Most common ‘sporadic’ cancers have a significant germline genetic component

Yi Lu1,†, Weronica E. Ek1,†, David Whiteman2, Thomas L. Vaughan6, Amanda B. Spurdle3, Douglas F. Easton7,8, Paul D. Pharoah7,8, Deborah J. Thompson7, Alison M. Dunning8, Nicholas K. Hayward4, Georgia Chenevix-Trench5, Q-MEGA and AMFS Investigators†, ANECS-SEARCH†, UKOPS-SEARCH†, BEACON Consortium† and Stuart Macgregor1,*

1Statistical Genetics, 2Cancer Control Group, 3Molecular Cancer Epidemiology, 4Oncogenomics and 5Cancer Genetics, QIMR Berghofer Medical Research Institute, Herston, QLD, Australia, 6Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, 7Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, and 8Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK

Received March 16, 2014; Revised May 22, 2014; Accepted June 16, 2014

Common cancers have been demarcated into ‘hereditary’ or ‘sporadic’ (‘non-hereditary’) types historically. Such distinctions initially arose from work identifying rare, highly penetrant germline mutations causing ‘hereditary’ cancer. While rare mutations are important in particular families, most cases in the general population are ‘sporadic’. Twin studies have suggested that many ‘sporadic’ cancers show little or no heritability. To quantify the role of germline mutations in cancer susceptibility, we applied a method for estimating the importance of common genetic variants (array heritability, \( h^2_g \)) to twelve cancer types. The following cancers showed a significant (\( P < 0.05 \)) array heritability: melanoma USA set \( h^2_g = 0.19 \) (95% CI = 0.01–0.37) and Australian set \( h^2_g = 0.30 \) (0.10–0.50); pancreatic \( h^2_g = 0.18 \) (0.06–0.30); prostate \( h^2_g = 0.81 \) (0.32–1); kidney \( h^2_g = 0.18 \) (0.04–0.32); ovarian \( h^2_g = 0.30 \) (0.18–0.42); esophageal adenocarcinoma \( h^2_g = 0.24 \) (0.14–0.34); esophageal squamous cell carcinoma \( h^2_g = 0.19 \) (0.07–0.31); endometrial UK set \( h^2_g = 0.23 \) (0.01–0.45) and Australian set \( h^2_g = 0.39 \) (0.02–0.76). Three cancers showed a positive but non-significant effect: breast \( h^2_g = 0.13 \) (0–0.56); gastric \( h^2_g = 0.11 \) (0–0.27); lung \( h^2_g = 0.10 \) (0–0.24). One cancer showed a small effect: bladder \( h^2_g = 0.01 \) (0–0.11). Among these cancers, previous twin studies were only able to show heritability for prostate and breast cancer, but we can now make much stronger statements for several common cancers which emphasize the important role of genetic variants in cancer susceptibility. We have demonstrated that several ‘sporadic’ cancers have a significant inherited component. Larger genome-wide association studies in these cancers will continue to find more loci, which explain part of the remaining polygenic component.

INTRODUCTION

Common cancers are frequently demarcated into ‘hereditary’ (‘familial’) or ‘sporadic’ (‘non-hereditary’) types (1). Such distinctions initially arose from work identifying rare highly penetrant germline mutations causing ‘hereditary’ cancer (such as CDKN2A mutations in melanoma and BRCA1/2 mutations in breast cancer). These rare mutations are important in particular families but do not explain many cases in the general population. The more common ‘sporadic’ cancers were thought to have little or no germline genetic component, as the large twin studies failed to detect significant heritability for many cancers, with the exceptions of breast, colorectal and prostate cancers (2,3). Studies
examinining affected relative pairs have shown a familial component to some cancers (4,5), but unlike twin studies, these cannot reliably distinguish between the effects of genes and the effects of common environment on heritability estimates. Recently, genome-wide association studies (GWAS) have identified some novel cancer susceptibility loci (~1–120, depending on the cancer type), but these typically explain only a small proportion of the heritability. Some studies have considered the contribution of all genetic variants simultaneously instead of just few known loci, using approaches such as polygenic risk prediction (6), but only showed a small heritability for prostate cancer (7,8) and not for breast cancer (8). Thus, it remains unclear whether common germline genetic variants, on the whole, make significant contribution to ‘sporadic’ cancers.

