sample size. For example, we estimated oropharyngeal and esophageal cancers to be two of the most polygenic sites, though the GWAS sample sizes for these two sites were relatively small. Further, Q-Q plots of observed and expected p-values indicate that the inferred models for effect-size distributions explain observed GWAS summary-statistics well, regardless of GWAS sample size. Another important limitation is that we only included data from subjects of European ancestry, since GWAS data for other ancestries are currently too small to permit reliable projections for most cancer sites. In addition, several cancers (e.g., lung, ovary, glioma, and breast) consist of etiologically heterogeneous subtypes that were not considered in our analyses due to lack of adequate sample sizes for appropriate subtypes for most of these cancer sites. Further studies of ancestry- and subtype-specific genetic architectures are needed to address these limitations.

In our projections, we assume standard agnostic association analysis of SNPs without incorporating any external information on population genetics or functional characteristics of SNPs. It is, however, possible to incorporate various types of external information to improve power for discovery of associations[33–36] and genetic risk prediction[37] . We have evaluated the merit of future GWAS only in terms of their ability to explain heritability and improve risk prediction. However, current and future discoveries have other major implications, including

provident insights to biological pathways and mechanisms, potential gene-environment interactions and understanding causal relationships through Mendelian Randomization analyses[38]. A number of these cancers are known to have rare high-penetrant risk variants, but for this study we have focused on estimating effect-size distribution associated with common variants. Furthermore, heritability analysis indicate that uncommon and rare variants could explain a substantial fraction of the variation of complex traits[39] , and thus, it is likely that there are many unknown uncommon and rare variants associated with these cancers as well. In the future, characterization of heritability and effect-size distribution associated with the full spectrum of allele frequencies will require individual level sequencing data on a substantially larger number of cases and controls.

The observed differences in the underlying genetic architecture of susceptibility across cancers could be due to various factors, including the effect of negative selection[30,40], tissue-specific genetic regulation of gene-expression[41], cell of origin[42], the number of biological steps needed to transition from normal to malignant tissue[43], mediation of genetic effects by underlying environmental exposures[44], and the presence of heterogeneous cancer-specific subtypes[21,25,27,28]. A number of cancer types, including those of lung, oropharynx and esophagus, which were associated with large numbers of SNPs with small average effect sizes, have known strong environmental risk factors and distinct etiologic subtypes. It is also noteworthy that testicular cancer also stands out for a large number of discoveries in cross-tissue expression quantitative trait loci analyses, likely indicating a stronger association of SNPs on gene expression levels for this tissue compared to others[41].

In conclusion, our comprehensive analysis of fourteen cancer sites in adults of European ancestry reveals that while all sites have polygenic influences, there is substantial diversity observed in their underlying genetic architectures, which reflects important biology and also influences the utility of polygenic risk prediction for individual cancers. Our projections for future yields of GWAS across these cancers provide a roadmap for important returns from future investment in research, including the potential clinical utility of polygenic risk prediction for stratification of absolute risks in the population.

## Methods

**Description of GWAS studies.** We analyzed summary data from GWAS studies across fourteen cancer types. For select cancer sites[26,28], we downloaded publicly available genome-wide summary-level statistics from the latest consortium-based analyses. For others, we obtained access to data through collaborative efforts with individual consortia. Details about individual studies, including the number of cases and controls, are provided in **Supplementary Table 1**.

**Quality control for summary GWAS data.** Across all cancers, we applied several filtering steps analogous to those used earlier for estimation of heritability[45,46] and effect-size distribution using summary-level data[15]. First, we restricted analysis to SNPs within a set of reference ~1.07 million SNPs included in the HapMap3 and which had minor allele frequency > 0.05 in the 1000

Genome European Ancestry sample. Second, we excluded SNPs having substantial amounts of missing genotype data: sample sizes less than 0.67 times the 90$^{th}$ percentile of the distribution of sample sizes across all SNPs. Third, we excluded SNPs within the major histocompatibility complex (MHC) region (i.e., SNPs between 26,000,000 and 34,000,000 base pairs on chromosome six) which is known to have very complex allelic architecture and can have uncharacteristically large effects on some traits. Fourth, we removed regions that have SNPs with extremely large effect sizes to reduce possible undue influence of them on estimation of parameters associated with overall effect-size distributions. We identify all top SNPs which have associated chi-square statistics greater than 80 (i.e., OR (in standardized scale) >2.19) and removed all SNPs which were within 1MB distance of or had an estimated larger than 0.1 with the top SNPs. We added back the contribution of these top independent SNPs in the final reporting of the total number of susceptibility SNPs, estimates of total heritability, and various projections we made as a function of sample size of the GWAS.

