Gene-level heritability analysis explains the polygenic architecture of cancer

Viola Fanfani [*], Luca Citi [†], Adrian L. Harris, [‡]Francesco Pezzella, [§]
Giovanni Stracquadanio [¶]

August 3, 2020

Abstract

Genome-wide association studies (GWAS) have found hundreds of single nucleotide polymorphisms (SNPs) associated with increased risk of cancer. However, the amount of heritable risk explained by these variants is limited, thus leaving most of cancer heritability unexplained.

Recent studies have shown that genomic regions associated with specific biological functions explain a large proportion of the heritability of many traits. Since cancer is mostly triggered by aberrant genes function, we hypothesised that SNPs located in protein-coding genes could explain a significant proportion of cancer heritability.

To perform this analysis, we developed a new method, called Bayesian Gene HERitability Analysis (BAGHERA), to estimate the heritability explained by all the genotyped SNPs and by those located in protein coding genes directly from GWAS summary statistics.

By applying BAGHERA to the 38 cancers reported in the UK Biobank, we identified $1,146$ genes explaining a significant amount of cancer heritability. We found these genes to be tumour suppressors directly involved in the hallmark processes controlling the transformation from normal to cancer cell; moreover, these genes also harbour somatic driver mutation for many tumours, suggesting a two-hit model underpinning tumorigenesis.

Our study provides new evidence for a functional role of SNPs in cancer and identifies new targets for risk assessment and patients' stratification.

1 Introduction

Decades of research have shown that inherited genomic mutations affect the risk of individuals of developing cancer [1, 49]. In cancer syndromes, mutations in susceptibility genes, such as the *tumour protein 53* (*TP53*) [30], and the BRCA1/2 DNA Repair Associated (*BRCA1, BRCA2*) genes[34, 58], confer up to an 8-fold increase in cancer risk in first degree relatives [49]. However, these inherited mutations are rare and highly penetrant and explain only a small fraction of the relative risk for all cancers [32].

[*] Institute of Quantitative Biology, Biochemistry, and Biotechnology, SynthSys, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3BF, UK

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

[‡] Molecular Oncology Laboratories, Department of Oncology, The Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK.

[§] Nuffield Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Oxford, UK

[¶] Institute of Quantitative Biology, Biochemistry, and Biotechnology, SynthSys, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3BF, UK. Phone: +44 (0) 131 6507193, Email: giovanni.stracquadanio@ed.ac.uk. Corresponding author.

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 [55], including breast [13, 52, 19], prostate [51, 14], testicular [27, 57, 29], chronic lymphocytic leukaemia [12, 46, 26], acute lymphocytic leukaemia [53, 36] and several lymphomas [11, 15]. However, the vast majority of SNPs account only for a limited increase in cancer risk [49, 47] and are usually filtered out by multiple hypotheses correction procedures applied in GWAS analysis [37].

Although most SNPs have only subtle effects, there is mounting evidence suggesting that they still contribute to the risk of developing cancer [55, 3]. 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 [60]. Moreover, the Pan-Cancer Analysis of Whole Genomes (PCAWG) study found that 17% of all patients have rare germline variants associated with cancer [7]. It is now becoming apparent that quantifying the contribution of low-penetrance inherited mutations can improve our understanding of cancer risk and the aetiology of the disease.

Heritability analysis provides the statistical framework to estimate the contribution of all common SNPs to cancer risk regardless of their statistical significance [54]. The study of heritability is now becoming a crucial step in recent cancer GWAS studies and has already provided insights on the risk of developing many malignancies [41], including prostate [31], cervical [9], testicular germ cell tumour [28] and breast cancer [42, 16].

However, since the functional impact of the SNPs is context-dependent [43], it is important to quantify the amount of heritability explained by genomic regions associated with well-characterised biological functions [17, 18, 45]. For cancer, in particular, which is mostly driven by mutations in genes rather than regulatory regions, estimating the heritability of SNPs in protein-coding genes could provide novel insights into the aetiology 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 eQTL [18, 45].

Here we developed a new method, called BAyesian Gene HERitability Analysis (BAGHERA), which implements a hierarchical Bayesian model to obtain simultaneous estimates of the heritability explained by all genotyped SNPs (genome-wide heritability) and by those in protein coding genes (gene-level heritability). BAGHERA is specifically designed to analyse traits with putative low heritability, such as cancer, and to use GWAS summary statistics rather than genotype data; this facilitates cancer heritability analysis across different studies and cancer types. We performed extensive simulations to assess whether BAGHERA was suitable to study cancer GWAS and found that our method provides robust, unbiased genome-wide heritability estimates, and simultaneously identifies genes explaining a higher proportion of heritability, while controlling the false discovery rate. Comparison with other state-of-the-art methods clearly showed that BAGHERA provides significantly more accurate heritability estimates for diseases with heritability lower than 10%.

We then used BAGHERA to analyse the 38 histologically different malignancies reported in the UK Biobank cohort [6]. Here we provide new genome-wide estimates for all cancers and a map of $1,146$ genes that have a significant contribution to the heritability of at least 1 cancer. We then showed that the vast majority of these genes are tumour suppressors and are directly involved in the hallmark processes controlling the transformation from normal to cancer cells. While we observed pleiotropy across cancers at the functional level, we did not observe pleiotropy at the gene level; this result suggests that while the functional mechanisms mediating risk are common to all cancers, the genes affecting these processes are cancer specific.

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

2 Results

2.1 Simulation results

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. Finally, we compared our method with state-of-the-art approaches for genome-wide and local heritability analysis. All our datasets were calibrated to simulate low heritability traits ($h^2_{SNP} \leq 0.5$), which is a reasonable assumption for cancer. Our analyses show that BAGHERA provides robust and unbiased genome-wide and gene-level heritability.

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

We generated genotype data for $M = 100,000$ SNPs of $N = 50,000$ subjects using haplotypes of chromosome 1 from European populations (See Supplementary Methods). Here we studied two different heritability effects, denoted as dense and gene-level effects; while the former defines constant per-SNP heritability $\hbar^2 = h^2_{SNP}/M$, the latter condition the amount of explained heritability as a function of the gene harbouring the SNP.

Our analyses shows that BAGHERA provides robust unbiased genome-wide estimates under both dense (Fig. 1A) and gene-level heritability (Fig. 1B) models. Interestingly, while extreme values of gene-level heritability might affect genome-wide estimates, we found that BAGHERA returns robust estimates both as the median of the posterior genome-wide heritability distribution (Fig. 1A-B, green diamond marker) and the sum of single gene-level heritability contributions (Fig. Fig. 1A-B, red square marker).