In the current study, we apply a recently developed method to several cancer GWAS datasets. This method links case–control status to relatedness derived from dense genetic marker data (9), allowing the estimation of the magnitude of the germline genetic component. Since most of the genotyping arrays used to date only include single-nucleotide polymorphisms (SNPs), which represent the common variation (frequency > 1–5%), the resultant estimate of ‘array heritability’ represents a lower bound for the overall heritability. Importantly, unlike studies that determine the contribution of genes based on estimates of risk among family members (e.g. sibling relative risk), our method excludes closely related individuals, yielding estimates less likely to be biased by shared environment.

RESULTS

For twelve cancer types, we were able to obtain genome-wide array data for >1000 cases and controls (Table 1). Eight of the twelve cancers studied demonstrated a significant polygenic component underlying cancer risk (Table 2). As with previous twin studies, we found a substantial genetic component for the most common cancer, prostate cancer, although our estimate ($h_g^2 = 0.81$, 95% CI = 0.32–1), albeit with wide confidence interval, was markedly higher than those reported in twin studies (0.42 in Lichtenstein et al. (2) and this estimate was updated to 0.58 using the Nordic Twin Registry of Cancer (NorTwinCan) (10); Table 2). For most other cancers, previous twin studies had insufficient sample size to draw any conclusions (2), but that was improved in the analysis of NorTwinCan (10).

Using GWAS data, we showed that endometrial cancer, esophageal adenocarcinoma, esophageal squamous cell carcinoma (ESCC), melanoma, pancreatic, renal cell carcinoma (RCC) and ovarian cancer all had significant array heritability (range from ~0.10 to ~0.40). Our data showed positive but non-significant estimates for breast, gastric and lung cancers; thus, we could not make a clear statement on these three cancers. Despite there being published genome-wide significant loci for bladder cancer, these only account for <1% of the variance, and overall this cancer had only a small polygenic component (estimate 0.01, with upper 95% confidence limit of 0.11). Significant X chromosome effects were seen in esophageal adenocarcinoma, ESCC and gastric cancer (GC), under all three dosage compensation models (Supplementary Material, Table S1).

When the effects of genome-wide significant SNPs were removed, the variance explained for prostate cancer and melanoma decreased substantially, although only in the smaller of the two melanoma datasets was the remaining polygenic variance not statistically significant (Table 2). In most other cases, the genome-wide significant SNPs accounted for very little of the estimated polygenic variance.

Following the removal of genetic relatedness and the adjustment of the first 20 principal components, we found the remaining population structure only contributed to a negligible proportion of genetic variance in each dataset. In most cases, the parameters indicating cryptic relatedness ($b_0$ and population stratification ($b_1$) were not significantly different from zero (Supplementary Material, Table S2). Hence, we conclude that it is unlikely that our estimates of $h_g^2$ were spuriously inflated by uncontrolled for population structure.

Furthermore, there was no significant array heritability in the first two control–control contrast studies after applying the standard analysis protocol (Supplementary Material, Table S3). We found, however, false array heritability in the comparison of two subsets of Welcome Trust Case-Control Consortium (WTCCC) controls (Supplementary Material, Table S3). This was removed by applying stringent quality control (QC) in addition to the standard protocol, which underlies the necessity of applying stringent QC to the analyses using WTCCC as controls. Overall, these results demonstrated that our method is robust.