**Statistical model.** We inferred common variant genetic architecture of the different cancers using GENESIS[15], a method we recently developed to characterize underlying effect-size distributions in terms of the total number of susceptibility SNPs (polygenicity) and a normal mixture model for the distribution of their effects. Specifically, it is assumed that standardized effects of common SNPs in an underlying logistic regression model on the risk of a cancer can be specified in the mixture distribution in the form $\beta_m \sim (1-\pi_c)\delta_0 + \pi_c N(0,\sigma^2)$ (two-component model), or $\beta_m \sim (1-\pi_c)\delta_0 + \pi_c\left[p_1 N\left(0,\ \sigma_1^2\right) + p_2 N(0,\sigma_2^2)\right]$ (three-component model) where $\delta_0$ is the Dirac delta function indicating that a fraction, $1-\pi_c$, of the SNPs have null effects, and remaining $\pi_c$ fraction of SNPs have non-null effects. Under the three-component model, $p_2 = 1-p_1$ denotes the proportion of SNPs allocated to mixture component with larger variance component (assuming $\sigma_2^2 > \sigma_1^2$) models. Under these models, $M\pi_c$ characterizes the degree of polygenicity, i.e., the number of susceptibility SNPs with independent effects on disease risk. Under both models, we defined "GWAS heritability" of a disease as $h^2 = M\pi_c E(\beta^2)$, where $E(\beta^2)$ denotes the average variance size of the non-null SNPs. We observed that under the above model, $h^2$ is also the population variance of the underlying "true" polygenic risk score, defined as $PRS = \sum_{m=1}^{M}\beta_m G_m$, where $G_m$ denotes the standardized genotype associated with the m-th SNP. Under the two-component model, which assumes a single normal distribution for the effect of all susceptibility SNPs, $E(\beta^2) = \sigma^2$. Under the three-component model, which allows mixture of two-normal distributions with distinct variance components and thus can better accommodate the presence of a group of susceptibility SNPs with much larger effects than others, we have $p_1\sigma_1^2 + p_2\sigma_2^2$. Under the three-component model, we use the fraction $v = p_1\sigma_1^2/(p_1\sigma_1^2 + p_2\sigma_2^2)$ to characterize the proportion of heritability explained by SNPs associated with the larger variance component parameter. As we removed SNPs with extremely large effects ($\chi_i^2 > 80$) and the associated regions from the analysis, in reporting the final heritability estimates, we added back the contribution of the top SNPs from these excluded regions as $\sum_i(\hat{\beta}_i^2 - \tau_i^2)$ where $\hat{\beta}$ is the estimate of log-odds ratio (in standardized scale) and $\tau_i$ is the corresponding standard error for the $i$-th SNP.

**Genetic variance projection.** Given the estimated effect-size distribution, we calculated expected discoveries and genetic variance explained using $ED = M\hat{\pi}_c \int_\beta pow_{\alpha,n}(\beta) \sum_{h=1}^{H} \hat{p}_h N(0, \hat{\sigma}_h^2)\, d\beta$ and $EV = M\hat{\pi}_c \int_\beta \beta^2 pow_{\alpha,n}(\beta) \sum_{h=1}^{H} \hat{p}_h N(0, \hat{\sigma}_h^2)\, d\beta$, respectively, at $\alpha = 5 \times 10^{-8}$ for a GWAS of sample size $n$, where $pow_{\alpha,n}(\beta) = 1 - \Phi\left(c_{\frac{\alpha}{2}} - \sqrt{n}\beta\right) + \Phi(-c_{\frac{\alpha}{2}} - \sqrt{n}\beta)$ with $\Phi(\cdot)$ the standard normal cumulative density function and $c_\alpha = \Phi^{-1}(1-\alpha)$ the $\alpha$-th quantile for the standard normal distribution. Similar to heritability calculations, we added back the contributions of top SNPs with very large effects to the number of expected discoveries and associated variances explained by the quantities $\sum_i pow_{\alpha,n}(\hat{\beta}_i)$ and $h^{-2}\sum_i(\hat{\beta}_i^2 - \tau_i^2) pow_{\alpha,n}(\hat{\beta}_i)$. We observed that for projections involving sample sizes bigger than the current study $pow_{\alpha,n}(\hat{\beta}_i)$ for the large effect SNPs will all be very close to 1.0.

**Projection for AUC and relative risk at top 1%.** As we quantify heritability in terms of the variability of the underlying "true" polygenic risk-score, we used the formula[12,47,48], $\mathrm{AUC} = \Phi\left(\sqrt{\frac{h^2}{2}}\right)$ to characterize the best discriminatory power achievable in limiting using common variant PRS. We used the same formula to calculate the AUC associated with PRSs that could be built using SNPs either reaching genome-wide significance (p-value < $5\times 10^{-8}$) or a weaker but optimized threshold, for a GWAS of given sample size based on the projected variance of the respective PRS. Given sample size of GWAS and an effect-size distribution for the underlying cancer, an optimal threshold for SNP selection that will maximize the expected predictive performance of PRS is calculated using analytic formula we have derive earlier[48]. The relative risk for those estimated to be at the 99$^{th}$ percentile or higher of the distribution of a PRS (compared to the average risk of the population) was calculated using the formula[12] $exp\left(-\frac{h^2}{2} + \Phi^{-1}(0.99)\sqrt{h^2}\right)$ where $h^2$ is the population variance of the PRS.