We then assessed whether BAGHERA was able to identify heritability genes, that is genes harbouring SNPs with a contribution to heritability higher than expected under a constant per-SNP heritability contribution. To do that, we selected 1% of the genes on chromosome 1 (≈ 13) as heritability genes and computed Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves at varying levels of genome-wide heritability. Here we found that BAGHERA correctly identified heritability genes (Fig. 1C), although precision and recall were significantly higher for higher genome-wide heritability levels.

However, our simulated datasets have two limitations. First, they take into account $M \approx 100,000$ SNPs from a single chromosome, whereas more than 1M are routinely genotyped in modern studies. This was a necessary restriction to reduce the time and memory required to simulate genotypes, which is a computationally taxing task. Moreover, fine tuning gene-level heritability is not trivial with genotype data; low and high gene-level heritability enrichments produce either undetectable signal or extremely skewed statistics, while LD patterns might produce spurious signal difficult to control.

We addressed these limitations by developing new model for simulating summary statistics using only linkage disequilibrium information (see Supplementary Materials). 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.

Here we found that BAGHERA correctly identifies heritability genes, even with fold-changes in heritability as low as $f_c = 5$ (see Fig. 1D and supplementary material), although the true positive rate was significantly lower for low heritability levels. Nonetheless, we found BAGHERA to be conservative with a low false discovery rate in all scenarios; this result suggests that our method is suitable for exploratory analyses, and that significant results are due to true biological signal.

2.1.2 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 analyse low heritability traits and to provide heritability estimates at the gene level. Nonetheless, a number of methods have been proposed to estimate genome-wide and local heritability, thus we proceeded to compare our approach to state-of-the-art methods to perform these analyses.

For the genome-wide analysis, we compared our results with LD score regression (LDsc) [5]. Since LDsc is routinely applied to estimate heritability for the traits in the UK Biobank data, we directly retrieved the results for all 38 cancers. Unsurprisingly, since BAGHERA uses a similar genome-wide estimator, we found strong consensus between the estimates of the two methods (Supplementary Materials). Nonetheless, BAGHERA is more robust for low heritability traits, since our Bayesian formulation always provides correct genome-wide heritability estimates, whereas LDsc usually provides negative values.

For local heritability analysis, we compared our performances with the Heritability Estimation from Summary Statistics (HESS) method [45], by using BAGHERA to estimate the heritability of 1703 regions of the original article (See Supplementary Materials). Here we focused on breast and prostate cancer data, since they are those with higher h^2_{SNP} estimates and would not favour either method. We found a statistically significant correlation between the local estimates of BAGHERA and HESS, with all loci reported as significant by HESS also reported by BAGHERA. However, BAGHERA is able to identify more loci with increased heritability, while providing more robust heritability estimates for regions with low h^2_{SNP} (see Supplementary Materials).

Taken together, we have shown that BAGHERA is more robust than existing methods on low heritability traits and can provide useful insights into the disease risk, being able to scale up to $15,000$ different loci across the genome.

2.2 Genome-wide estimates of cancer heritability in the UK Biobank

We used BAGHERA to analyse 38 cancers in the UK Biobank [6], a large-scale prospective study aiming at systematically screening and phenotyping more than $500,000$ individuals, with age ranging between 37 and 73 years.

We obtained summary statistics for $N = 361,194$ individuals ([35], see Table 1), including subjects whose tumours were histologically characterised according to the ICD10 classification, where malignant neoplasms are identified with codes ranging from C00 to C97 (see Supplementary Material). The number of cases varies significantly across cancers, ranging from 102 individuals, for malignant neoplasm of base of tongue and other, to 9086 individual, for malignant neoplasms of the skin. 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 [4].

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, similar to other studies, we expected to have sufficient polygenic signal for our analysis if the average χ^2 was greater than 1 [18]. 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), as these usually comprise tumours of basal and squamous cells, which are mostly caused by sun exposure. Thus, we restricted our analysis to 16 cancers for which we had enough power to perform our analysis, although we also report results for the other cancers in the Supplementary Materials.

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). While comparison between cancer heritability estimates are 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 [22, 41, 31, 28]. The heritability of testicular cancer ($h^2_{SNP_L} = 0.3158$) is the highest among all malignancies, consistent with the hypothesis that germline variants have stronger effects in early onset cancers. However, 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 tumours (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.

2.3 Cancer heritability genes across 16 malignancies

We identified 783 heritability genes ($\eta > 0.99$), harbouring $1,146$ protein-coding genes, across 16 cancers (Fig. 2), with 53 heritability genes per malignancy on average, ranging from 5 genes in mesothelioma, to 271 genes for prostate (see Table 1, Figure 3A). It is worth noting that we are here using the term heritability genes when referring to the genomic, non-overlapping, regions tested by BAGHERA. 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 to 276-fold for the *fibroblast growth factor receptor 2 (FGFR2)* gene in breast cancer. Interestingly, 87% of heritability genes show per-SNP heritability 10-fold higher than the genome-wide estimate. Only 3 genes have fold changes below 5 and more than 99% of genes with fold-changes below 10 are found in the breast and prostate datasets, which have $h^2_{SNP} > 0.01$. Importantly, based on our simulations for datasets with similar heritability enrichment, our set of heritability genes are expected to have a limited number of false positives.

Heritability genes represent less than 1% of all the genes in the genome, but they are significantly more than those harbouring genome-wide significant SNPs (see Supplementary Materials), consistent with cancer being polygenic. Although we identified a polygenic signal, heritability genes 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 (Figure 3A). Consistent with our hypotheses, when we looked at the contribution of SNPs outside our protein-coding regions, we did not observe any difference compared to the genome-wide estimate.

We then tested whether heritability genes 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 genes 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 tumour sequencing studies, which have shown that pleiotropic effects are limited to few master regulators, such as *TP53* [2]. 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, including 4 genes (CLPTM1L, APAF1, THADA, AGBL1) involved in apoptosis and 3 genes (PCDH15, DLG2, POU5F1B) involved in cell division, migration and tumorigenesis [44, 21]. It is important to note that the *cisplatin resistance-related protein 9* (CLPTM1L) is the heritability gene found in most cancers (4) and is one of the gene in the 5p15.33 locus (the other being TERT), which has been consistently associated with 17 different cancer types [39].

Taken together, our analysis found 784, harbouring $1,146$ protein-coding genes, having a significant contribution to the heritable risk of at least 1 cancer. We denoted these $1,146$ loci as cancer heritability genes (CHGs).s

2.4 Cancer heritability genes are recurrently mutated in tumours

Tumour sequencing projects, including the 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 harbouring germline and somatic mutations can mediate cancer phenotypes [38, 60, 47], thus we tested whether cancer heritability genes are significantly enriched among known cancer driver genes. To do that, we built a curated list of driver genes using the COSMIC Cancer Gene Census (Supplementary Table 2). We found that 60 of the $1,146$ CHGs ($\approx 5\%$) are significantly enriched among 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 *Tumour 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 B).

