The landscape of the heritable cancer genome

Viola Fanfani[1]. Luca Citi[2], Adrian L. Harris[3], Francesco Pezzella[4], and Giovanni Stracquadanio[1]

[1]Institute of Quantitative Biology, Biochemistry, and Biotechnology, SynthSys, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

[2]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom

[3]Molecular Oncology Laboratories, Department of Oncology, The Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom

[4]Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom

Running title
The landscape of the heritable cancer genome.

Corresponding Author: Giovanni Stracquadanio, School of Biological Sciences, University of Edinburgh, Max Born Crescent, Edinburgh, EH9 3BF Edinburgh, United Kingdom. Phone: +44 (0) 131 6507193, Email: giovanni.stracquadanio@ed.ac.uk.

Declaration of interests
The authors declare no competing interests.

Abstract

Genome-wide association studies (GWAS) have found hundreds of single nucleotide polymorphisms (SNP) associated with increased risk of cancer. However, the amount of heritable risk explained by SNPs is limited, leaving most of cancer heritability unexplained. Tumor sequencing projects have shown that causal mutations are enriched in genic regions. We hypothesized that SNPs located in protein coding genes and nearby regulatory regions could explain a significant proportion of the heritable risk of cancer. To perform gene-level heritability analysis, we developed a new method, called Bayesian Gene HERitability Analysis (BAGHERA), to estimate the heritability explained by all genotyped SNPs and by those located in genic regions using GWAS summary statistics. BAGHERA was specifically designed for low heritability traits such as cancer and provides robust heritability estimates under different genetic architectures. BAGHERA-based analysis of 38 cancers reported in the UK Biobank showed that SNPs explain at least 10% of the heritable risk for 14 of them, including late onset malignancies. We then identified 1,146 genes, called cancer heritability genes (CHG), explaining a significant proportion of cancer heritability. CHGs were involved in hallmark processes controlling the transformation from normal to cancerous cells. Importantly, 60 of them also harbored somatic driver mutations, and 27 being tumor suppressors. Our results suggest that germline and somatic mutation information could be exploited to identify subgroups of individuals at higher risk of cancer in the broader population and could prove useful to establish strategies for early detection and cancer surveillance.

Significance

This study describes a new statistical method to identify genes associated with cancer heritability in the broader population, creating a map of the heritable cancer genome with gene-level resolution.

Introduction

Decades of research have shown that inherited genomic mutations affect the risk of individuals of developing cancer (1). In cancer syndromes, mutations in susceptibility genes, such as the *tumor protein 53* (*TP53*) (2), and the BRCA1/2 DNA Repair Associated (*BRCA1, BRCA2*) genes (3,4), confer up to an 8-fold increase in cancer risk in first degree relatives (1). However, these inherited mutations are rare and highly penetrant and explain only a small fraction of the relative risk for all cancers (5).

It has been hypothesized that part of cancer risk could be apportioned to high-frequency low-penetrant variants, such as single nucleotide polymorphism (SNPs). Genome-Wide Association Studies (GWAS) have been instrumental in identifying SNPs associated with increased risk of cancer in the broader population (6), including breast (7), prostate (8), testicular (9), and blood malignancies (10,11). However, the vast majority of SNPs account only for a limited increase in cancer risk (1) and are usually filtered out by multiple hypotheses correction procedures applied in GWAS analysis (12), which ultimately leaves most of the cancer risk unexplained.

Although most SNPs have only subtle effects, there is mounting evidence suggesting that they still contribute to the risk of developing cancer (6,13). Recently, we have shown that low-penetrant germline mutations in p53 pathway genes can directly control cancer related processes, including p53 activity and response to chemotherapies (14). Moreover, the Pan-Cancer Analysis of Whole Genomes (PCAWG) study found that 17% of all patients have rare germline variants associated with cancer (15). It is now becoming apparent that quantifying the contribution of low-penetrance but high-frequency inherited mutations could further improve our understanding on how inherited mutations mediate cancer risk and tumorigenesis.

Heritability analysis provides the statistical framework to estimate the contribution of all common SNPs to cancer risk regardless of their statistical significance and effect size (16). Studying heritability is now becoming a crucial step in cancer GWAS studies and has provided insights on the risk of developing many malignancies (17), including prostate (18), cervical (19), testicular germ cell tumor (20) and breast cancer (21,22).

However, since the functional impact of the SNPs is context-dependent (23), it is important to quantify the amount of heritability explained by genomic regions associated with well-characterized biological functions (24,25). Recently, the PCAWG study has shown that driver mutations are mostly located in protein-coding rather than regulatory regions (26), albeit few mutations in cis-regulatory regions, such as the TERT promoter, can still mediate cancer phenotypes. Thus, we reasoned that estimating the heritability of SNPs in protein-coding genes and proximal regulatory regions could provide novel insights into the etiology of this disease. However, developing analytical methods for estimating heritability at the gene-level has been challenging, and current methods allow only the estimation of heritability for large functional regions or SNP categories, such as histone marks or expression Quantitative Trait Loci (eQTL) (25,27).

Here we developed a new method, called BAyesian Gene HERitability Analysis (BAGHERA), which, to the best of our knowledge, is the first method to enable heritability analysis both at genome-wide and at gene-level resolution. We performed extensive simulations to validate the robustness of BAGHERA estimates and assess whether our method was prone to false discoveries. Comparison with other state-of-the-art methods (27,28) clearly showed that BAGHERA provides significantly more accurate heritability estimates for traits with heritability lower than 10%, such as cancer.

We then used BAGHERA to analyze all the 38 histologically different malignancies reported in the UK BioBank cohort (29). Our genome-wide heritability analysis showed that SNPs account for at least 10% of the heritable risk of 14 tumors, including late onset malignancies, such as prostate and bladder, which are not thought to be driven by high-frequency inherited mutations. We then used gene-level heritability analysis to build a panel of 1,146 genes, called cancer heritability genes (CHGs), that have a significant contribution to the heritability of at least one cancer. Interestingly, a significant proportion of CHGs are known tumor suppressors or are directly involved in the hallmark processes controlling the transformation from normal to cancer cells.

Our study provides new methods to analyze GWAS data and genetic evidence of a causal role for high-frequency inherited mutations in cancer.

Materials and Methods

Estimation of heritability at the gene level

Narrow sense heritability, h^2, is defined as the amount of phenotype variance explained by additive genetic effects. Genome-wide association studies (GWAS) provide unique opportunities to study heritability of many diseases; in particular, with the advent of high-density arrays, where more than $500{,}000$ SNPs are genotyped, the heritability explained by these variants, h^2_{SNP}, represents a reasonable estimate for h^2.

Our goal is to identify the amount of h^2_{SNP} explained by a protein-coding gene and its proximal regulatory regions. To obtain unbiased heritability estimates and control the number of false positives, we require SNPs to be uniquely assigned to genes.