**Absolute risk projection.** For each cancer site, we projected the distribution of residual lifetime risk (up to age 80 years) for Non-Hispanic White individuals in the general US population according to PRSs which could be built from GWAS of different sample sizes. For any given age, we first obtain the distribution of residual lifetime risks based on a model for absolute risks developed using the iCARE tool that we have described earlier[12,29]. The iCARE tool uses projected standard deviations of PRS at different GWAS sample sizes and age-specific cancer incidence rates available from the US National Cancer Institute-Surveillance, Epidemiology, and End Results Program (NCI-SEER) (2015) to obtain absolute risk distributions. In deriving absolute risks, we adjusted for competing risk of mortality due to other causes using the age-specific mortality rates from the Center for Disease Control (CDC) WONDER database (2016). We then weighted the projected residual lifetime risk distribution at different baseline ages (in five-year categories) based on the US population distribution of ages within 30 to 75 years, as observed in the estimated 2016 US Census. For cancers of the reproductive system, weights were based on the age distributions among males or females, as appropriate.

## Acknowledgement

The research was supported by a RO1 grant from NHGRI (1 R01 HG010480-01) and the intramural program of the National Cancer Institute.

## Author Contribution

N.C. and M.G.C. conceived the project. Y.Z. and A.W. performed main analyses. Y.Z., N.C. and M.G.C. wrote the first draft of the manuscript. L.T.A., C.I.A., S.B., B.M.B., N.J.C., S.J.C., F.D., M.H.G., L.H., M.M.I., E.J.J., M.J., M.H.L, S.M., K.A.M., R.L.M., T.A.O., P.P.C, U.P., P.D.P.P., M.P.P., H.A.R., M.K.S., S.L.S., T.A.S., A.B.S., R.Z.S.S., D.J.T., S.V.W. and M.R.W. commented on earlier drafts of manuscript. All authors reviewed and approved drafts of manuscripts.

## References

1. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* **17**, 692-704 (2017).
2. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* (2019).
3. Law, P. J. et al. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat Commun* **8**, 14175 (2017).
4. Litchfield, K. et al. Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor. *Nat Genet* **49**, 1133-1140 (2017).
5. Mucci, L. A. et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315**, 68-76 (2016).
6. Maas, P. et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol* **2**, 1295-1302 (2016).
7. Mavaddat, N. et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21-34 (2019).
8. Jeon, J. et al. Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology* **154**, 2152-2164.e19 (2018).
9. Seibert, T. M. et al. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**, j5757 (2018).
10. Garcia-Closas, M. et al. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res* **73**, 2211-2220 (2013).
11. Turnbull, C., Sud, A. & Houlston, R. S. Cancer genetics, precision prevention and a call to action. *Nat Genet* **50**, 1212-1218 (2018).
12. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* **17**, 392-406 (2016).
13. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569 (2010).
14. Sampson, J. N. et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *J Natl Cancer Inst* **107**, djv279

(2015).
15. Zhang, Y., Qi, G., Park, J. H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet* **50**, 1318-1326 (2018).
16. Berndt, S. I. et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat Commun* **7**, 10933 (2016).
17. Wang, Z. et al. Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor. *Nat Genet* **49**, 1141-1147 (2017).
18. Lesseur, C. et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* **48**, 1544-1550 (2016).
19. Klein, A. P. et al. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nat Commun* **9**, 556 (2018).
20. Scelo, G. et al. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat Commun* **8**, 15724 (2017).
21. Melin, B. S. et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet* **49**, 789-794 (2017).
22. Law, M. H. et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet* **47**, 987-995 (2015).
23. O'Mara, T. A. et al. Identification of nine new susceptibility loci for endometrial cancer. *Nat Commun* **9**, 3166 (2018).
24. Schumacher, F. R. et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* **6**, 7138 (2015).
25. Phelan, C. M. et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **49**, 680-691 (2017).
26. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928-936 (2018).
27. McKay, J. D. et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **49**, 1126-1132 (2017).
28. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94 (2017).
29. Choudhury, P. P. et al. iCARE: An R Package to Build, Validate and Apply Absolute Risk Models. *bioRxiv* p.079954 (2018).
30. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* **50**, 746-753 (2018).
31. Stahl, E. A. et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* **44**, 483-489 (2012).
32. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).
33. Schork, A. J. et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet* **9**, (2013).
34. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* **94**, 559-573 (2014).
35. Andreassen, O. A. et al. Improved detection of common variants associated with