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 ([7]), the Precision Oncology Knowledge Base (OncoKB, [8]), Memorial Sloan Kettering Impact and Heme gene panels [10], and the curated list of cancer genes by Vogelstein et al. [56]. Here we found that CHGs are significantly enriched in each cancer driver gene annotation analysed, with an enrichment ranging from $OR = 1.55$ for the PCAWG annotation to 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 genes 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 tumours.

2.5 Cancer heritability genes underpin biological processes affecting tumorigenesis

Our gene-level heritability analysis identified $1,146$ loci 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 hypothesised that cancer heritability genes could be involved in molecular functions and biological processes affecting tumorigenesis.

To do that, we characterized CHGs by gene ontology enrichment analysis, using the slim gene ontology for human (Supplementary Table 1). We found a statistically significant enrichment for 21 terms (Fisher's exact test; False Discovery Rate, FDR $< 10\%$, Figure 4C and Supplementary Table 1), 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 ($OR : 1.43,\ P : 2.40 \times 10^{-3}$), differentiation ($OR : 1.20,\ P : 7.7 \times 10^{-3}$), which includes the *mammalian target of rapamycin (mTOR)* (*MTOR*) gene, growth ($OR : 1.55,\ P : 2.62 \times 10^{-4}$) and cell proliferation ($OR : 1.3,\ P : 3.4 \times 10^{-3}$), which includes the *Poly [ADP-ribose] polymerase 1* (*PARP1*) gene. We also observed a significant enrichment of genes associated with cytoskeleton organization ($OR : 1.40,\ P : 2.39 \times 10^{-2}$) and anatomical structure development ($OR : 1.32,\ P : 6.13 \times 10^{-3}$), which includes members of the SWI/SNF complex, such as the *AT-Rich Interaction Domain 2* (*ARID2*).

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 [48], 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 [20]. These lists of biological processes include proliferative signalling, suppression of growth, escaping immunic 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 ($OR : 2.062,\ P : 3{\times}10^{-4}$). Consistent with our previous analysis, cancer heritability genes are involved in escaping cell death, mediating proliferative signalling, invasion and metastasis (Figure 4D and Supplementary Table 3). We then went further to study whether CHGs mediate these cancer processes by acting either as tumour suppressor genes (TSGs) or oncogenes (see Fig. 4E). To do that, we used the Precision Oncology Knowledge Base (OncoKB, [8]), 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, $P : 7.9{\times}10^{-6}$), 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). Tumour suppressor CHGs include well-known cancer driver genes, such as *CDKN2A* and *MTOR* which regulate cell growth, and DNA repair genes, such as *MUTYH* and *FANCA* [25].

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 [24]; while inherited mutations associated with oncogene activation are likely to be under purifying selection, mutations in tumour suppressor genes can be observed at higher frequency because deleterious effects are only observed upon complete loss of function.

3 Discussion

Our study provides new fundamental evidence demonstrating a strong contribution of high-frequency inherited mutations to the heritable risk of cancer. Here we provide a high resolution map of the heritable cancer genome consisting of $1,146$ genes showing a significant contribution to the heritable risk of 16 malignancies. We showed that these loci harbour tumour suppressors controlling growth, cell morphogenesis and proliferation, which are fundamental processes required for tumorigenesis.

Ultimately, our results support a two-hit model, where inherited mutations in tumour suppressor genes could create a favourable 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. Preliminary support for this model has been provided by studies in hereditary diffuse gastric cancer (HDGC) [33] and, more recently, by germline variant burden analyses [38].

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. 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 analyse 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, which limits our power of detecting tissue-specific contributions. On this point, as the genes may be expressed in different cellular compartments, they may contribute to the stromal niches in which cancers develop and their role in tissue specificity of mutations will be of interest to analyse experimentally.

Taken together, our study provides a new view of the genetic architecture of cancer with gene-level resolution. We anticipate that the availability of genome editing techniques will enable testing of the functional mechanisms mediated by cancer heritability genes. We also expect that integrating our results with tumour sequencing data will provide new venues for personalized treatment and patients' stratification.

Notes.

G.S. and V.F. conceived the study. V.F., G.S. and L.C. designed the model. V.F. wrote the software and performed all analyses, supervised by G.S. . G.S., A.L.H. and F.P. analysed the functional roles of cancer heritability genes. G.S. and V.F. wrote the manuscript with contributions from all authors. The authors also declare no competing interests. The results of our analyses have been deposited in CSV format on Zenodo at: https://doi.org/10.5281/zenodo.3968269.

Acknowledgments.

The authors would like to thank Dr Yongjin Park at the MIT for the useful discussions about GWAS data simulation.

ICD10	Malignancy	Cases	Prevalence	$\hat{\chi}^2$	h^2_{SNP}	$h^2_{SNP_L}$	HG
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 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 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 genes (HG) reported by BAGHERA as significant for $\eta > 0.99$. In bold, we denote the cancers used to build the cancer heritability genes (CHGs) panel.