DISCUSSION

The goal of the current study was to quantify the genetic contribution to the lifetime risk of cancer. In eight of the twelve cancers studied, there was a significant polygenic component, that is, a large number of genes of weak effect are involved. Given a polygenic basis to disease, it has been shown that most people who develop cancer would be ‘sporadic cases’ (no affected first-, second- or third-degree relatives) (11). This may go some way toward explaining the misconception that most ‘sporadic’ cancers have little or no germline genetic component. The polygenic components estimated for these datasets were not largely attributable to the genome-wide significant variants identified by recent GWASs. Even in conditions such as prostate cancer, where the polygenic component decreased substantially when genome-wide significant loci were removed, a significant polygenic component remained. One corollary of this is that conducting larger GWASs in most cancers will continue to find ever larger numbers of loci, which will explain more of the polygenic component. It has been suggested that an enhanced predictive power, gained from incorporating a larger number of genetic loci, will help fulfill the promise of personalized risk prediction (12).

In common with methods for estimating heritability from twin studies, we estimated the variance attributable to common SNPs (array heritability, $h_g^2$) on an assumed underlying quantitative liability scale where individuals are modeled as becoming affected once they exceed a particular threshold. However, there are two ways in which estimates of heritability derived from twin studies differ from array heritability. First, when using twin/family data, expected sharing of the genome among close relatives is used, meaning that all of the genome contributes to estimates of heritability. In contrast, estimating the genetic relationship between unrelated individuals uses information on only the portion of the genome tagged by SNPs on the microarray.
(i.e. common variants). Rare genetic variants (such as rare BRCA1/2 mutations in breast cancer) are generally not tagged on commercial microarrays, and hence, they do not contribute to our array heritability estimates. Second, twin studies sample from the general population and hence provide heritability estimates for the cancer up to the age at which the twins were ascertained (13) (twin studies typically ascertain individuals who are 85 years of age as assumed for our \( h^2_g \) calculations), which means that some twin study-derived estimates of heritability are artificially lowered for late onset cancers such as prostate. For example, when adding 10 years of follow-up (plus expanding the original twin cohorts from the Nordic countries), the heritability of prostate cancer was found significantly higher than the previous estimate (approximately one-third increase), which may be due to different age ascertainment (Table 2).

Results from our study should be interpreted in the context of limitations. Despite having ensured of > 1000 cases and controls for all 12 cancer types, the statistical power for detecting a significant polygenic component varies dramatically with disease prevalence and true heritability. All but bladder, breast, gastric and lung cancer sets had reasonable statistical power (≥ 0.7); hence, the estimates of array heritability we obtained here should be relatively robust. However, for those four cancers, we had limited power (< 0.5) to draw clear conclusion on their genetic variance.

It is worth noting that, since the information for the estimation of array heritability derives from the differences between cases and controls, when cases form a more extreme sample from the population [i.e. cancers with lower lifetime risk (K)], the information content of a sample is increased. Mathematically, for \( K < 0.5 \), the variance of the array heritability reduces as \( K \) reduces [is proportional to \( (K (1 - K))^2 \)] (9). Hence, for the two most common cancer types, prostate and breast cancer, a sample size of 1000 cases and controls are less informative than for rarer cancer types; thus, the estimates had wide confidence intervals. The prostate cancer dataset is from the Prostate, Lung, Colon and Ovarian (PLCO) Cancer Screening Trial. It contains a larger proportion of aggressive disease than observed in the population (aggressive disease was defined as cases with a Gleason Sore of \( \geq 7 \) or Stage of \( \geq III \) at the time of diagnosis; 60% in the current sample set as compared with ~20–30% in prostate cancer patients from the general population). To our