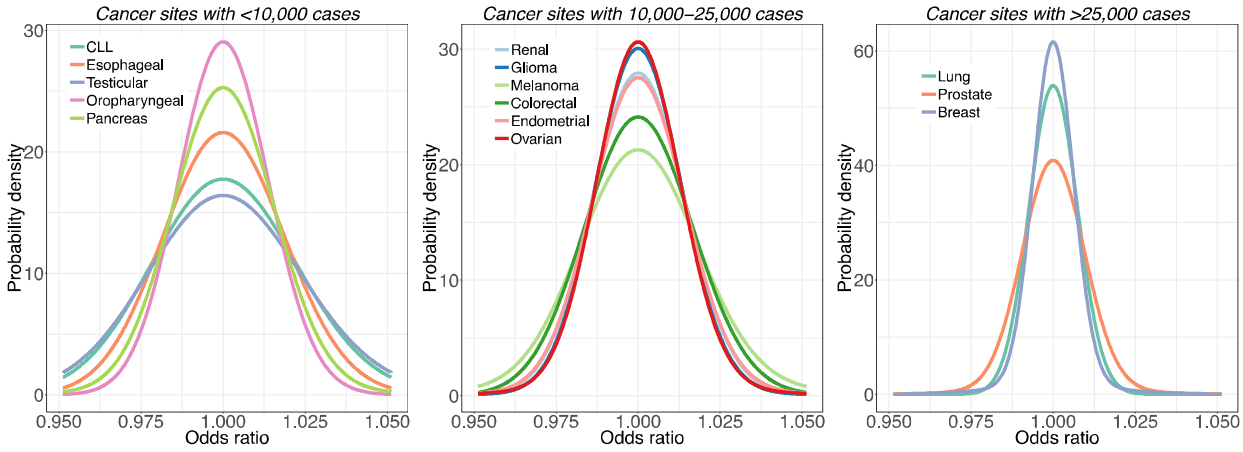
schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet* **92**, 197-209 (2013).
36. Andreassen, O. A. et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet* **9**, e1003455 (2013).
37. Hu, Y. et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput Biol* **13**, e1005589 (2017).
38. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89-98 (2014).
39. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv* (2019).
40. O'Connor, L. J. et al. Polygenicity of complex traits is explained by negative selection. *bioRxiv* **420497**, (2018).
41. GTEx, C. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
42. Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314-322 (2011).
43. Rizzo, A. A., Strickland, D. & Bouchard, S. The challenge of using virtual reality in telerehabilitation. *Telemed J E Health* **10**, 184-195 (2004).
44. Hutter, C. M. et al. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genet Epidemiol* **37**, 643-657 (2013).
45. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).
46. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
47. Pharoah, P. D. et al. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33-36 (2002).
48. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* **45**, 400-5, 405e1 (2013).

**Table 1: Estimated number of independent common susceptibility variants and heritability across 14 cancer sites using the best fitted (two- or three-component) normal mixture model for effect-size distributions.** All results are reported with respect to a reference panel of 1.07 million common SNPs included in the Hapmap3 panel after removal of MHC region.

| Number of cases in the analysis | Cancer site | Total Number of susceptibility SNPs[a] (SE[b]) | Total heritability[c] (SE) | Average heritability explained per susceptibility SNP[d] (SE), in $10^{-4}$ | Number of SNPs associated with larger variance component (SE) | % of heritability explained by SNPs with larger variance component | AUC associated with the best PRS[e] (SE) |
|---|---|---|---|---|---|---|---|
| <10,000 | CLL[f] | 2025 (1501) | 1.62 (0.37) | 7.2 (4.4) | 52 (15) | 41 | 0.82 (0.03) |
| | Esophageal | 3641 (2515) | 1.24 (0.36) | 3.4 (1.9) | NA[g] | NA | 0.78 (0.03) |
| | Testicular | 2598 (2088) | 2.81 (0.40) | 9.2 (6.6) | 196 (75) | 54 | 0.88 (0.02) |
| | Oropharyngeal | 3623 (2060) | 0.68 (0.27) | 1.9 (0.5) | NA | NA | 0.72 (0.04) |
| | Pancreas | 1757 (1490) | 0.60 (0.16) | 3.2 (2.2) | 47 (27) | 31 | 0.71 (0.03) |
| 10,000 - 25,000 | Renal | 2220 (1555) | 0.57 (0.12) | 2.4 (1.4) | 46 (36) | 24 | 0.70 (0.02) |
| | Glioma | 2364 (1593) | 0.87 (0.11) | 2.2 (1.2) | 61 (25) | 55 | 0.75 (0.01) |
| | Melanoma | 1098 (533) | 0.65 (0.09) | 4.4 (1.6) | 106 (58) | 52 | 0.72 (0.01) |
| | Colorectal | 1484 (696) | 0.43 (0.10) | 2.9 (0.8) | 14 (11) | 7 | 0.68 (0.02) |
| | Endometrial | 1052 (772) | 0.27 (0.07) | 2.5 (1.3) | 46 (34) | 26 | 0.64 (0.02) |
| | Ovarian | 1015 (715) | 0.24 (0.06) | 2.2 (1.1) | 49 (31) | 36 | 0.64 (0.02) |
| >25,000 | Lung | 6096 (2750) | 0.39 (0.06) | 0.6 (0.2) | 15 (7) | 15 | 0.67 (0.01) |
| | Prostate | 4530 (1052) | 0.77 (0.04) | 1.1 (0.2) | 276 (99) | 51 | 0.73 (0.01) |
| | Breast | 7599 (1615) | 0.60 (0.03) | 0.6 (0.1) | 587 (133) | 56 | 0.71 (0.00) |