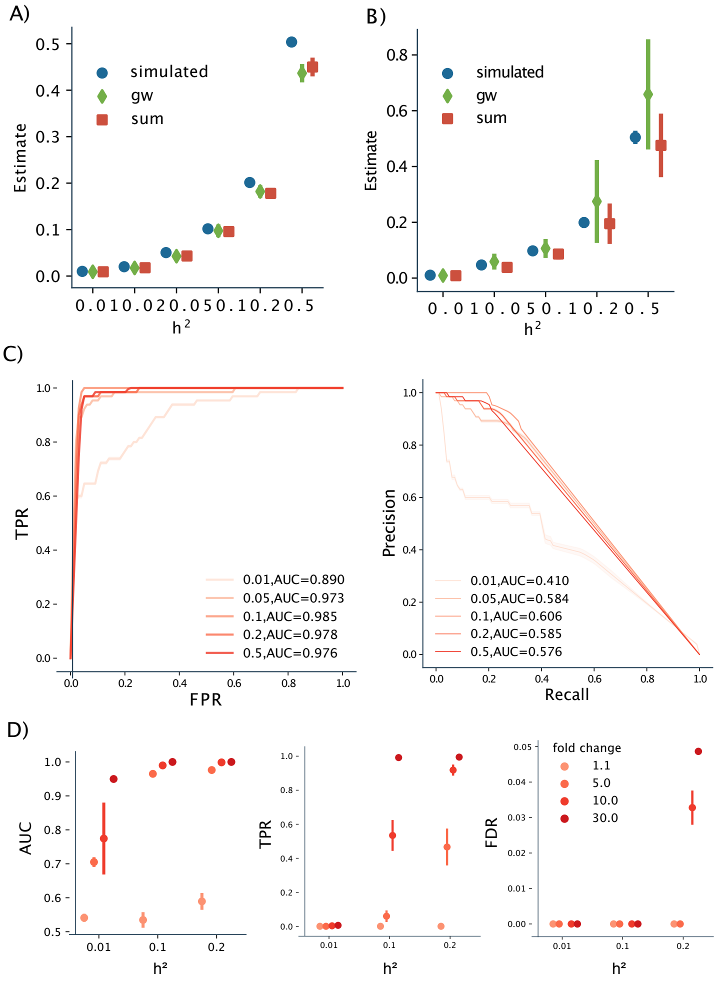

Figure 1: **Performance on simulated data.** A) Genome-wide heritability estimates for dense effects. For each value of h^2, we plot the simulated value, the target, the genome-wide (gw) estimate, which is the median of the posterior of genome-wide heritability term, and the gene-level estimate which is the sum of all median gene heritability estimate (sum). B) Genome-wide heritability estimates for datasets with varying gene-level heritability. For each value of h^2 we plot the simulated value, the target, the gw estimate which is the median of the prior heritability term, and the gene-level estimate which is the sum of all median gene heritability estimate C) Receiver Operator Characteristic curves and Precision Recall curves, for the performance of BAGHERA at retrieving positive genes for different values of genome-wide h^2. D) Performance of BAGHERA for different values of h^2 and gene level enrichment. We show the AUCs of the ROC curves, the True Positive Rate and False Discovery Rate (FDR) for $\eta > 0.99$. A-B-C show the performance on simulated genotype data on chromosome 1. D is showing the performance on the data simulated from summary statistics.

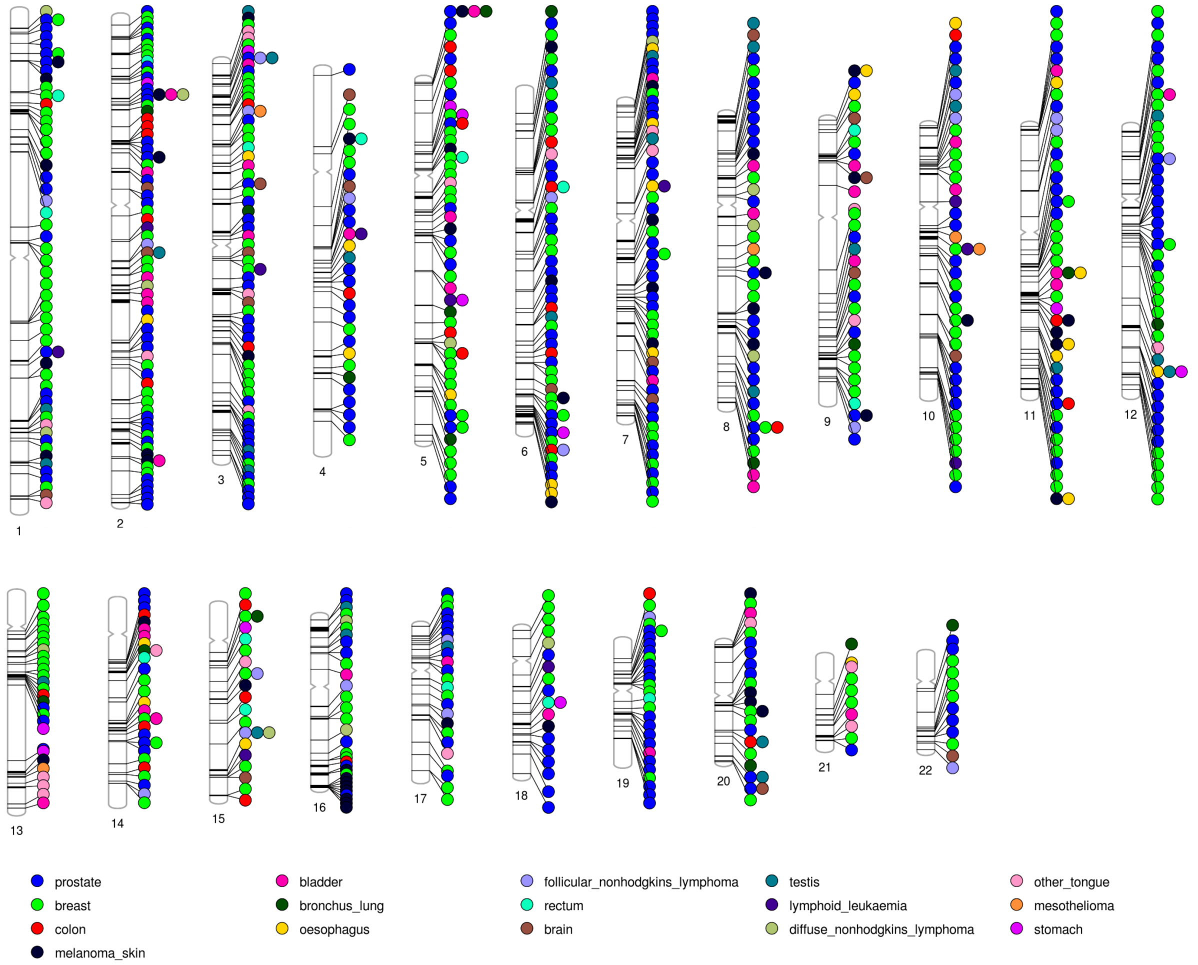

Figure 2: **Cancer heritability genes across the human genome.** For each cancer heritability gene, we report its locus and the associated cancer.