Hereby, we denote as genome-wide heritability the amount of heritability explained by all genotyped SNPs, M, whereas we refer to the amount of heritability explained by the SNPs in a gene as gene-level heritability. In a model where each SNP has equal contribution to the genome-wide heritability, the per-SNP heritability is $\hbar^2 = h^2_{SNP}/M$. Conversely, if variants can have varying contribution to the genome-wide heritability, we can model the per-SNP heritability as a random variable, $\hbar^2_M$, whose expectation is $\hbar^2_M = \mathrm{E}\left[\hbar^2_j\right]_{j=1,\cdots,M}$, where M denotes the number of SNPs used to average the per-SNP contribution to heritability.

We hereby demonstrate that the genome-wide heritability can be expressed as the sum of the gene-level contribution and that the per-SNP genome-wide heritability is the expectation of the per-SNP gene-level heritability. Let K be the number of non-overlapping genes in the human genome, each of them with M_k SNPs, the genome-wide heritability can be expressed as $h^2_{SNP} = \sum_{k=1}^{K} \sum_{j\in k} \hbar^2_j = \sum_{k=1}^{K} M_k\, \hbar^2_{M_k}$ where $M_k \hbar^2_{M_k}$ is the amount of heritability explained by all the SNPs in the k-th gene. Thus, let the number of SNPs in each gene and the gene-level per-SNP heritability be independent random variables, it is straightforward to prove that the expectation of the gene-level per-SNP heritability is the per-SNP genome-wide estimate $h^2_{SNP}/M = E\left[\hbar^2_{M_k}\right]_K$. However, estimating h^2_{SNP} only from SNPs assigned to genes would lead to biased estimates, since the contribution of the SNPs in intergenic regions would be neglected; thus, SNPs outside genic regions are assigned to a single intergenic locus, such that the heritability is correctly estimated from all genotyped SNPs.

A hierarchical Bayesian model for heritability estimation

The estimation of heritability can be modelled as a hierarchical Bayesian regression problem, which provides a robust approach to simultaneously estimate the genome-wide heritability, h^2_{SNP}, and the gene-level heritability, h^2_k, from the observed data Y. Our base Bayesian regression model can be defined as follows:

$$\begin{array}{ccc} h^2_{SNP} & \sim & F_1(\quad)\, with\; supp\big(F_1(\quad)\big) \in [0,1] \\ h^2_k |\, h^2_{SNP} & \sim & F_2(h^2_{SNP}) \\ Y |\, h^2_k & \sim & F_3\big(h^2_k\big) \end{array} \qquad (1)$$

where F_1, F_2, F_3 are suitable distributions.

SNP heritability, h^2_{SNP}, is the ratio of the variance of the additive genetic effects, σ^2_g, and the phenotypic variance, σ^2_P. Let $\sigma^2_P = \sigma^2_g + \sigma^2_e$, where σ^2_e are the non-additive and environmental effects, these quantities can be modelled as random variables with $\sigma^2_g \sim \Gamma(\alpha,\theta)$ and $\sigma^2_e \sim \Gamma(\beta,\theta)$, respectively. Since $\Gamma\,(\alpha,\theta)/\big(\Gamma(\alpha,\theta) + \Gamma(\beta,\theta)\big) \sim \mathrm{Beta}(\alpha,\beta)$, a suitable distribution for F_1, in Eq. 1, would be an uninformative Beta distribution, e.g. $\mathrm{Beta}(1{,}1)$. In practice, the use of a Beta distribution as prior for h^2_{SNP} allows us to obtain accurate heritability estimates in the unit range even for low-heritability diseases, where classical methods are usually inaccurate (28).

The gene-level heritability, h^2_k, can be modelled as a random variable following a Gamma distribution with shape $\alpha = h^2_{SNP}$ and rate $\beta = 1$. It is worth noting that h^2_k/M is the per-SNP heritability of gene k, whereas the amount of heritability explained by the gene is $M_k\big(h^2_k/M\big)$, where M_k are the SNPs in gene k. While theoretically the Gamma distribution is unbounded, in practice, for $M_k \ll M$, the likelihood of obtaining an estimate h^2_k s.t. $M_k\big(h^2_k/M\big) > 1$ is negligible. Therefore, for $F_2 = Gamma(h^2_{SNP}, 1)$, the expectation would be h^2_{SNP}, which is an unbiased estimator of the genome-wide heritability.

Finally, our model requires a suitable estimator to regress h_k^2 from the observed data. Recently, many methods have been proposed to estimate heritability from GWAS data (30); however, the vast majority requires genotype data, which are both difficult to obtain, due to privacy concerns, and computationally taxing to analyze, because of high dimensionality. Thus, we adopted the LD-score (LDsc) regression model (28), which allows estimation of heritability from GWAS summary statistics, such as regression coefficients and standard errors, which are readily available (12).

Thus, for F_3, we rewrote the LDsc model to estimate gene-level heritability, from summary statistics of M SNPs in a GWAS with N subjects, as follows:

$$\chi_{jk}^2 \sim N\left(Nl_j\, h_k^2/M + e, \sqrt{l_j}\right) \tag{2}$$

where χ_{jk}^2 and l_j are the χ^2 statistic and LD score associated with SNP j in gene k, respectively. The LD score is a quantity defined as $l_j = \sum_z r_{jz}^2$, where r_{jz}^2 is the linkage disequilibrium between variant j and variant z within a certain genomic window (e.g. 1Mb) in a given population. Importantly, LD scores can be conveniently computed from large scale genetic studies, such as the 1000 Genomes project.

Finally, setting the standard deviation to the LD score of the j-th SNP allows us to control for heteroskedasticity of the test statistics due to linkage disequilibrium, somehow similar to the weighting scheme used in LDsc, and a term e accounting for confounding biases, which is modelled using an uninformative normal prior.