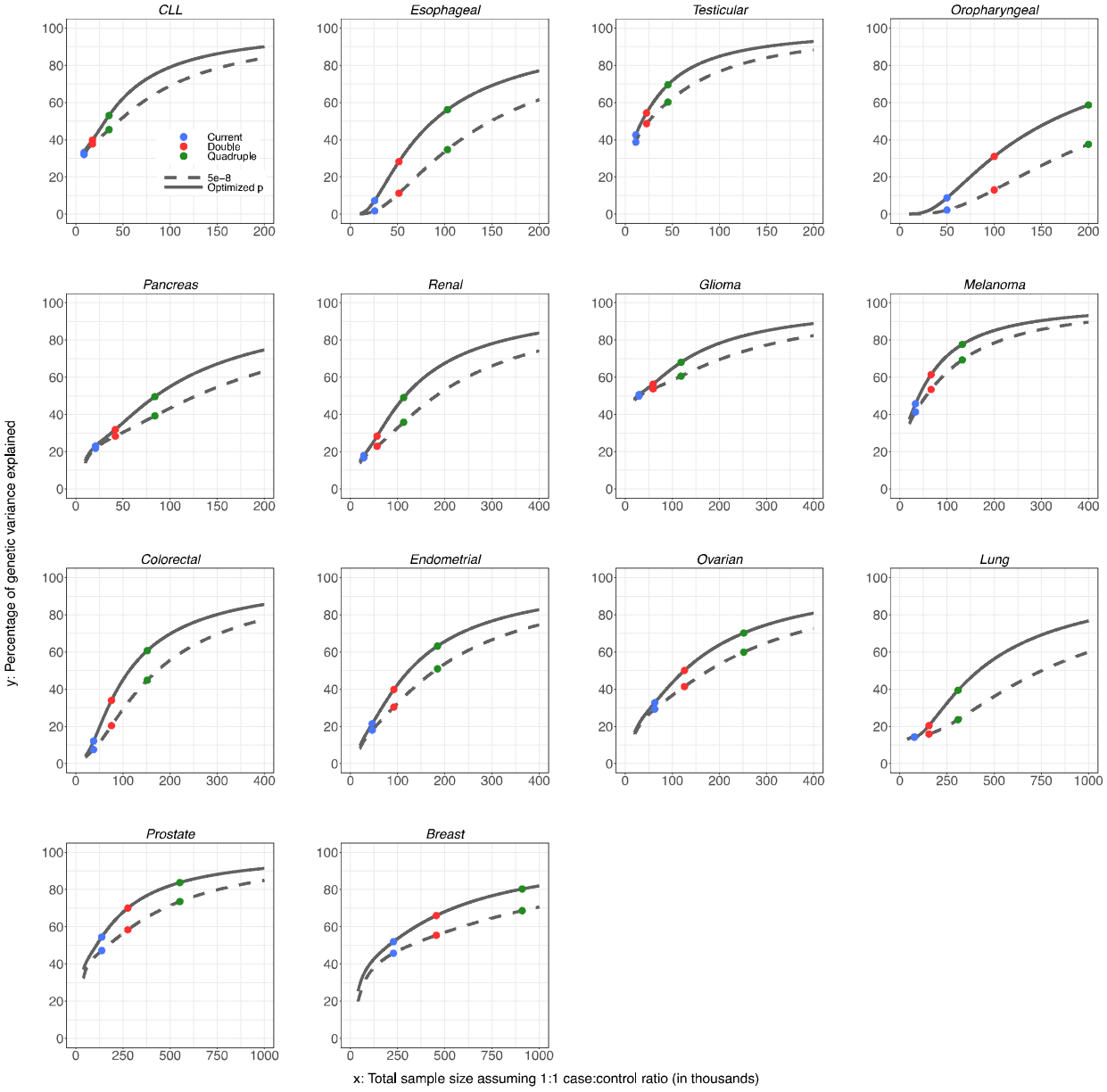
[a]SNP: single nucleotide polymorphism. [b]Standandard errors. [c]Total heritability is characterized by population variance of the underlying true PRS as $h^2 = Var(\sum_{m=1}^{M} \beta_m G_m) = M\pi_c E(\beta^2)$, where $E(\beta^2)$ denotes per-SNP effect-size of the non-null SNPs. [d]Average heritability explained per susceptibility SNP excludes SNPs with extremely large effects (see **Methods**). [e]Area under the curve (AUC) associated with best PRS is calculated using the formula AUC=$\Phi(\sqrt{h^2/2})$ where $\Phi(\cdot)$ is the cumulative density function of standard normal distribution. [f]CLL = chronic lymphocytic leukemia. [g]NA indicates a two-component model is favorable compared to three-component model.

**Fig. 1. Estimated effect-size distributions for susceptibility SNPs across 14 cancer sites**. Effect-size distribution of susceptibility SNPs is modelled using a two-component normal mixture model for all sites, except esophagus, oropharynx and melanoma. For these sites, effect-sizes are modelled using a single normal distribution that provided similar fit as the two-component normal mixture model (see **Supplementary Fig. 1 & 2**). SNPs with extremely large effects are excluded from the analysis (see **Methods**). Plots are stratified by sample size of the GWAS for comparability. Distributions with fatter tails imply the underlying traits have relatively greater number of susceptibility SNPs with larger effects. Note here the effect-size distribution is plotted on the log scale of odds ratio (x-axis).