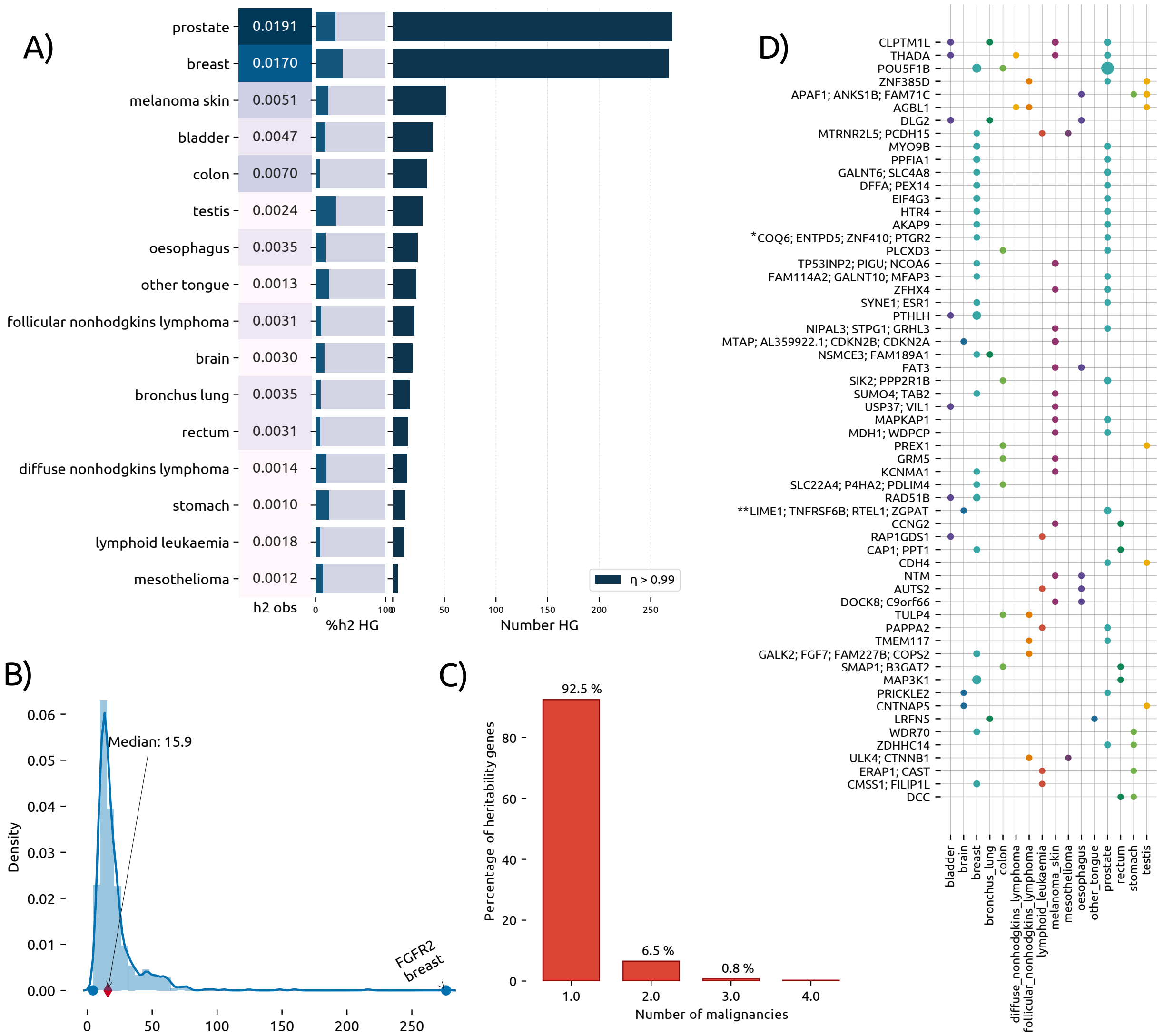

Figure 3: **Heritability genes across 16 cancers in the UK Biobank** A) For each malignancy, we report the observed heritability (h^2_{SNP}, left box), the percentage of h^2_{SNP} explained by heritability genes (central barplot, dark blue is the percentage explained by HGs) and the number of heritability genes (right barplot). B) Gene-level heritability density distribution across heritability genes, expressed as fold-change with respect to the genome-wide estimate. Highlighted are the top genes and the median fold-change across all cancers. C) Percentage of cancer heritability genes associated with multiple cancers. Approximately 8% of HGs are common to multiple malignancies. D) Cancer heritability genes associated with multiple cancers. We report the 59 HGs common to at least 2 cancers; here the size of the dot is proportional to the heritability enrichment of the gene in the specific cancer.

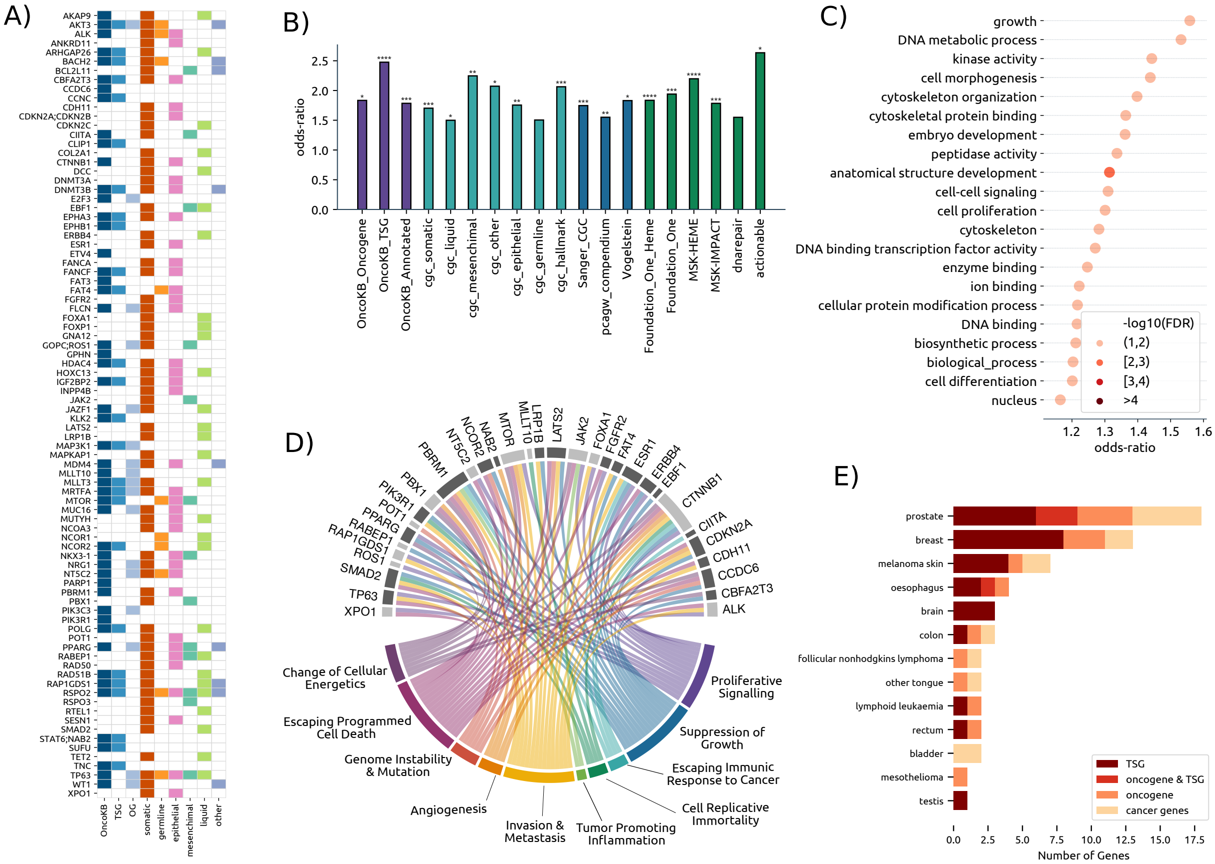

Figure 4: **Functional characterisation of cancer heritability genes.** A) List of CHGs reported as cancer driver genes across multiple annotations. With the blue hue, first three columns, we report the genes annotated by OncoKB, specifying whether they are tumour suppressors (TSG) or oncogenes (OG). 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 $-log_{10}$(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. E) tumour suppressor and oncogene CHGs across cancers. For each cancer type (y-axis), we report the number of genes (x-axis) reported as tumour suppressors (TSGs) and/or oncogenes in OncoKB (colour codes).