### Table 1. Sample size (post-QC), number of SNPs (post-QC) and genotyping arrays for each cancer

<table>
<thead>
<tr>
<th>Cancer types</th>
<th>Cases</th>
<th>Controls</th>
<th>SNPs</th>
<th>Arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder</td>
<td>2848</td>
<td>3159</td>
<td>505729</td>
<td>HumanHapmap550</td>
</tr>
<tr>
<td>Breast</td>
<td>1081</td>
<td>1085</td>
<td>489247</td>
<td>HumanHapmap550</td>
</tr>
<tr>
<td>Australian endometrial cancer</td>
<td>564</td>
<td>574</td>
<td>540995</td>
<td>Human610-Quad, Human1M-Duo</td>
</tr>
<tr>
<td>UK endometrial cancer</td>
<td>670</td>
<td>2558</td>
<td>495905</td>
<td>Human610-Quad, Human1M-Duo</td>
</tr>
<tr>
<td>Esophageal adenocarcinoma</td>
<td>1397</td>
<td>1947</td>
<td>795160</td>
<td>Omni1-Quad</td>
</tr>
<tr>
<td>ESCC</td>
<td>1388</td>
<td>1614</td>
<td>442551</td>
<td>Human660W-Quad</td>
</tr>
<tr>
<td>GC</td>
<td>1342</td>
<td>1624</td>
<td>451186</td>
<td>Human660W-Quad</td>
</tr>
<tr>
<td>Kidney</td>
<td>1080</td>
<td>2489</td>
<td>498235</td>
<td>HumanHapmap550, Human610-Quad, Human660W-Quad</td>
</tr>
<tr>
<td>Lung</td>
<td>2557</td>
<td>2895</td>
<td>493060</td>
<td>HumanHapmap550, HumanHapmap240, HumanHapmap300, Human610-Quad, Human1M-Duo</td>
</tr>
<tr>
<td>Melanoma QLD</td>
<td>2009</td>
<td>1963</td>
<td>298477</td>
<td>Omni1-Quad</td>
</tr>
<tr>
<td>Melanoma USA</td>
<td>1849</td>
<td>965</td>
<td>809287</td>
<td>Omni1-Quad</td>
</tr>
<tr>
<td>Ovary</td>
<td>1710</td>
<td>2562</td>
<td>471552</td>
<td>Human610-Quad, Human1M-Duo</td>
</tr>
<tr>
<td>Pancreas</td>
<td>1926</td>
<td>1971</td>
<td>499792</td>
<td>HumanHapmap550, HumanHapmap610</td>
</tr>
<tr>
<td>Prostate</td>
<td>1093</td>
<td>992</td>
<td>493308</td>
<td>HumanHapmap300, HumanHapmap240</td>
</tr>
</tbody>
</table>

### Table 2. Array heritability explained by all autosomes (\( h^2_g \)) compared with the total heritability estimates from the largest twin studies

<table>
<thead>
<tr>
<th>Cancer types</th>
<th>Total heritability from twins studies</th>
<th>Mucci (2013) %</th>
<th>( h^2_g ) (95% CI) %</th>
<th>( h^2_g ) (95% CI), after removing known loci %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder</td>
<td>Lichtenstein (2000)</td>
<td>n.a.</td>
<td>1 (0–11)</td>
<td>0 (0–10)</td>
</tr>
<tr>
<td>Breast</td>
<td>27 (4–41)</td>
<td>28 (12–52)</td>
<td>13 (0–56)</td>
<td>5 (0–46)</td>
</tr>
<tr>
<td>Australian endometrial cancer</td>
<td>0 (0–42)</td>
<td>24 (14–87)</td>
<td>39 (2–76)</td>
<td>39 (2–76)</td>
</tr>
<tr>
<td>UK endometrial cancer</td>
<td>n.a.</td>
<td>n.a.</td>
<td>23 (1–45)</td>
<td>23 (1–45)</td>
</tr>
<tr>
<td>Esophageal adenocarcinoma</td>
<td>n.a.</td>
<td>n.a.</td>
<td>24 (14–34)</td>
<td>24 (14–34)</td>
</tr>
<tr>
<td>GC</td>
<td>n.a.</td>
<td>n.a.</td>
<td>10 (0–24)</td>
<td>8 (0–22)</td>
</tr>
<tr>
<td>Kidney</td>
<td>26 (0–49)</td>
<td>25 (12–44)</td>
<td>30 (10–50)</td>
<td>21 (4–41)</td>
</tr>
<tr>
<td>Lung</td>
<td>19 (1–37)</td>
<td>19 (1–37)</td>
<td>8 (0–28)</td>
<td>8 (0–28)</td>
</tr>
<tr>
<td>Melanoma QLD</td>
<td>2009</td>
<td>28 (15–47)</td>
<td>30 (18–42)</td>
<td>29 (17–41)</td>
</tr>
<tr>
<td>Melanoma USA</td>
<td>36 (0–53)</td>
<td>18 (6–30)</td>
<td>16 (4–28)</td>
<td>16 (4–28)</td>
</tr>
<tr>
<td>Pancreas</td>
<td>42 (29–50)</td>
<td>58 (52–63)</td>
<td>81 (32–100)</td>
<td>59 (12–100)</td>
</tr>
</tbody>
</table>