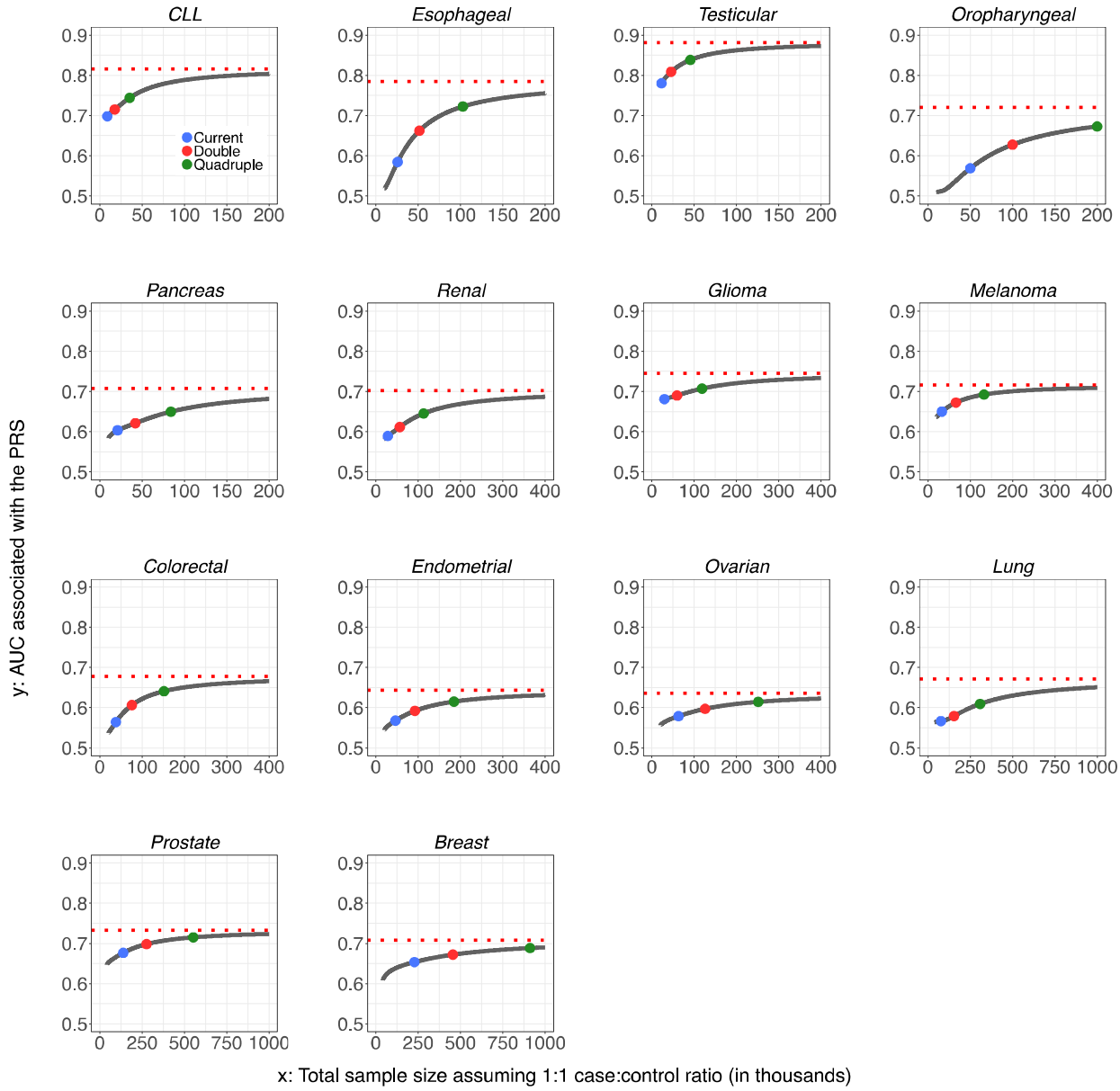

CLL = chronic lymphocytic leukemia.

**Fig. 2. Projections of percentage of GWAS heritability explained by SNPs as sample size for GWAS increases**. Results are shown for projections including SNPs at the optimized p-value threshold (solid curve) and at genome-wide significance ($P<5 \times 10^{-8}$) level (dashed curve). Colored dots correspond to sample size for largest published GWAS and those for doubled and quadruped sizes.