References

[1] David E. Anderson. "Genetic study of breast cancer: Identification of a high risk group". In: *Cancer* 34.4 (1974), pp. 1090–1097. ISSN: 0008-543X. DOI: 10.1002/1097-0142(197410)34:4<1090::AID-CNCR2820340419>3.0.CO;2-J. URL: http://doi.wiley.com/10.1002/1097-0142{\%}28197410{\%}2934{\%}3A4{\%}3C1090{\%}3A{\%}3AAID-CNCR2820340419{\%}3E3.0.CO{\%}3B2-J.

[2] Matthew H Bailey et al. "Comprehensive Characterization of Cancer Driver Genes and Mutations." In: *Cell* 173.2 (2018), 371–385.e18. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.02.060. URL: http://www.ncbi.nlm.nih.gov/pubmed/29625053http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6029450.

[3] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186. ISSN: 10974172. DOI: 10.1016/j.cell.2017.05.038. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867417306293.

[4] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." In: *CA: a cancer journal for clinicians* (2018). ISSN: 1542-4863. DOI: 10.3322/caac.21492. URL: http://www.ncbi.nlm.nih.gov/pubmed/30207593.

[5] Brendan K Bulik-Sullivan et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature genetics* 47.3 (2015), p. 291.

[6] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0579-z. URL: http://www.nature.com/articles/s41586-018-0579-z.

[7] Peter J. Campbell et al. "Pan-cancer analysis of whole genomes". In: *Nature* 578.7793 (2020), pp. 82–93. ISSN: 14764687. DOI: 10.1038/s41586-020-1969-6.

[8] Debyani Chakravarty et al. "OncoKB: a precision oncology knowledge base". In: *JCO precision oncology* 1 (2017), pp. 1–16.

[9] Dan Chen et al. "Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability". In: *Carcinogenesis* 36.9 (2015), pp. 992–998. ISSN: 0143-3334. DOI: 10.1093/carcin/bgv083. URL: https://academic.oup.com/carcin/article-lookup/doi/10.1093/carcin/bgv083.

[10] Donavan T Cheng 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". In: *The Journal of molecular diagnostics* 17.3 (2015), pp. 251–264.

[11] W. Cozen et al. "A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus". In: *Nature Communications* 5.1 (2014), p. 3856. ISSN: 2041-1723. DOI: 10.1038/ncomms4856. URL: http://www.nature.com/articles/ncomms4856.

[12] Maria Chiara Di Bernardo et al. "A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia". In: *Nature Genetics* 40.10 (2008), pp. 1204–1210. ISSN: 1061-4036. DOI: 10.1038/ng.219. URL: http://www.nature.com/articles/ng.219.

[13] Douglas F. Easton et al. "Genome-wide association study identifies novel breast cancer susceptibility loci". In: *Nature* 447.7148 (2007), pp. 1087–1093. ISSN: 0028-0836. DOI: 10.1038/nature05887. URL: http://www.nature.com/doifinder/10.1038/nature05887.

[14] Rosalind A Eeles et al. “Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array”. In: *Nature Genetics* 45.4 (2013), pp. 385–391. ISSN: 1061-4036. DOI: 10.1038/ng.2560. URL: http://www.nature.com/articles/ng.2560.

[15] Victor Enciso-Mora et al. “A genome-wide association study of Hodgkin’s lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3)”. In: *Nature Genetics* 42.12 (2010), pp. 1126–1130. ISSN: 1061-4036. DOI: 10.1038/ng.696. URL: http://www.nature.com/articles/ng.696.

[16] Viola Fanfani et al. “Dissecting the heritable risk of breast cancer: from statistical methods to susceptibility genes”. In: *Seminars in cancer biology* (2020). ISSN: 10963650. DOI: 10.1016/j.semcancer.2020.06.001.

[17] Hilary K. Finucane et al. “Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types”. In: *Nature Genetics* 50.4 (2018), pp. 621–629. ISSN: 15461718. DOI: 10.1038/s41588-018-0081-4. URL: http://dx.doi.org/10.1038/s41588-018-0081-4.

[18] Hilary K Finucane et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature Genetics* 47.11 (2015), pp. 1228–1235. ISSN: 1061-4036. DOI: 10.1038/ng.3404. URL: http://www.nature.com/articles/ng.3404.

[19] Maya Ghoussaini et al. “Genome-wide association analysis identifies three new breast cancer susceptibility loci”. In: *Nature Genetics* 44.3 (2012), pp. 312–318. ISSN: 1061-4036. DOI: 10.1038/ng.1049. URL: http://www.nature.com/articles/ng.1049.

[20] Douglas Hanahan and Robert A Weinberg. “Hallmarks of cancer: the next generation”. In: *cell* 144.5 (2011), pp. 646–674.

[21] H Hayashi et al. “The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer”. In: *Oncogene* 34.2 (2015), pp. 199–208.

[22] Xia Jiang et al. “Shared heritability and functional enrichment across six solid cancers”. In: *Nature Communications* 10.1 (2019), p. 431. ISSN: 2041-1723. DOI: 10.1038/s41467-018-08054-4. URL: http://www.nature.com/articles/s41467-018-08054-4.

[23] A. E. Kennedy, U. Ozbek, and M. T. Dorak. “What has GWAS done for HLA and disease associations?” In: *International Journal of Immunogenetics* 44.5 (2017), pp. 195–211. ISSN: 1744313X. DOI: 10.1111/iji.12332. URL: http://doi.wiley.com/10.1111/iji.12332.

[24] Alfred G Knudson. “Mutation and cancer: statistical study of retinoblastoma”. In: *Proceedings of the National Academy of Sciences* 68.4 (1971), pp. 820–823.

[25] Sabine S Lange, Kei-ichi Takata, and Richard D Wood. “DNA polymerases and cancer”. In: *Nature reviews cancer* 11.2 (2011), p. 96.