The Bayesian Gene HERitability Analysis (BAGHERA) software

We implemented our hierarchical model (see Eq. 3) as part of the BAGHERA software, which allows simultaneous estimation of genome-wide and gene-level heritability, also called heritability loci, which are genes and proximal regulatory regions with a per-SNP heritability higher than the genome-wide estimate (see Fig. 1). Since fitting the Beta-Gamma model is computationally taxing, we relaxed our requirements by modelling h_k^2 as a random variable following a Normal distribution whose mean is the genome-wide heritability, h_{SNP}^2, and the standard deviation is controlled by an uninformative Inverse-Gamma prior. While this formulation might provide gene-level heritability estimates outside the unit domain, we found this problem to be well controlled in practice.

$$\begin{array}{ccc} e & \sim & N(1,1) \\ W & \sim & \text{Inv-Gamma}(1,1) \\ h_{SNP}^2 & \sim & \text{Beta}(1,1) \\ h_k^2 | h_{SNP}^2, W & \sim & N(h_{SNP}^2, W^2) \\ \chi_{jk}^2 | h_k^2, e, l_j, N, M & \sim & N\left(Nl_j\, h_k^2/M + e, \sqrt{l_j}\right) \end{array} \tag{3}$$

BAGHERA predicts heritability genes by computing the posterior distribution of $\eta_k \sim I\left(h_k^2 > h_{SNP}^2\right)$, where I is a function that returns 1 if the evaluated condition is true, and 0 otherwise. The expectation of the posterior distribution of η_k, $\mathrm{E}[\eta_k]$, is the probability of the heritability of a gene k of being higher than the genome-wide estimate; specifically, we report as heritability genes, those with $\mathrm{E}[\eta_k] > 0.99$. For each gene, we also report effect sizes in terms of fold-change with respect to the genome-wide heritability estimate, as $fc_k = h_k^2/h_{SNP}^2$.

We use the No-U-Turn Sampler as implemented in PyMC 3.4 (31), using 4 chains with 10^4 sweeps each and a burnin step consisting of $2{,}000$ samples. Convergence of the sampling process was assessed based on the Gelman-Rubin convergence criterion.

BAGHERA is released as a Python software package under MIT license, and it is available on GitHub (https://github.com/stracquadaniolab/baghera), as a package on Anaconda, and as a Docker image. BAGHERA also implements the Beta-Gamma model described in the previous section, called BAGHERA-Γ. Alongside the source code, we also provide a Snakemake workflow (https://github.com/stracquadaniolab/workflow-baghera) to run the pipeline presented in our study.

UK BioBank summary statistics processing and curation

We used summary statistics of the UK BioBank GWAS for cancers classified using the ICD10 disease classification (source: https://nealelab.github.io/UKBB_ldsc/); importantly, data are uniformly processed with state-of-the-art methods, which prevents any methodological bias. Here, we developed a custom pipeline to assign LD scores to SNPs, and SNPs to human genes (see Fig. 1). Specifically, we used pre-computed LD scores for SNPs on autosomal chromosomes with minor allele frequency MAF > 0.01 in the European population (EUR) of the 1000 Genomes project. We then removed the SNPs on chr6:26,000,000-34,000,000, since this region contains the Major Histocompatibility Complex (MHC) that have unusual genetic patterns and is known to affect GWAS result interpretation (28,32). Ultimately, our analysis is conducted on $1{,}285{,}620$ SNPs over 22 chromosomes.

We then used Gencode v31 to determine the genomic coordinates of protein coding genes in the GRCh37 human genome. First, we merged overlapping genes by creating a new multi-gene locus, whose name denotes the overlapping genes and whose boundaries are defined as the first and last base-pair of these loci. We then assigned to a locus all SNPs within or no more than ± 50kb away from its boundaries (Fig. 1); this strategy allows us to account for cis-regulatory elements while retaining gene-level resolution. All other SNPs are assigned to the intergenic locus. Overall 55% of SNPs were mapped to a locus, while the rest of them are assigned to the intergenic term. Finally, in order to mitigate false positives due to poorly genotyped regions, we considered only gene-loci harboring at least 10 variants. Ultimately, our dataset consists of 15025 loci, 12042 (80.1%) of them are harboring more than 10 SNPs, which were considered in our heritability study. The results of our analyses are deposited in CSV format on Zenodo (doi: 10.5281/zenodo.3968269).

Enrichment analyses

We used a one-tailed Fisher's exact test for all enrichment analyses, with p-values adjusted using the Benjamini-Hochberg procedure, since we are interested in testing whether genes associated with a given category (e.g. molecular function, gene panel) are overrepresented in our set of significant heritability loci. Importantly, since loci in our analysis might represent overlapping protein-coding regions, we post-processed our gene lists by converting each multi-gene locus into the set of its genes. For the Gene Ontology analysis, we used a GO slim annotation to obtain a high-level view of the processes and functions mediated by a set of genes. All external datasets, with their respective date of download, are detailed in the Supplementary Methods.

Results

Simulations assessing robustness of genome-wide and gene-level estimates for low heritability traits

We performed extensive testing of our method on simulated data to assess i) the robustness of genome-wide estimates for low heritability traits and ii) the false discovery rate (FDR) associated with gene-level predictions. All our datasets were calibrated to simulate low heritability traits ($h^2_{SNP} \leq 0.5$), which is a reasonable assumption for cancer. We generated genotype data for $M = 100{,}000$ SNPs of $N = 50{,}000$ subjects using haplotypes of chromosome 1 from European populations under different heritability models (See Supplementary Methods).

Our analyses show that BAGHERA provides robust unbiased genome-wide estimates (see Supplementary Methods); interestingly, while extreme values of gene-level heritability might affect genome-wide estimates, we found that BAGHERA returns correct estimates both as the median of the posterior genome-wide heritability distribution and as the sum of gene-level heritability contributions.

We then assessed whether BAGHERA was able to identify heritability loci, that is loci harboring SNPs with a contribution to heritability higher than expected under a constant per-SNP heritability contribution. To do that, we selected 1% of the loci on chromosome 1 (≈ 13) as heritability loci and computed Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves at varying levels of genome-wide heritability (see Supplementary Methods). For all curves, we evaluate the Area Under the Curve (AUC). Here we found that BAGHERA correctly identified heritability loci (ROC AUC 0.89), although precision and recall were consistently higher for higher genome-wide heritability levels (PR AUC: 0.41 for $h_2 = 0.01$, >0.58 for $h_2 \geq 0.01$; Supplementary Figures 1,2).

However, our simulated datasets have limitations; since simulating genotype data is a computationally taxing task, we restricted the number of simulated SNPs to $M \approx 100{,}000$ SNPs from a single chromosome, whereas more than 1M are routinely genotyped in modern studies.

We addressed this limitation by simulating summary statistics using only linkage disequilibrium information (see Supplementary Methods). This approach provides a tractable framework to test varying levels of heritability enrichment, reported in terms of fold-change with respect to the genome-wide estimate, and to simulate SNPs across the entire genome, rather than a single chromosome.

We then assessed the performance by computing ROC and PR curves, the True Positive Rate (TPR), and the False Discovery Rate (FDR). BAGHERA correctly identifies heritability loci, even with fold-changes in heritability as low as $f_c = 5$ (ROC AUC range:$0.70 - 0.99$). Importantly, we found BAGHERA to be conservative with a low false discovery rate across all scenarios (FDR range: $0 - 5\%$); this result suggests that our method is suitable for exploratory analyses, and that significant results are associated to true biological signal (see Supplementary Figures 3,4,5).