Ones in bold are significantly different from zero (\( P < 0.05 \)).
knowledge, there is no existing heritability estimate for prostate cancer stratified by aggressive and nonaggressive disease. We estimated the array heritability separately for these two diseases and found that the aggressive disease ($h^2_a = 0.80, 95\%\ CI = 0.15-1, P = 0.009, based on 626 aggressive disease and 992 controls) had a higher heritability than the nonaggressive disease ($h^2_n = 0.55, 95\%\ CI = 0-1, P = 0.1, based on 467 nonaggressive cases and same controls). Due to the limited sample size, we cannot distinguish whether the heritability of aggressive and nonaggressive disease were significantly different. It warrants further investigation in larger prostate cancer datasets. If the heritability of aggressive disease is confirmed to be higher than that of nonaggressive disease, our dataset with a large proportion of aggressive disease might partly explain why our estimate is higher than those in the literature. A further consideration with prostate cancer is that there are very few familial genes identified. This is in contrast with, for example, breast and ovarian cancers, where the effect of high- to moderate-penetrance familial relative risk ($\gamma$). Since the contribution of rare high-penetrance genes to prostate cancer appears to be small, our estimate of the array heritability explained by common variants might be high and very close to the total heritability.

The heritability of breast cancer from the twin studies was consistently $\sim0.3$ (2,10). The estimate was higher in family studies (0.4–0.6) (15), but this may be due to shared environmental factors. Our estimate of array heritability accounts for approximately half of the estimates from twin studies, which is not surprising given familial genes such as BRCA1/2 with high penetrance account for a sizable proportion of the total heritability.

For these common cancers, ideally the Genome-wide Complex Trait Analysis (GCTA) approach is applied to large datasets from the same or very similar populations, through consortium effort such as Collaborative Oncological Gene-environment Study (COGS) (16). At present, the genotyping array (iCOGS) does not have sufficient coverage as it contains mainly the fine-mapping loci without tagging the whole genome. If applying GCTA to these data, the estimate of array heritability would be much smaller due to insufficient array coverage. However, it might be feasible when the second OncoArray with reasonable tagging of the whole genome becomes available.

Our estimate of array heritability for bladder cancer was 0–0.11. Despite there are known risk variants, these only explain a very small proportion of heritability. The current sample set was underpowered to detect a low heritability: with the sample size of 2848 cases and 3159 controls, we had 80% power to detect the true heritability of 0.1. Although we did not exclusively show a significant heritability, the result did suggest that the heritability of bladder cancer is likely to be much lower than other cancer types.

We have used publically available data for a number of cancer types, including bladder, breast, esophagus squamous cell carcinoma, gastric, kidney, lung, melanoma (USA), pancreas and prostate. Some of these datasets consisted of multiple small cohorts, which were genotyped on different arrays. Using different arrays for cases and controls, respectively, may inflate estimates of array heritability. However, our robustness checks indicated that such inflation was unlikely to be large.