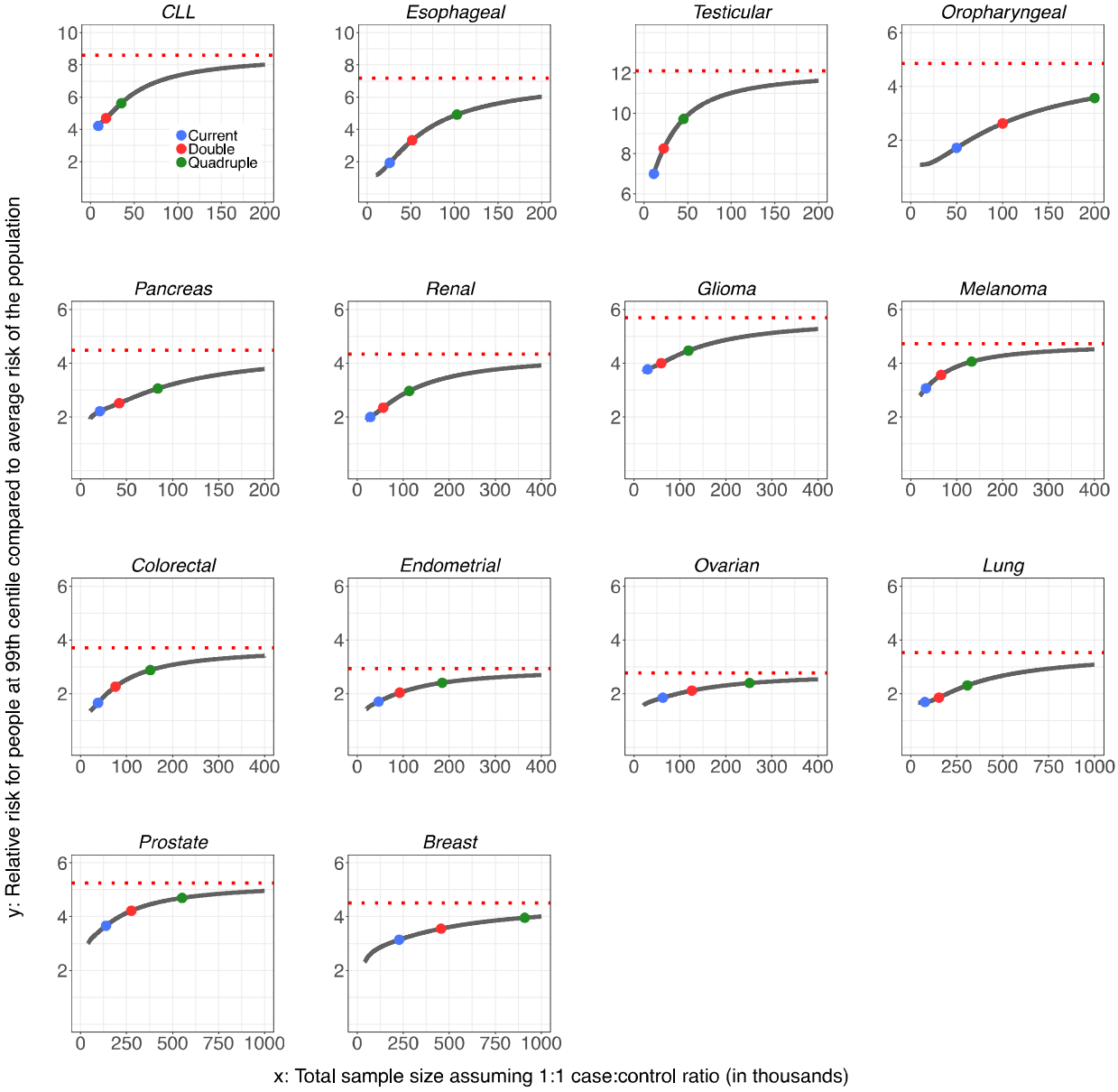

For oropharyngeal cancer, the projections at the "current sample size" are based on a sample size of 25K cases and 25K controls. For breast and esophageal cancer, the projections at the "current sample size" are based on the current largest GWAS sample sizes: 123K cases and 106K controls, and 10K cases and 17K controls, respectively. For all other cancer sites, the projections at the "current sample size" are based on the GWAS sample sizes in **Supplementary Table 1.** CLL = chronic lymphocytic leukemia.

**Fig. 3. Projections of area under the curve (AUC) characterizing predictive performance of PRS as sample size for GWAS increases.** Results are shown for PRS including SNPs at the optimized p-value threshold. The dotted horizontal red line indicates the maximum AUC achievable according to the estimate of GWAS heritability. Colored dots correspond to sample size for largest published GWAS and those for doubled and quadruped sizes.