[26] Philip J. Law et al. “Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia”. In: *Nature Communications* 8 (2017), p. 14175. ISSN: 2041-1723. DOI: 10.1038/ncomms14175. URL: http://www.nature.com/doifinder/10.1038/ncomms14175.

[27] Kevin Litchfield et al. “Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor”. In: *Nature Genetics* 49.7 (2017), pp. 1133–1140. ISSN: 15461718. DOI: 10.1038/ng.3896.

[28] Kevin Litchfield et al. "Quantifying the heritability of testicular germ cell tumour using both population-based and genomic approaches". In: *Scientific Reports* 5.1 (2015), p. 13889. ISSN: 2045-2322. DOI: 10.1038/srep13889. URL: http://www.nature.com/articles/srep13889.

[29] Kevin Litchfield et al. "The genomic landscape of testicular germ cell tumours: From susceptibility to treatment". In: *Nature Reviews Urology* 13.7 (2016), pp. 409–419. ISSN: 17594820. DOI: 10.1038/nrurol.2016.107.

[30] D Malkin et al. "Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms." In: *Science (New York, N.Y.)* 250.4985 (1990), pp. 1233–8. ISSN: 0036-8075. URL: http://www.ncbi.nlm.nih.gov/pubmed/1978757.

[31] Nicholas Mancuso et al. "The contribution of rare variation to prostate cancer heritability". In: *Nature Genetics* 48.1 (2016), pp. 30–35. ISSN: 1061-4036. DOI: 10.1038/ng.3446. URL: http://www.nature.com/articles/ng.3446.

[32] Teri A. Manolio et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (2009), pp. 747–753. ISSN: 00280836. DOI: 10.1038/nature08494. URL: http://dx.doi.org/10.1038/nature08494.

[33] Alisha M Mendonsa, Tae-Young Na, and Barry M Gumbiner. "E-cadherin in contact inhibition and cancer". In: *Oncogene* (2018), p. 1.

[34] Y Miki et al. "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1". In: *Science* 266.5182 (1994), pp. 66–71. ISSN: 0036-8075. DOI: 10.1126/science.7545954. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.7545954.

[35] Benjamin Neale. *UKBB Heritability*. URL: https://nealelab.github.io/UKBB_ldsc/.

[36] L Orsi et al. "Genetic polymorphisms and childhood acute lymphoblastic leukemia: GWAS of the ESCALE study (SFCE)". In: *Leukemia* 26.12 (2012), pp. 2561–2564. ISSN: 0887-6924. DOI: 10.1038/leu.2012.148. URL: http://www.nature.com/articles/leu2012148.

[37] Bogdan Pasaniuc and Alkes L. Price. "Dissecting the genetics of complex traits using summary association statistics". In: *Nature Reviews Genetics* 18.2 (2017), pp. 117–127. ISSN: 14710064. DOI: 10.1038/nrg.2016.142.

[38] Tao Qing et al. "Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden". In: *Nature Communications* 11.1 (2020), pp. 1–8.

[39] Thorunn Rafnar et al. "Sequence variants at the TERT-CLPTM1L locus associate with many cancer types". In: *Nature genetics* 41.2 (2009), pp. 221–227.

[40] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016), e55.

[41] Joshua N Sampson et al. "Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types." In: *Journal of the National Cancer Institute* 107.12 (2015), djv279. ISSN: 1460-2105. DOI: 10.1093/jnci/djv279. URL: http://www.ncbi.nlm.nih.gov/pubmed/26464424http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4806328.

[42] Yadav Sapkota. "Germline DNA variations in breast cancer predisposition and prognosis: a systematic review of the literature." In: *Cytogenetic and genome research* 144.2 (2014), pp. 77–91. ISSN: 1424-859X. DOI: 10.1159/000369045. URL: http://www.ncbi.nlm.nih.gov/pubmed/25401968.

[43] Andrew J Schork et al. "All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs". In: *PLoS genetics* 9.4 (2013), e1003449.

[44] Yang W Shao et al. "Cross-species genomics identifies DLG2 as a tumor suppressor in osteosarcoma". In: *Oncogene* 38.2 (2019), pp. 291–298.

[45] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. "Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data". In: *American Journal of Human Genetics* 99.1 (2016), pp. 139–153. ISSN: 15376605. DOI: 10.1016/j.ajhg.2016.05.013. URL: http://dx.doi.org/10.1016/j.ajhg.2016.05.013.

[46] Helen E Speedy et al. "A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia". In: *Nature Genetics* 46.1 (2014), pp. 56–60. ISSN: 1061-4036. DOI: 10.1038/ng.2843. URL: http://www.nature.com/articles/ng.2843.

[47] Giovanni Stracquadanio et al. "The importance of p53 pathway genetics in inherited and somatic cancer genomes". In: *Nature reviews Cancer* 16.4 (2016), p. 251.

[48] Christina H. Stuelten, Carole A. Parent, and Denise J. Montell. "Cell motility in cancer invasion and metastasis: Insights from simple model organisms". In: *Nature Reviews Cancer* 18.5 (2018), pp. 296–312. ISSN: 14741768. DOI: 10.1038/nrc.2018.15. URL: http://dx.doi.org/10.1038/nrc.2018.15.

[49] Amit Sud, Ben Kinnersley, and Richard S. Houlston. "Genome-wide association studies of cancer: current insights and future perspectives". In: *Nature reviews. Cancer* 17.11 (Nov. 2017), pp. 692–704. ISSN: 14741768. DOI: 10.1038/nrc.2017.82. URL: https://www.nature.com/articles/nrc.2017.82.

[50] Peter H Sudmant et al. "An integrated map of structural variation in 2,504 human genomes". In: *Nature* 526.7571 (2015), p. 75.

[51] Gilles Thomas et al. "Multiple loci identified in a genome-wide association study of prostate cancer". In: *Nature Genetics* 40.3 (2008), pp. 310–315. ISSN: 1061-4036. DOI: 10.1038/ng.91. URL: http://www.nature.com/articles/ng.91.

[52] Clare Turnbull et al. "Genome-wide association study identifies five new breast cancer susceptibility loci". In: *Nature Genetics* 42.6 (2010), pp. 504–507. ISSN: 1061-4036. DOI: 10.1038/ng.586. URL: http://www.nature.com/articles/ng.586.

[53] Jayaram Vijayakrishnan et al. "Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia". In: *Nature Communications* 9.1 (2018), p. 1340. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03178-z. URL: http://www.nature.com/articles/s41467-018-03178-z.