Comparison with state-of-the-art methods for genome-wide and local heritability estimation

To the best of our knowledge, BAGHERA is the first method specifically designed to analyze low heritability traits and to provide heritability estimates with gene-level resolution. However, since our methods can estimate both genome-wide and local heritability up to gene-level resolution, we decided to compare its performance to state-of-the-art methods designed to estimate genome-wide and local heritability.

Genome-wide estimates were compared with LD score regression (LDsc) results (28). Gold-standard methods require raw data; however, previous studies have shown that LDsc has comparable performance in most scenarios (22). Since LDsc is routinely used to estimate heritability for the traits in the UK BioBank, we retrieved the results for all 38 cancers and compared them to BAGHERA estimates. We found strong consensus between the estimates of the two methods (see Supplementary Figure 6), consistent with the fact that BAGHERA uses a similar genome-wide estimator. Nonetheless, BAGHERA is more robust for low heritability traits, since our Bayesian formulation guarantees correct heritability estimates in the unit domain, whereas LDsc incorrectly provides negative values.

Performances on local heritability analysis were compared with the Heritability Estimation from Summary Statistics (HESS) method (27), which is the only available approach to estimate local heritability from summary statistics. Here, we used BAGHERA to estimate the heritability of 1703 regions, as defined in the HESS original study (see Supplementary Methods). We then restricted our analysis to breast and prostate cancer data, since these malignancies are those with the highest h^2_{SNP} estimates; this was necessary to ensure a fair comparison between the two methods, since HESS is not designed for low heritability traits. Here we found a statistically significant correlation between HESS and BAGHERA estimates (Pearson’s ρ: 0.76 for prostate and 0.78 for breast, see Supplementary Figures 7,8). However, since BAGHERA provides robust estimates for as much as 15000 regions, it enables more detailed analyses compared to HESS.

Taken together, we have shown that BAGHERA provides robust estimates for low heritability traits and can identify loci with heritability enrichment up to gene-level, which represent a 10-fold increase in genomic resolution compared to existing methods.

Genome-wide estimates of cancer heritability in the UK Biobank

We used BAGHERA to analyze 38 cancers in the UK Biobank (29), a large-scale prospective study aiming at systematically screening and phenotyping more than $500{,}000$ individuals, with a reported age at the assessment centre ranging between 37 and 73 years.

We obtained summary statistics for $N = 361{,}194$ individuals (see Table 1), including subjects whose tumors were histologically characterized according to the ICD10 classification, where malignant neoplasms are identified with codes ranging from C00 to C97 (see Supplementary Methods). The number of cases varies significantly across cancers, ranging from 102 individuals, for malignant neoplasm of base of tongue (C01), to 9086 individual, for other malignant neoplasms of the skin (C44). In this cohort, cancer prevalence ranges between 0.29% and 2.51%, with higher estimates for common malignancies in European populations, such as breast and prostate cancer (33).

Estimating heritability from non-targeted cohorts can be challenging, due to the small prevalence of the disease. To test whether we had sufficient signal for each cancer, we reasoned that if the SNP test statistic follows a χ^2 distribution with 1 degree of freedom, under the null hypothesis of no association, its expected value is $E[\chi^2] = 1$; thus, similarly to other studies, we expected to have sufficient polygenic signal for our analysis if the average χ^2 was greater than 1 (25). Here we found the vast majority of cancers to have an average $\chi^2 \approx 1$, with only 17 having a deviation greater than 1% from the expected value of the test statistic. We also did not consider cancers assigned to other malignant neoplasm of the skin (C44), since i) most tumors belong to unspecified anatomical regions (C44.3, C44.9), ii) are predominantly caused by sun exposure in Europeans and iii) and includes poorly characterized rare skin cancers. Ultimately, we restricted our study to 16 cancers for which we had sufficient power to perform our analysis. Nonetheless, all our results are consistently aligned with those we obtained when considering all 38 cancer types (see Supplementary Figures 9,10,11,14,15, Supplementary Tables 5,6,7).

We then estimated genome-wide heritability of each cancer by computing the median of the posterior distribution of h^2_{SNP} and transforming this value on to the liability scale, $h^2_{SNP_L}$, to obtain estimates independent from prevalence and comparable across malignancies. We found cancer heritability to be $h^2_{SNP_L} = 14.7\%$ on average, ranging from 8% for non-Hodgkin's lymphoma and up to 31% for testis (see Table 1) consistent with other available estimates for this cohort (see Supplementary Materials and Supplementary Figures 16,17,18, Supplementary Table 8). While comparison between cancer heritability estimates is usually difficult across studies, due to differences in histological classification and genetic confounders, we found our heritability estimates on the liability scale to be consistent with those reported for other cohorts, in particular for breast, prostate, testes and bladder (17,18,20,34). The heritability of testicular cancer is the highest among all malignancies ($h^2_{SNP_L} = 0.3158$), consistent with the hypothesis that germline variants have stronger effects in early onset and young adult cancers. However, lethal early onset cancers are underrepresented in the UK Biobank, since children and young adults were not enrolled in the study, and thus an accurate estimation of the correlation between age of onset and heritability is not possible. Nonetheless, it is interesting to note that many malignancies with onset in late adulthood, such as prostate or bladder, still display a significant heritable component, ranging from $h^2_{SNP_L} = 0.25$ for brain tumors (age of onset: 59) to $h^2_{SNP_L} = 0.08$ for diffuse non-Hodgkin's lymphoma (age of onset: 60). Overall, 14 out of 16 cancers (87%) show heritability higher than 10% suggesting a consistent contribution of SNPs to the heritable risk of cancer.

Heritability loci across 16 malignancies

We identified 783 heritability loci ($\eta > 0.99$), harbouring $1{,}146$ protein-coding genes, across 16 cancers (see Figure 2), with 53 heritability loci per malignancy on average, ranging from 5 loci in mesothelioma, to 271 loci for prostate (see Table 1, Figure 3A); here we are using the term heritability loci when referring to the non-overlapping genomic regions tested by BAGHERA, which might also include multi-gene loci. Gene-level heritability across the selected 16 cancers has a long-tail distribution (Figure 3B), with a median 16-fold increase compared to the genome-wide estimate, ranging from 4.4-fold for the *Phosphodiesterase 4D (PDE4D)* gene locus to 276-fold for the *fibroblast growth factor receptor 2 (FGFR2)* gene locus in breast cancer. Interestingly, 87% of heritability loci show per-SNP heritability 10-fold higher than the genome-wide estimate. Only 3 loci have fold changes below 5 and more than 99% of loci with fold-changes below 10 are found in the breast and prostate datasets, which have $h^2_{SNP} > 0.01$. Based on our simulations, our set of heritability loci are expected to have a limited number of false positives.