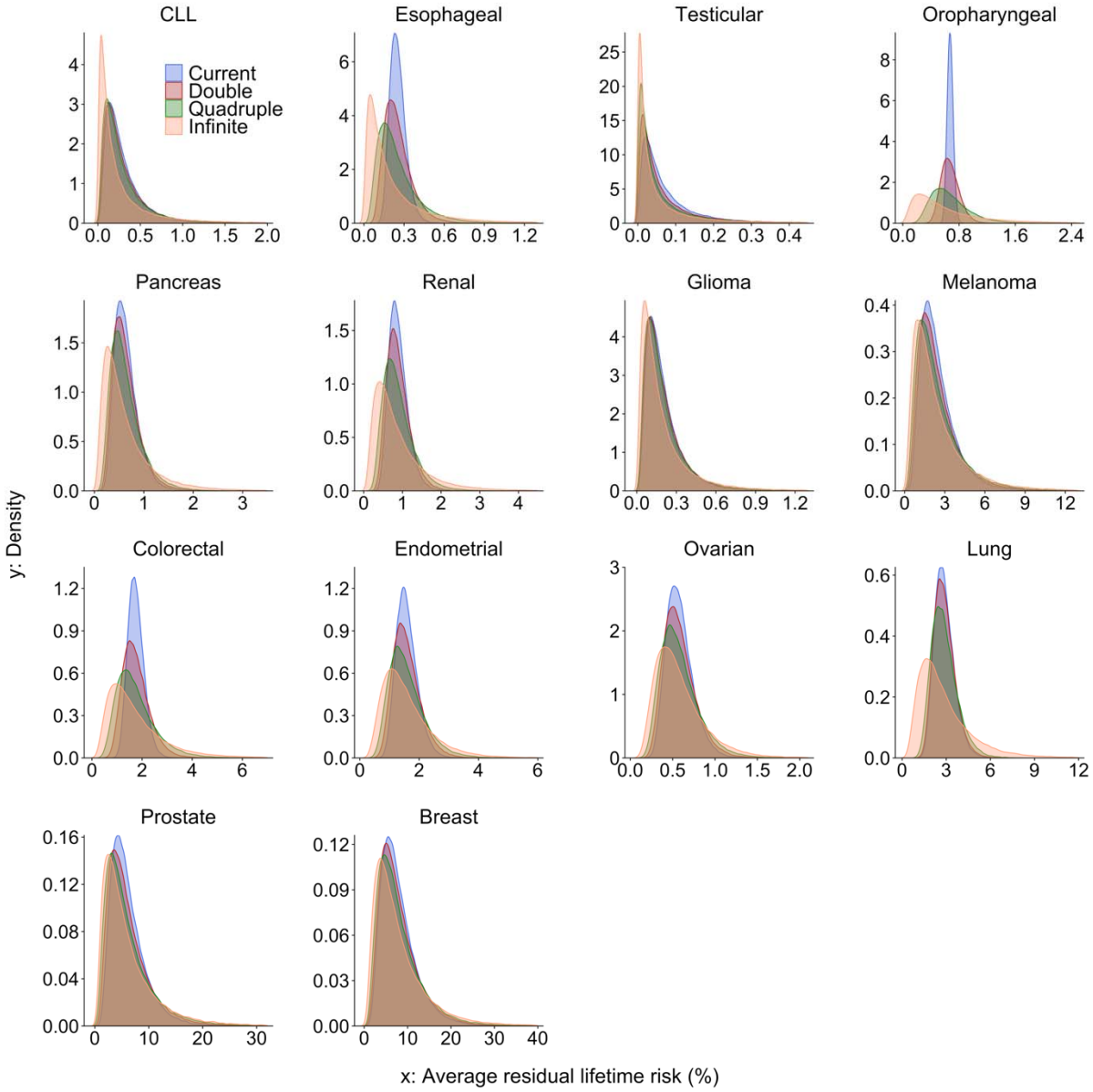

For oropharyngeal cancer, the projections at the "current sample size" are based on a sample size of 25K cases and 25K controls. For breast and esophageal cancer, the projections at the "current sample size" are based on the current largest GWAS sample sizes: 123K cases and 106K controls, and 10K cases and 17K controls, respectively. For all other cancer sites, the projections at the "current sample size" are based on the GWAS sample sizes in **Supplementary Table 1.** CLL = chronic lymphocytic leukemia.

**Fig. 4. Projections of relative risks for individuals at or higher than 99$^{th}$ percentile of PRS distribution (compared to average risk) as sample size for GWAS increases**. Results are shown where PRS is built based on SNPs at optimized p-value threshold. The dotted horizontal red line indicates the maximum relative risk achievable according to estimate of GWAS heritability. Colored dots correspond to sample size for largest published GWAS and those for doubled and quadruped sizes.




For oropharyngeal cancer, the projections at the "current sample size" are based on a sample size of 25K cases and 25K controls. For breast and esophageal cancer, the projections at the "current sample size" are based on the current largest GWAS sample sizes: 123K cases and 106K controls, and 10K cases and 17K controls, respectively. For all other cancer sites, the projections at the "current sample size" are based on the GWAS sample sizes in **Supplementary Table 1.** CLL = chronic lymphocytic leukemia.

**Fig. 5. Projected distribution of average residual lifetime risk in the US population of Non-Hispanic Whites aged 30 to 75 years, according to variation of polygenic risk scores.** The projections are shown for PRS built based on GWAS with current, doubled and quadrupled sample sizes and the best PRS that corresponds to limits defined by heritability. The projections are obtained by combining information on projected population variance of PRS, age-specific population incidence rate, competing risk of mortality and current distribution of age according to US 2016 census.




For oropharyngeal cancer, the projections at the "current sample size" are based on a sample size of 25K cases and 25K controls. For breast and esophageal cancer, the projections at the "current sample size" are based on the current largest GWAS sample sizes: 123K cases and 106K controls, and 10K cases and 17K controls, respectively. For all other cancer sites, the projections at the "current sample size" are based on the GWAS sample sizes in **Supplementary Table 1.** CLL = chronic lymphocytic leukemia.