# Supplementary Materials: The landscape of the heritable cancer genome

Viola Fanfani[1], Luca Citi[2], Adrian L. Harris[3], Francesco Pezzella[4], and Giovanni Stracquadanio[1,5]

[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*
[5] *Corresponding author. Phone: +44 (0) 131 6507193, Email: giovanni.stracquadanio@ed.ac.uk.*

## Contents

# 1 Supplementary Methods

## 1.1 Simulated datasets

We performed extensive simulations to assess the performance of our hierarchical Bayesian model, as implemented in BAGHERA.

First, we generated datasets with a realistic genetic architecture and linkage disequilibrium patterns using data from the 1000 Genomes Project (see Supplementary Methods 1.1.1). Since these simulations are computationally taxing and existing tools do not scale for genome-wide simulations, we restricted our analyses to SNPs located on chromosome 1. We used these datasets to test the accuracy of the genome-wide heritability estimates returned by BAGHERA, and its performances for gene-level heritability analysis.

Nonetheless, we also wanted to explore the performance of our method on whole genome datasets, which is the common use case for our method. Thus, we simulated whole genome summary statistics with a varying number of heritability loci and enrichments (see Supplementary Methods 1.1.2).

When assessing the performance of BAGHERA in detecting heritability loci. We remind the reader that our model estimates the posterior distribution of $\eta_k$, whose value is the probability of the per-SNP heritability of gene $k$ to be higher than the per-SNP genome-wide estimate; thus, we can test how many heritability loci are discovered as a function of $\eta_k$. Since heritability loci are known a-priori in our simulations, we derived Receiver Operating Characteristic (ROC) curves and computed the corresponding Area Under the Curve (AUC) for each type of simulation. While ROC curves allow straightforward comparison of different experimental conditions, they can be problematic for interpreting genomic data, since the number of positive samples is significantly smaller than the negatives. For this reason, we also derived Precision and Recall (PR) curves as a more accurate approach to control Type 1 errors.

Hereby, we describe the procedures implemented to generation our simulated datasets and the main results of the simulation analysis.

### 1.1.1 Simulated datasets with a realistic genetic architecture

We simulated N = $50,000$ subjects and M = $100,000$ SNPs on chromosome 1 from 1000 Genome reference data from $503$ European ancestry subjects, using HAPGEN2 [4] and haplotype data downloaded from the IMPUTE website (`https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#download`). We then filtered out SNPs with minor allele frequency (MAF) smaller than 0.01, leading to a final dataset consisting of 99,586 SNPs.

We then controlled whether the simulated genetic architecture was coherent with the one observed in Europeans. To do that, we estimated the correlation between the observed MAF in the 1000 Genomes data and our simulated data; here we found a statistically significant correlation between the two datasets (Pearson correlation coefficient $\rho = 0.9929$, $P \leq 10^{-5}$), suggesting that our strategy was appropriate to generate a realistic genetic architecture.

Summary statistics were then simulated following a dense and gene-level effect size model. First, we used the dense effect model to test the robustness of the genome-wide heritability estimates. To do that, we explicitly set the variance of the SNPs to be $\hbar^2 = h^2/M$, with $h^2 = \{0.01, 0.1, 0.2, 0.5\}$; for each parameter setting, we generated $5$ different datasets. BAGHERA correctly estimates genome-wide heritability both as the median of genome-wide term $h^2_{SNP}$ and as the sum of the contributions of all genes (see Supplementary Figure 1). Performance drops for larger $h^2$ values, which are outside the working conditions of our method.

We then assessed BAGHERA as a method for discovering heritability loci. To do that, we set as causal only those SNPs that are located in a predefined set of loci. With this setting, we tested whether BAGHERA was able to identify heritability loci under different genome-wide and

local heritability levels. Out of all loci $L$, we selected a fraction of them, $s_L$, as significant, with $L_{sig} = L \times s_L$ being the total amount of significant loci. We then assigned $90\%$ of the variance to the $M_{sig}$ SNPs falling into the $L_{sig}$ loci, while the remaining $10\%$ variance is equally distributed to the other loci. We simulated data with $h^2 = \{0.01, 0.05, 0.1, 0.2\}$ and $s_L = 0.01$ $(1\%)$; taken together, we obtained $L_{sig} = 13$ heritability loci out of $1322$ loci with more than $10$ SNPs on chromosome 1. For each parameters combination, we simulated $5$ datasets. Here we found BAGHERA to provide accurate $h^2_{SNP}$ estimates, both as the median of the posterior of the $h^2_{SNP}$ term and the sum of the gene level heritability (see Supplementary Figure 2A). Similar to the results for dense-effect simulations, performance is more unstable for larger values of heritability. However, in the worst case scenario, $h^2_{SNP}$ tends to be overestimated, which leads towards more conservative statistical testing. Importantly, BAGHERA performs extremely well in retrieving significant loci with AUCs above $90\%$ for ROC analysis and above $50\%$ for most PR analysis (see Supplementary Figure 2B and C).