For some cancers, ongoing research is providing a molecular taxonomy for cancer, which is currently in a state of evolution, making it unclear how to best categorize cancer cases so they form a homogeneous group. In breast cancer and ovarian cancer, there is evidence for different subtypes having distinct genetic bases (17,18). Ideally for the polygenic analysis, subtypes would be tested separately but we lacked sufficient sample size to, for example, differentiate between different ovarian cancer histologies. Similarly, there are multiple ways of dealing with the smoking and non-smoking groups in lung cancer, but further investigation is limited by the sample size. There are differences in prevalence by sex in some cancers. We modeled this by fitting sex as a covariate. In addition to modeling an overall difference in prevalence between sexes, array heritability can be estimated in each sex separately, although for most of the cancers presented here, we had insufficient data to make a clear statement on sex-specific polygenic effects. We kept Australian and United States melanoma samples as separate groupings because the lifetime risk, a key part of the array heritability estimate, varies markedly between countries. The UK and Australian endometrial cancer sets were also analyzed separately because different QC was applied.

In summary, we show that common genetic variants contribute significantly to susceptibility to most of the cancer types that we examined. For most cancers examined here, the descriptor ‘sporadic’ or ‘non-hereditary’ should be replaced by ‘polygenic’.

**MATERIALS AND METHODS**

Individual study descriptions are listed as follows.

**Melanoma (Australian)**

Data consist of a subset of samples from an Australian melanoma study (19). Melanoma cases in individuals of European descent ($n = 2168$) were selected from the Queensland study of Melanoma: Environment and Genetic Associations (Q-MEGA) (20) and the Australian Melanoma Family Study (21). Samples from Australian individuals of European descent from three different sources were used as controls ($n = 2146$) (20–22).

**Melanoma (United States)**

Data consist upon an extensive resource of melanoma cases and hospital based controls collected over several years at the U.T. M.D. Anderson Cancer Centre. This dbGaP study (dbGaP accession number phs000187.v1.p1) contains samples from 1965 European ancestry cases and 1038 European ancestry controls using the Illumina OMNI-Quad SNP chip (23,24).

**Breast cancer**

Data consist of a subset from the Nurses’ Health Study (25,26). Controls were matched to cases based on age, blood collection variables (time, date and year of blood collection, as well as recent (<3 months) use of postmenopausal hormones), ethnicity (all cases and controls are self-reported Caucasians) and menopausal status (all cases and controls were menopausal at blood draw). The dbGaP accession number is phs000147.v1.p1.
Prostate cancer

Data consist of a subset from the Cancer Genetic Markers of Susceptibility prostate cancer GWAS (27) with patients of European ancestry drawn from the PCLO cancer screening trial. The dbGaP accession number is phs000207.v1.p1. The aggressive subtype was defined as cases with a Gleason Score of ≥7 or Stage of >III at the time of diagnosis, whereas the nonaggressive subtype was defined as cases with a Gleason Score of <7 and Stage of <III.

Pancreatic cancer

Participants were drawn from 12 cohort studies and 8 case–control studies (28,29). Cases were defined as those individuals having primary adenocarcinoma of the exocrine pancreas. Those with non-exocrine pancreatic tumors were excluded from the study. The datasets used for the analyses described in this manuscript were obtained from dbGaP through dbGaP accession number phs000206.v3.p2.

Lung cancer

Data consist of a subset from the national cancer institute (NCI) GWAS study, specifically the consent group ‘Cancer in all age groups, other diseases in adults only, and methods’, containing a total of 7622 subjects (30). Subjects were drawn from one population-based case–control study and three cohort studies. Lung cancer cases were included based on a lung cancer diagnosis established on clinical criteria and confirmed by pathology reports from surgery, biopsy or cytology samples in ~95% of cases and on clinical history and imaging for the remaining 5%. The datasets used for the analyses described in this manuscript were obtained from dbGaP through dbGaP accession number phs000336.v1.p1.