[54] Peter M. Visscher, William G. Hill, and Naomi R. Wray. "Heritability in the genomics era - Concepts and misconceptions". In: *Nature Reviews Genetics* 9.4 (2008), pp. 255–266. ISSN: 14710056. DOI: 10.1038/nrg2322. arXiv: arXiv:1011.1669v3. URL: http://www.nature.com/doifinder/10.1038/nrg2322.

[55] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22. ISSN: 00029297. DOI: 10.1016/j.ajhg.2017.06.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409.

[56] Bert Vogelstein et al. "Cancer genome landscapes". In: *science* 339.6127 (2013), pp. 1546–1558.

[57] Zhaoming Wang et al. “Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor”. In: *Nature Genetics* 49.7 (2017), pp. 1141–1147. ISSN: 1061-4036. DOI: 10.1038/ng.3879. URL: http://www.nature.com/doifinder/10.1038/ng.3879.

[58] R Wooster et al. “Identification of the breast cancer susceptibility gene BRCA2.” In: *Nature* 378.6559 (1995), pp. 789–92. ISSN: 0028-0836. DOI: 10.1038/378789a0. URL: http://www.ncbi.nlm.nih.gov/pubmed/8524414.

[59] Jian Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[60] Ping Zhang et al. “Germline and somatic genetic variants in the p53 pathway interact to affect cancer risk, progression and drug response”. In: *bioRxiv* (2019), p. 835918.

4 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 $500,000$ single nucleotide polymorphisms (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 portion of h^2_{SNP} explained by each protein-coding gene, which requires a unique assignment of SNPs to genes to avoid biased estimates.

We denote as genome-wide 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 simply $\hbar^2 = h^2_{SNP}/M$. 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 $\hbar^2_M = \mathbf{E}\left[\hbar^2_j\right]_{j=1,\cdots,M}$, 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 $h^2_{SNP}/M = \mathrm{E}\left[\hbar^2_{M_k}\right]_K$. 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 nuisance gene, 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:

$$\begin{array}{rcl} h^2_{SNP} & \sim & \mathcal{F}_1() \quad \text{with} \quad \operatorname{supp}(\mathcal{F}_1()) \in [0,1] \\ h^2_k | h^2_{SNP} & \sim & \mathcal{F}_2(h^2_{SNP}) \\ \mathrm{Y}|h^2_k & \sim & \mathcal{F}_3(h^2_k) \end{array} \tag{1}$$

where $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ 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 $\Gamma(\alpha,\theta)/(\Gamma(\alpha,\theta) + \Gamma(\beta,\theta)) \sim$ Beta(α, β), a suitable distribution for F_1,in Eq. 1, would be an uninformative Beta distribution, e.g. Beta$(1, 1)$. In practice, the use of a Beta distribution as prior for h^2_{SNP} allow us to obtain accurate estimates of heritability in the unit range even for low-heritability diseases, where classical methods are usually unreliable [5].

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 $\beta = 1$. Therefore, for $\mathcal{F}_2 = \mathrm{Gamma}(h^2_{SNP}, 1)$, 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 [59]; however, the vast majority requires genotype data, which are both difficult to obtain, due to privacy concerns, and computationally taxing to analyse, because of high-dimensionality. Thus, we adopted the LD-score (LDsc) regression model [5], which allows estimation of heritability from GWAS summary statistics, such as regression coefficients and standard errors, which are readily available [37].

Thus, for $\mathcal{F}_3$, we rewrote the LDsc model to estimate gene-level heritability, from summary statistics of M SNPs in a GWAS with N subjects, as follows:

$$\chi^2_{j,k} \sim N(Nl_j h_k^2/M + e, \sqrt{l_j}) \tag{2}$$

where $\chi^2_{j,k}$ 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 $l_j = \sum_z r^2_{jz}$, where r^2_{jz} is the linkage disequilibrium between variant j and variant z in a given population [50]. Moreover, setting the standard deviation to the LD score of the j-th SNP allow us to control for heteroskedasticity of the test statistics due to linkage disequilibrium, somehow similar to the weighting scheme used in LDsc. The e term accounts for confounding biases and it 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, and predicts heritability genes, that are genes with a per-SNP heritability higher than the genome-wide estimate (see Supplementary Figure 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^2_{SNP}, 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.

$$\begin{array}{rcl} e & \sim & \mathcal{N}(1,1) \\ W & \sim & \text{Inv-Gamma}(1,1) \\ h^2_{SNP} & \sim & \text{Beta}(1,1) \\ h^2_k | h^2_{SNP}, W & \sim & \mathcal{N}(h^2_{SNP}, W^2) \\ \chi^2_{j,k} | h^2_k, e, l_j, N, M & \sim & \mathcal{N}(Nl_j h^2_k/M + e, \sqrt{l_j}) \end{array} \tag{3}$$

BAGHERA predicts heritability genes by computing the posterior distribution of $\eta_k \sim I(h^2_k > h^2_{SNP})$, 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, $\mathbf{E}[\eta_k]$, is the probability of the heritability of gene k of being higher than the genome-wide estimate; specifically, we report as heritability genes, those with $\mathbf{E}[\eta_k] \geq 0.99$. For each gene, we also report effect sizes in terms of fold-change with respect to the genome-wide heritability estimate, $fc_k = h^2_k / h^2_{SNP}$.

We use the No-U-Turn Sampler as implemented in PyMC 3.4 [40], 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 installable 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).

UK BioBank summary statistics processing and curation

We used summary statistics of the UK BioBank GWAS for ICD10 classified cancer types [35]. We developed a custom pipeline to assign LD scores to SNPs, and SNPs to human genes (Supplementary Material). For each dataset, we mapped each SNP to a precomputed LD score. 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 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 [5, 23]. Overall, our analysis is conducted on 1285620 SNPs over 22 chromosomes.

We then used Gencode v31 to determine the genomic coordinates of protein coding genes in the GRCh37 human genome. We then merged overlapping genes by creating a new pseudo-gene, whose name reports the merged gene names and whose boundaries are defined as the first and last base-pair of the overlapping genes. We assigned a SNP to a gene if it is within ± 50kb from the gene boundaries, which allow us to account for cis-regulatory elements, overall 55% of SNPs were mapped to a gene. For BAGHERA we considered only those genes harboring at least 10 variants. Our dataset consists of 15025 genes, 12042 of them are harboring more than 10 SNPs, then they are used for the UKBB analysis.

4.1 Enrichment analyses

We used a one-tailed Fisher's exact test for all enrichment analyses, with p-value adjusted using the Benjamini-Hochberg procedure. Since genes in our analysis might represent overlapping protein-coding regions, we post-processed our gene lists by converting each composed region into the set of its genes for functional characterization and annotation. The overlap with cancer datasets has instead been tested with individual one-tailed Fisher's exact tests.