Interestingly, heritability loci represent less than 1% of all the loci in the genome, but they are significantly more than those harboring genome-wide significant SNPs (see Supplementary Materials, Supplementary Figures 12,13, Supplementary Table 4); this result is consistent with cancer being polygenic. Although we identified a polygenic signal, heritability loci account for up to 38% of all the heritable risk (breast cancer), suggesting that a significant amount of heritability could be explained by only few loci across the genome (Figure 3A). Consistent with our hypotheses, when we looked at the contribution of SNPs in intergenic regions, we did not find any heritability enrichment.

We then tested whether heritability loci were shared among multiple cancers to identify any potential genomic hotspot for pan-cancer heritability. We found that only 59 ($\approx 8\%$) of the 783 heritability loci show a significant heritability enrichment in at least 2 cancers, and 8 (≈1%) in 3 or more (Figure 3C-D). This observation is consistent with results from tumor sequencing studies, which have shown that pleiotropic effects are limited to few master regulators, such as *TP53* (35). Nonetheless, after performing literature curation, we found evidence for a cancer mediating role for 7 of the 11 unique protein coding genes found in at least 3 cancers (see Supplementary Table 1), including 4 genes (CLPTM1L, APAF1, THADA, AGBL1) involved in apoptosis and 3 genes (PCDH15, DLG2, POU5F1B) involved in cell division, migration and tumorigenesis (36,37). It is important to note that the *cisplatin resistance-related protein 9* (CLPTM1L) is the heritability locus found in most cancers (4 out of 16) and is one of the gene in the 5p15.33 locus (the other being TERT), which has been consistently associated with different cancer types (38).

Taken together, our analysis found 783 loci, harbouring 1,146 protein-coding genes, having a significant contribution to the heritable risk of at least 1 cancer. We denoted these 1,146 genes as cancer heritability genes (CHGs).

Cancer heritability genes are recurrently mutated in tumors

Tumor sequencing projects, including The Cancer Genome Atlas (TCGA) program and the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, have identified a number of driver genes, which promote tumorigenesis when acquiring a somatic mutation.

There is also increasing evidence that genes harboring germline and somatic mutations can mediate cancer phenotypes (14,39), thus we tested whether cancer heritability genes are significantly enriched among known cancer driver genes. To do that, we obtained a curated list of driver genes using the COSMIC Cancer Gene Census (Supplementary Table 2). Interestingly, we found that a significant proportion of cancer heritability genes, 60 out of 1,146 ($\approx 5\%$), are also known cancer driver genes ($OR = 1.75$, $P: 1.3 \times 10^{-4}$). These genes include members of the p53 pathway, such as *CDKN2A*, the *Tumor Protein 63* (*TP63*) and *MDM4 regulator of p53* (*MDM4*), as well as genes mutated across multiple types of cancer, including *FGFR2* and the *anaplastic lymphoma kinase (Ki-1)* (*ALK*) gene (Figure 4A and 4B).

However, the number of cancer driver genes is extremely variable across malignancies and studies, thus we tested whether the enrichment of CHGs in cancer driver genes was independent from the cancer driver gene annotation used. To do that, we collected lists of cancer driver genes from multiple studies, including the PCAWG project (15), the Precision Oncology Knowledge Base (OncoKB, (40)), Memorial Sloan Kettering Impact and Heme gene panels (41), and the curated list of cancer genes by Vogelstein et al. (42). Here we found that CHGs are significantly enriched in each cancer driver gene annotation analyzed, with an enrichment ranging from OR=1.55 for the PCAWG annotation to OR=2.47 for OncoKB tumour suppressors (Supplementary Table 2). Interestingly, we did not find any enrichment of CHGs in genes carrying germline driver mutations; this is consistent with the fact that most germline driver mutations are rare, and thus are unlikely to be genotyped in GWAS studies.

Taken together, we found 60 cancer heritability loci that are also recurrently mutated in multiple tumours; this result suggests that SNPs in cancer heritability genes might affect the same biological programs altered by somatic mutations in tumors.

Cancer heritability genes underpin biological processes affecting tumorigenesis

Our gene-level heritability analysis identified 1,146 genes explaining a significant proportion of the heritable risk of at least 1 cancer. We then showed that cancer heritability genes are enriched in known cancer driver genes, suggesting that loci recurrently mutated in tumours also harbour high-frequency inherited mutations that could mediate cancer risk. Thus, we hypothesized that cancer heritability genes could be involved in molecular functions and biological processes affecting tumorigenesis.

To do that, we characterized CHGs by gene ontology (GO) enrichment analysis (see Table 2). We found a statistically significant enrichment for 21 terms (Fisher's exact test; False Discovery Rate, FDR < 10%, Figure 4C), with an average odds ratio of 1.31 and up to 1.55 for growth. CHGs are genes predominantly involved in biological processes driving cell morphogenesis, differentiation, proliferation and growth, which include the *mammalian target of rapamycin (mTOR)* and the *Poly [ADP-ribose] polymerase 1* (*PARP1*) genes. We also observed a significant enrichment of genes associated with cytoskeleton organization and anatomical structure development, which includes the *Mothers against decapentaplegic homolog 2* (SMAD2) gene.

While these molecular processes drive normal cell fate, survival and proliferation, they are recurrently hijacked by cancer cells to gain growth advantage and spread through the body through metastases (43), a process that is considered an hallmark of cancer. We then tested whether cancer heritability genes are associated with any other hallmark of cancer, which are processes, common to all malignancies, controlling the transformation of normal into cancer cells (44). These lists of biological processes include proliferative signaling, suppression of growth, escaping immune response, cell replicative immortality, promoting inflammation, invasion and metastasis, angiogenesis, genome instability and mutation, and escaping cell death. Interestingly, we found 33

CHGs associated with at least one hallmark (OR: 2.062, P: 3×10^{-4}). Consistent with our previous analysis, cancer heritability loci are involved in escaping cell death, mediating proliferative signaling, invasion and metastasis (Figure 4D and Supplementary Table 3). We then went further to understand whether CHGs mediate these cancer processes by acting either as tumor suppressor genes (TSGs) or oncogenes (see Fig. 4E). To do that, we used the Precision Oncology Knowledge Base (OncoKB, (40)), a curated list of 519 cancer genes, including 197 tumour suppressor genes (TSGs), 148 oncogenes and other cancer genes of unknown function. We found that 27 CHGs are tumour suppressors (OR: 2.47, P: 7.9×10^{-6}), whereas 17 are reported as oncogene (OR: 1.83, P: 0.0198) of which 4 can function both as TSG and as oncogene (Figure 4A, D and E and Supplementary Tables 2 and 3); importantly, this result has been also confirmed when using the COSMIC Cancer Gene Census TSG annotation (OR: 2.036, P: 2.07×10^{-4}). Tumour suppressor CHGs include well-known cancer driver genes, such as *CDKN2A and SMAD2*, which regulate cell growth, and DNA repair genes, such as *MUTYH* and *FANCA* (45).