Kidney cancer (renal cell carcinoma)

Data consist of a subset of the NCI GWAS of RCC (31). The GWAS study included 1453 RCC cases and 3531 controls of European background from 4 studies (3 prospective cohorts and 1 case–control study). All cases were diagnosed with pathologically confirmed renal cancer (International Classification of Diseases for Oncology, Second Edition, Topography C64). The datasets used for the analyses described in this manuscript were obtained from dbGaP through dbGaP accession number phs000351.v1.p1.

Bladder cancer

Data consist of a subset of the GWAS study (32). Cases and controls used in this study were all from European ancestry (3532 cases and 5119 controls) (32). Cases were defined as individuals having histologically confirmed primary carcinoma of the urinary bladder. Scan data were obtained from two case–control studies carried out in Spain and the United States and three prospective cohort studies in Finland and the United States. Samples were genotyped on Illumina SNP array (Human_1M, HumanHap550, HumaHap250, HumanHap300 or Human610-Quad). In this study, we only used individuals typed on the 550 chip. The datasets used for the analyses described in this manuscript were obtained from dbGaP through dbGaP accession number phs000346.v1.p1.

Ovarian cancer

All cases were from the UK and were confirmed as invasive epithelial ovarian cancer (33). Genotyping for the cases was conducted using the Illumina Infinium 610 K array at Illumina Corporation. Existing data from the WTCCC, genotyped using the Illumina Human1M-Duo array, were used as controls.

Esophageal adenocarcinoma

The data used were collected by investigators in the Barrett’s and Esophageal Adenocarcinoma Consortium (BEACON). A Subset of individuals with European ancestry from 17 epidemiologic studies from three cohorts (Australia, Europe and North America) was selected for this study (1509 cases and 3202 controls). Cases were clinically diagnosed and histologically confirmed with esophageal adenocarcinoma.

Gastric adenocarcinoma and esophageal squamous cell carcinoma

Data used for these two cancers were collected from the Asian UGI GWAS (34,35). Esophageal squamous cell carcinomas include incident, primary esophageal cancers with squamous histology. Gastric cancers include incident, primary gastric adenocarcinomas. Controls were all cancer-free at the time of enrolment (and blood draw). The datasets used for the analyses described in this manuscript were obtained from dbGaP through dbGaP accession number phs000361.v1.p1.

Endometrial cancer

GWAS was conducted using cases with endometrioid histology endometrial cancer from Australia and the UK (36). Genotyping was conducted using the Human 610 K array on the Illumina Infinium platform. We extracted control data for SNPs included on the 610-K platform from existing Illumina 1.2 M genome-wide scan data for controls of European ancestry from two UK population-based studies genotyped by the WTCCC (37). We also matched Australian cases on the basis of ancestry with 1244 individuals from the Hunter Community Study (38).

Methods

We applied a standardized QC pipeline to all datasets: individuals with >5% missing genotypes were excluded, as were SNPs with minor allele frequency of <0.01 or call rate of <0.99 or P-value from testing Hardy–Weinberg equilibrium of <0.0001. Specific to two datasets, UK endometrial cancer and ovarian cancer, both using the WTCCC as controls, we have applied more stringent QC regarding differential missingness between cases and controls (P < 0.001) and two-locus QC (P < 0.02) as described previously (39). The numbers of SNPs passed QC were summarized in Table 1. We used GCTA (9,40) to estimate the genetic relatedness between pairs of individuals that were derived from all autosomal
SNPs post-QC. By relating observed case/control status to the estimated genetic relationship matrix, the variance explained by all autosomal SNPs was estimated using restricted maximum likelihood in GCTA (9). We define array heritability ($h^2_a$) as variance explained by autosomal SNPs divided by the total variance. Intuitively, $h^2_a$ is high if pairs of individuals with high genetic relatedness (estimated from SNPs) have more similar phenotypes than those with lower genetic relatedness.

All datasets were restricted to individuals with European ancestry, except for the GC and ESCC, which were restricted to individuals with Asian ancestry. To limit the impact of population structure (cryptic relationship and population stratification) on the estimation of array heritability, we included only those strictly unrelated samples (genetic relatedness < 0.025, approximately equivalent to second cousin; the numbers of cases and controls after removing genetic relatedness were summarized in Table 1) and adjusted for the first 20 principal components. We also applied the simple approach based on genome partitioning (41) to quantify remaining population substructure. This approach involves modeling the differences in by-chromosome $h^2_g$, which are estimated from separately fitting GRM of one chromosome at a time and from jointly fitting the GRMs of all 22 autosomes as random effects, against the chromosome length, i.e. $h^2_{c,sep} - h^2_{c,joint} = b_0 + b_1 \times L + e$, with $b_0$ indicating inflation due to cryptic relationship (irrespective of chromosomal length $L$), and $b_1$ indicating inflation due to population stratification (correlates with $L$). The genetic variance attributable to remaining population structure can thus be calculated from $h^2_g = b_0 \times 22/21 + b_1 \times \text{sum}(L)/21$ (41).

To model differing prevalence between sexes, sex was fitted as a fixed effect if appropriate. We assume causal variants have a similar frequency spectrum as genotyped SNPs (42). Variance explained by X chromosome SNPs, defined as $h^2_g_x$, was estimated separately, under a full dosage compensation model (that is, each allele in females has only a half of the effect in males, $h^2_g_x$ for females = $1/2 \times h^2_g_x$ for males) (41). We also estimated $h^2_g_x$ under alternative models: no dosage compensation (that is, each allele has a similar effect on the trait; thus, the genetic variance in females is twice of that in males, $h^2_g_x$ for females = $2 \times h^2_g_x$ for males) and equal variance between males and females (that is, $h^2_g_x$ for females = $h^2_g_x$ for males) (41).

We initially estimated variance explained on the observed scale (cancer case or control). For such binary traits, heritability is usually parameterized on an unobserved continuous `liability' scale (typically, a Normal distribution, where individuals above a threshold are `affected'). This is desirable as it allows the heritability to describe the relative importance of genetic and non-genetic factors independent of disease prevalence. Hence, to allow us to compute a lower bound for heritability (on the continuous scale), we transform the estimate obtained on the observed scale using the estimated lifetime risk from population studies as the `prevalence' (Supplementary Material, Table S4).

To assess how much array heritability has already been explained by the specific SNPs unambiguously identified by GWAS, for each cancer, we analyzed the data with the known GWAS loci removed. To do this, for each SNP in the GWAS catalog (http://www.genome.gov/gwastudies/ accessed 21 August 2013), we removed the associated SNPs and all SNPs within 1 megabase. Since linkage disequilibrium very rarely extends beyond this distance, removing all SNPs in such regions will result in all signals at the locus being removed.

The methods for estimating the ‘polygenic’ component are potentially sensitive to QC issues; hence, we tested the robustness of this method by estimating array heritability in control sets from different studies: first, we contrasted controls from the breast and prostate cancer studies; second, we combined controls from the breast and the prostate cancer studies and compared this combined set with controls from the pancreatic cancer study; third, we split the WTCCC controls into two sets, 1958 British Birth cohort and the UK Blood Service control group. The analysis protocol was identical to the one used in case–control studies. We arbitrarily defined one set of controls as ‘cases’ and tested whether the array heritability was significantly different from zero. Since ‘lifetime risk’ in the situation cannot be defined, we did the calculation on the observed case/control scale; note the significance of the genetic component is not affected by scaling for ‘lifetime risk’.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**ACKNOWLEDGEMENTS**

We thank Jian Yang and Sang Hong Lee for helpful discussions. Additional acknowledgements for each study are listed in Supplementary Material.

**Conflict of Interest statement.** None declared.

**FUNDING**

S.M., N.K.H., A.B.S. and G.C.T. are supported by Australian National Health and Research Council fellowships. S.M. and D.W. are supported by an Australian Research Council fellowship.

**REFERENCES**