Taken together, we found evidence that cancer heritability genes directly mediate processes underpinning tumorigenesis; interestingly, while we did not observe pleiotropic effects at genomic level, we found that cancer heritability genes are involved in biological processes common to all cancers. It is then conceivable that inherited mutations in genes controlling these biological programs could provide a selective advantage to cancer cells, once they acquire a driver somatic mutation. Our results suggest a functional role for cancer heritability genes consistent with a two-hit model (46); while inherited mutations associated with oncogene activation are likely to be under purifying selection, mutations in tumor suppressor genes can be observed at higher frequency because deleterious effects are only observed upon complete loss of function.

Discussion

Our study provides new fundamental evidence demonstrating a strong contribution of high-frequency inherited mutations to the heritable risk of cancer. We found that SNPs account for at least 10% of the heritable risk of 14 malignancies, and their contribution is not only limited to early onset cancers, but also malignancies with a late age of onset, such as bladder and prostate.

We then went further and built a high-resolution map of the heritable cancer genome consisting of 1,146 genes showing a significant contribution to cancer heritability. We then showed that CHGs are responsible for controlling growth, cell morphogenesis and proliferation, which are fundamental processes required for tumorigenesis. Interestingly, we found that a significant proportion of CHGs (60 out of 1,146) are also recurrently mutated across many tumors, including well known driver genes such as *FGR2*, *CDKN2A* and *SMAD2*. Importantly, 27 out of 60 (45%) are known tumor suppressors genes (TSGs), suggesting that SNPs might support cancer by hijacking tumor suppressor functions. Ultimately, our results suggest that inherited mutations in tumor suppressor genes could create a favorable genetic background for tumorigenesis. It is conceivable that SNPs make normal cells more likely to evade the cell-cell contact inhibition of proliferation, to elude the anatomical constrains of their tissue and to achieve more easily independent motility in presence of other early oncogenic events; evidence supporting these mechanisms has been recently found in advanced urothelial cancer (47). Thus, combining germline and somatic genetic information of key cancer genes could facilitate the identification of subpopulations of patients at higher risk, differential response to treatment and risk of relapse. Nonetheless, determining the heritability threshold to justify the integration of genes carrying low penetrant mutations into clinical cancer genetics will require further investigation.

However, a causal role for many CHGs cannot be ascertained only by genetic analysis and will require further experimental validation. Of particular interest is the subset of CHGs belonging to the Solute carrier (SLC) family (48). SLCs might support cancer metabolism, and polymorphisms in these loci could provide a strong basis for interaction with environmental risk factors such as fats, carcinogens, metal ion deficiencies, and thus could be integrated with future dietary studies, since risk factors may be greater in subgroups of patients.

Obtaining a genomic map with gene-level resolution required the development of a new method, we called BAyesian Gene HERitability Analysis (BAGHERA), for estimating heritability of low heritability traits at the gene-level; to the best of our knowledge, BAGHERA is the first method to enable heritability analysis with gene-level resolution. We performed extensive simulations to show that our method provides robust genome-wide and gene-level heritability estimates across different genetic architectures and outperforms existing methods when used to analyze low heritability traits, such as cancer.

We also recognize the limitations of our work. While our method provides accurate estimates of genome-wide heritability, extremely low heritability diseases could lead to negative gene-level heritability estimates; this was a trade-off to ensure reasonable computational efficiency, although a rigorous model is provided as part of our software. Our analysis does not incorporate functional information, such as gene expression or stratified effects for synonymous/non-synonymous variants, which limits our power of detecting tissue-specific contributions and single causal variants. Finally, since BAGHERA works at single-gene level using summary statistics, analyzing tumors triggered by multi-hit events might still require genotype data.

Taken together, our study provides new insights on the genetic architecture of cancer with gene-level resolution. We expect that integrating heritability information of cancer genes, along with other cancer heritability genes linked to environmental risk factors and somatic information, will help define more effective early detection and surveillance strategies for the broader population.

Acknowledgments.

The authors would like to thank Dr Yongjin Park at the University of British Columbia for the useful discussions about GWAS data simulation.

References.

1. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nature reviews Cancer. 2017;17:692–704.

2. Malkin D, Li FP, Strong LC, Fraumeni JF, Nelson CE, Kim DH, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science (New York, NY). 1990;250:1233–8.

3. Miki Y, Swensen J, Shattuck-Eidens D, Futreal P, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266:66–71.

4. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995;378:789–92.

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. Nature Publishing Group; 2009;461:747–53.

6. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics. 2017;101:5–22.

7. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nature Genetics. 2012;44:312–8.

8. Eeles RA, Olama AAA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nature Genetics. 2013;45:385–91.

9. Wang Z, McGlynn KA, Rajpert-De Meyts E, Bishop DT, Chung CC, Dalgaard MD, et al. Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor. Nature Genetics . Nature Publishing Group; 2017;49:1141–7.

10. Law PJ, Berndt SI, Speedy HE, Camp NJ, Sava GP, Skibola CF, et al. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. Nature Communications . 2017;8:14175.

11. Vijayakrishnan J, Studd J, Broderick P, Kinnersley B, Holroyd A, Law PJ, et al. Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. Nature Communications . 2018;9:1340.

12. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics. 2017;18:117–27.

13. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell . 2017;169:1177–86.

14. Zhang P, Kitchen-Smith I, Xiong L, Stracquadanio G, Brown K, Richter P, et al. Germline and somatic genetic variants in the p53 pathway interact to affect cancer risk, progression and drug response. bioRxiv. Cold Spring Harbor Laboratory; 2019;835918.

15. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. Nature Research; 2020;578:82–93.

16. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era - Concepts and misconceptions. Nature Reviews Genetics . 2008;9:255–66.

17. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. Journal of the National Cancer Institute . Oxford University Press; 2015;107:djv279.

18. Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, et al. The contribution of rare variation to prostate cancer heritability. Nature Genetics . Nature Publishing Group; 2016;48:30–5.

19. Chen D, Cui T, Ek WE, Liu H, Wang H, Gyllensten U. Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability. Carcinogenesis; 2015;36:992–8.

20. Litchfield K, Thomsen H, Mitchell JS, Sundquist J, Houlston RS, Hemminki K, et al. Quantifying the heritability of testicular germ cell tumour using both population-based and genomic approaches. Scientific Reports . Nature Publishing Group; 2015;5:13889.

21. Sapkota Y. Germline DNA variations in breast cancer predisposition and prognosis: a systematic review of the literature. Cytogenetic and genome research . Karger Publishers; 2014;144:77–91.

22. Fanfani V, Zatopkova M, Harris AL, Pezzella F, Stracquadanio G. Dissecting the heritable risk of breast cancer: from statistical methods to susceptibility genes. Seminars in cancer biology. NLM (Medline); 2020;

23. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS genetics. Public Library of Science; 2013;9:e1003449.

24. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature Genetics . Springer US; 2018;50:621–9.

25. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics . Nature Publishing Group; 2015;47:1228–35.

26. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. Nature Publishing Group; 2020;578:102–11.

27. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. American Journal of Human Genetics . American Society of Human Genetics; 2016;99:139–53.

28. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. Nature Publishing Group; 2015;47:291.

29. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature . 2018;562:203–9.

30. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. The American Journal of Human Genetics. Elsevier; 2011;88:76–82.

31. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in python using PyMC3. PeerJ Computer Science. PeerJ Inc.; 2016;2:e55.

32. Kennedy AE, Ozbek U, Dorak MT. What has GWAS done for HLA and disease associations? International Journal of Immunogenetics . Blackwell Publishing Ltd; 2017;44:195–211.

33. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians . 2018

34. Jiang X, Finucane HK, Schumacher FR, Schmit SL, Tyrer JP, Han Y, et al. Shared heritability and functional enrichment across six solid cancers. Nature Communications . Nature Publishing Group; 2019;10:431.

35. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell . 2018;173:371–385.e18.

36. Shao YW, Wood GA, Lu J, Tang Q-L, Liu J, Molyneux S, et al. Cross-species genomics identifies DLG2 as a tumor suppressor in osteosarcoma. Oncogene. Nature Publishing Group; 2019;38:291–8.

37. Hayashi H, Arao T, Togashi Y, Kato H, Fujita Y, De Velasco M, et al. The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer. Oncogene. Nature Publishing Group; 2015;34:199–208.

38. Rafnar T, Sulem P, Stacey SN, Geller F, Gudmundsson J, Sigurdsson A, et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nature genetics. Nature Publishing Group; 2009;41:221–7.

39. Qing T, Mohsen H, Marczyk M, Ye Y, O'Meara T, Zhao H, et al. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. Nature Communications. Nature Publishing Group; 2020;11:1–8.

40. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A precision oncology knowledge base. JCO precision oncology. American Society of Clinical Oncology; 2017;1:1–6.

41. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. The Journal of molecular diagnostics. Elsevier; 2015;17:251–64.

42. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. science. American Association for the Advancement of Science; 2013;339:1546–58.

43. Stuelten CH, Parent CA, Montell DJ. Cell motility in cancer invasion and metastasis: Insights from simple model organisms. Nature Reviews Cancer . Nature Publishing Group; 2018;18:296–312.

44. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. cell. Elsevier; 2011;144:646–74.

45. Lange SS, Takata K, Wood RD. DNA polymerases and cancer. Nature reviews cancer. Nature Publishing Group; 2011;11:96.

46. Knudson AG. Mutation and cancer: Statistical study of retinoblastoma. Proceedings of the National Academy of Sciences. National Acad Sciences; 1971;68:820–3.

47. Vosoughi A, Zhang T, Shohdy KS, Vlachostergios PJ, Wilkes DC, Bhinder B, et al. Common germline-somatic variant interactions in advanced urothelial cancer. Nature Communications. Nature Publishing Group; 2020;11:1–3.

48. Zhang Y, Zhang Y, Sun K, Meng Z, Chen L. The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. Journal of molecular cell biology. Oxford University Press; 2019;11:1–3.

Tables

***Table 1. Genome-wide heritability of the** 38 **cancers in the UK BioBank.** For each cancer, we report the number of cases, the prevalence in the cohort, the average χ^2 of the SNPs considered in the GWAS analysis ($\hat{\chi}^2$), the genome-wide estimates of heritability, both on the observed (h^2_{SNP}) and the liability ($h^2_{SNP_L}$) scale, and the number of heritability loci (HL) reported by BAGHERA as significant for $\eta > 0.99$. In bold, we denote the 16 cancers that we used for the downstream analysis analysis and functional characterisation.*

ICD10	Malignancy	Cases	Prevalence	$\hat{\chi}^2$	h^2_{SNP}	$h^2_{SNP_L}$	HL
C44	Other malignant neoplasms of skin	9086	0.0252	1.1408	0.0341	0.2422	422
C50	**Malignant neoplasm of breast**	8304	0.0230	1.0869	0.0170	0.1285	267
C61	**Malignant neoplasm of prostate**	4342	0.0120	1.0765	0.0191	0.2320	271
C18	**Malignant neoplasm of colon**	2226	0.0062	1.0399	0.0070	0.1416	33
C43	**Malignant melanoma of skin**	1672	0.0046	1.0288	0.0051	0.1293	52
C15	**Malignant neoplasm of oesophagus**	519	0.0014	1.0236	0.0035	0.2296	24
C67	**Malignant neoplasm of bladder**	1554	0.0043	1.0222	0.0047	0.1254	39
C34	**Malignant neoplasm of bronchus and lung**	1427	0.0040	1.0208	0.0035	0.1010	17
C20	**Malignant neoplasm of rectum**	1118	0.0031	1.0130	0.0031	0.1091	15
C62	**Malignant neoplasm of testis**	221	0.0006	1.0120	0.0024	0.3158	29
C71	**Malignant neoplasm of brain**	368	0.0010	1.0116	0.0030	0.2578	19
C45	**Mesothelioma**	150	0.0004	1.0110	0.0012	0.2213	5
C91	**Lymphoid leukaemia**	349	0.0010	1.0109	0.0018	0.1646	11
C02	**Malignant neoplasm of other and unspecified parts of tongue**	152	0.0004	1.0106	0.0013	0.2475	23
C16	**Malignant neoplasm of stomach**	388	0.0011	1.0106	0.0010	0.0868	12
C83	**Diffuse non-Hodgkin's lymphoma**	587	0.0016	1.0104	0.0014	0.0824	14
C82	**Follicular (nodular) non-Hodgkin's lymphoma**	320	0.0009	1.0101	0.0031	0.3059	21
C90	Multiple myeloma and malignant plasma cell neoplasms	401	0.0011	1.0092	0.0013	0.1020	15
C56	Malignant neoplasm of ovary	693	0.0019	1.0063	0.0012	0.0616	13
C54	Malignant neoplasm of corpus uteri	988	0.0027	1.0063	0.0008	0.0295	14
C48	Malignant neoplasm of retroperitoneum and peritoneum	122	0.0003	1.0053	0.0009	0.2064	5
C64	Malignant neoplasm of kidney except renal pelvis	701	0.0019	1.0043	0.0009	0.0455	10
C01	Malignant neoplasm of base of tongue	102	0.0003	1.0043	0.0014	0.3596	10
C73	Malignant neoplasm of thyroid gland	278	0.0008	1.0042	0.0011	0.1254	13
C49	Malignant neoplasm of other connective and soft tissue	222	0.0006	1.0040	0.0017	0.2229	28
C80	Malignant neoplasm without specification of site	398	0.0011	1.0040	0.0016	0.1300	14
C53	Malignant neoplasm of cervix uteri	192	0.0005	1.0039	0.0005	0.0709	14
C22	Malignant neoplasm of liver and intrahepatic bile ducts	189	0.0005	1.0031	0.0009	0.1353	7
C21	Malignant neoplasm of anus and anal canal	139	0.0004	1.0027	0.0007	0.1436	23
C85	Other and unspecified types of non-Hodgkin's lymphoma	762	0.0021	1.0023	0.0013	0.0600	9
C09	Malignant neoplasm of tonsil	162	0.0004	1.0022	0.0006	0.1009	5
C92	Myeloid leukaemia	328	0.0009	1.0011	0.0008	0.0764	9
C17	Malignant neoplasm of small intestine	114	0.0003	1.0007	0.0015	0.3596	12
C19	Malignant neoplasm of rectosigmoid junction	498	0.0014	0.9992	0.0006	0.0390	10
C25	Malignant neoplasm of pancreas	403	0.0011	0.9991	0.0005	0.0402	12
C81	Hodgkin's disease	150	0.0004	0.9989	0.0003	0.0597	5
C69	Malignant neoplasm of eye and adnexa	137	0.0004	0.9970	0.0004	0.0705	14
C32	Malignant neoplasm of larynx	159	0.0004	0.9914	0.0003	0.0450	7

Table 2. Gene ontology enrichment analysis of cancer heritability genes. *We report the gene ontology terms significantly associated with cancer heritability genes, at 10%FDR. For each term, we report the number of annotated CHGs, the odds ratio, the p-value from the Fisher's Exact test and the adjusted p-value after applying the Benjamini-Hochberg procedure.*

GO term	No. CHGs	OR	p-value	FDR
anatomical structure development	352	1.31	0.000044	0.006133
kinase activity	126	1.44	0.000237	0.012169
growth	84	1.55	0.000263	0.012169
DNA metabolic process	82	1.53	0.000481	0.016723
cytoskeleton organization	120	1.39	0.000861	0.023924
ion binding	431	1.22	0.001248	0.028903
biosynthetic process	361	1.21	0.002711	0.041872
biological_process	505	1.20	0.002224	0.041872
cell morphogenesis	81	1.43	0.002419	0.041872
cell proliferation	146	1.30	0.003404	0.047312
cytoskeleton	141	1.28	0.005851	0.054216
cellular protein modification process	275	1.21	0.004476	0.054216
cell-cell signaling	123	1.30	0.005097	0.054216
peptidase activity	103	1.33	0.005513	0.054216
DNA binding transcription factor activity	160	1.27	0.005068	0.054216
enzyme binding	178	1.24	0.006568	0.057059
cell differentiation	268	1.20	0.007776	0.063577
embryo development	77	1.36	0.009437	0.069042
cytoskeletal protein binding	77	1.36	0.009173	0.069042
nucleus	347	1.16	0.014507	0.097916
DNA binding	174	1.21	0.014793	0.097916

Figures

***Figure 1**: **BAyesian Gene-level HERitability Analysis (BAGHERA) workflow.** Here we show the 4 steps required to run gene-level heritability analysis with BAGHERA. 1) In the preprocessing step, SNP summary statistics are retrieved, and genes are processed, such that a multi-gene locus is created when two or more genes are overlapping. 2) SNPs are assigned to the closest gene locus within 50kb. For example, the SNP marked with a star is within 50kb from both D;E and F, but it is assigned to locus F, which is closer. SNPs farther than 50kb from any gene locus are considered intergenic. 3) BAGHERA uses the No U-Turn Sampler (NUTS) (on the left) to fit our hierarchical Bayesian model to estimate genome-wide and gene-level heritability. The sampler estimates the posterior distributions of the heritability terms (on the right) and evaluates the indicator function to identify loci explaining a significant amount of heritability. When $\eta > 0.99$, the locus is considered significant. 4) Finally, results are saved into CSV format to facilitate downstream analyses. It is worth noting that h^2_{SNP} is the estimate for genome-wide heritability and it is calculated for the malignancy rather than per-locus.*

***Figure 2: Cancer heritability loci across the human genome.** For each chromosome, we report all cancer heritability loci with heritability enrichment in the top 1%. In case of a multi-gene locus, we report only the first gene name of the locus.*

***Figure 3: Heritability loci across cancers in the UK Biobank.** A) For each malignancy, we report the observed heritability (h^2_{SNP}, left box), the heritability on the liability scale ($h^2_{SNP_L}$, dark barplot, between 0 and 0.5), the percentage of h^2_{SNP} explained*

by heritability loci (central barplot, the percentage explained by HLs is highlighted with a darker shade) and the number of heritability loci (right barplot). B) Gene-level heritability distribution across heritability loci, expressed as fold-change with respect to the genome-wide estimate. The x-axis is bound to the minimum and maximum values of fold-change. We highlighted the top locus (FGFR2) and the median (15.9) fold-change across all cancers. C) Percentage of cancer heritability loci associated with multiple cancers. Approximately 8% of HLs are common to multiple malignancies. D) Cancer heritability loci associated with multiple cancers. We report the 59 HLs common to at least 2 cancers; here the size of the dot is proportional to the fold-change of the locus in the specific cancer.

***Figure 4: Functional characterization of cancer heritability genes.** A) List of CHGs reported as cancer driver genes across multiple annotations. With the blue hue (first three columns), we report the genes annotated by OncoKB, specifying whether they are tumor suppressors (TSG) or oncogenes (OG). With red and orange, 4-th and 5-th columns, we report the genes that are included in the COSMIC annotation as drivers and whether the reported mutation is somatic and germline. In the last four columns, we annotate each gene to the cancer type for which is denoted as driver in COSMIC. B) Enrichment of CHGs across cancer driver genes annotations; here we report OncoKB (purple), COSMIC database (light blue), different cancer driver sets (dark blue) and other sets (green) like DNA repair genes and known actionable targets. Stars indicate statistical significance, with multiple terms having P<10^{-4},. C) Gene Ontology enrichment analysis using Fisher's exact test. For each significant term, we report the odds-ratio (x-axis) and $-log10$(FDR) (color gradients). D) CHGs associated with the hallmark of cancers; genes in darker grey are tumour suppressors. Each gene is connected to the hallmarks that it mediates according to the Cancer Gene Census. E) Tumor suppressor and oncogene CHGs across cancers. For each cancer type (y-axis), we report the number of genes (x-axis) reported as tumor suppressors (TSGs) and/or oncogenes in OncoKB (color codes, cancer genes are known to be drivers, but their specific role is not reported).*