### 1.1.2 Whole genome simulated datasets

Restricting the analysis to chromosome 1 would not provide conclusive evidence about the performances of our method, which was designed to run on high-density genotype data. We then used a simpler model, which does not require genotype data, to generate simulated summary statistics for 22 chromosomes with a varying number of heritability loci and levels of heritability enrichment.

We assigned random effect sizes to SNPs with MAF $> 0.01$ in the European populations of the 1000 Genomes Phase 3 project by sampling from a normal distribution and weighting the random variate by $w_j = \sqrt{(1 + \frac{N}{M} h^2_k l_j)}$, where $h^2_k$ is the gene-level heritability and $l_j$ is the LD score of the $j$-th SNP in the dataset [1]. Using LD scores allow us to account for positional constraints and LD patterns without using genotype data. We then randomly selected a fraction of loci as heritability loci and set their heritability $h^2_k = fc_k \times h^2_{SNP}$, where $h^2_{SNP}$ is the genome-wide heritability, $fc_k$ is the fold-change in heritability in the locus $k$ compared to the genome-wide estimate.

In our experiments, we set the genome-wide heritability to $h^2_{SNP} = \{0.01, 0.1, 0.2\}$, to mimic a disease with a reasonably low heritability, such as cancer. We then considered $p = 1\%$ of the loci in the genome as heritability loci, and set the heritability fold-change as $fc_k = \{1.1, 5, 10, 30\}$, while fold-change value $fc = 1.1$ is used as control. For each possible parameter setting, we generated $3$ independent datasets, which resulted in a testbed consisting of $36$ datasets in total.

Our model obtained excellent results for fold-changes ranging from $5$ to $30$, when the genome wide heritability is at least $0.1$. While ROC performance drops for $5$ and $10$ fold-change for low heritability levels, TPR and FDR estimates prove that our testing procedure is actually conservative (see Supplementary Figure 3) and that our model has $FDR < 0.05$. Finally, for the control simulations $fc = 1.1$, as expected, the ROC and PR analyses show no significant difference with respect to a random classifier (see Supplementary Figures 3, 4, and 5).

It is worth noting that the ROC curves in Supplementary Figures 4 are the detail of the ROC AUC shown in Supplementary Figure 3.

## 1.2 Comparison with state-of-the-art methods

### 1.2.1 Comparison of genome-wide heritability estimates between BAGHERA and LDsc

We compared BAGHERA genome-wide estimates with the observed $h^2_{SNP}$ estimates of LD score regression (LDsc) [2]. It is straightforward to note that BAGHERA and LDsc estimates

follow a similar trend, although BAGHERA is more robust on low heritability malignancies, including $9$ cases where LDsc erroneously reported negative estimates (see Supplementary Figure 6).

### 1.2.2 Comparison of local heritability estimates between BAGHERA and HESS

We compared our estimates of local heritability with those obtained by HESS [3], which, to date, is the only method for the estimation of local heritability using summary statistics and can be applied on regions smaller than a chromosome.

First, we outline the main differences between the two methods, which could confuse the interpretation of the results. HESS has been shown to provide robust heritability estimates for genomic regions defined as LD independent. BAGHERA, instead, provides heritability estimates for any non overlapping set of genomic regions, including $\approx 15,000$ protein-coding genes in the human genome. Thus, BAGHERA can provide heritability estimates at a much higher genomic resolution.

It is also important to also note the different output returned by BAGHERA and HESS. We remind the reader that each region explains a portion of heritability $\ddot{h}^2_k = \sum_{j=1}^{M_k} \hbar^2_j$, where $\ddot{h}^2_k$ is the output of HESS. With the notation we introduced in our study, $\ddot{h}^2_k/M_k = h^2_k/M$, where $h^2_k$ is the gene-level heritability estimated by BAGHERA. Both methods, however, test whether the local single SNP heritability, either $h^2_k/M$ or $\ddot{h}^2_k/M_k$, is larger than the expected genome-wide heritability $\hbar^2_M$.

It is also worth mentioning that the two methods implement different testing strategies; after the estimation of local heritability, HESS converts the estimates to z-scores to obtain a p-value for each region, and then uses Bonferroni correction to control the family-wise error rate. BAGHERA instead uses a Bayesian hierarchical model to estimate the posterior distribution of the genome-wide and gene-level heritability, along with the posterior distribution of the indicator function, $\eta$, which is used to estimate the probability of the per-SNP heritability of gene $k$ to be higher than genome-wide estimate.

We then applied HESS and BAGHERA on the two cancer datasets from the UK Biobank with the highest heritability: breast (C50) and prostate (C61). In order to compare local heritability estimates of the two methods, we used the same set of SNPs and the $1703$ regions originally used by HESS, although we filtered out $10$ of them having less than $10$ SNPs. For each cancer (ICD10 code), we computed the genome-wide estimates $h^2$, the number of significant genomic loci, the number of significant loci found both by HESS and BAGHERA, the correlation between the local heritability estimates (Pearson's $\rho$) and the corresponding p-value (see table below).

| | HESS | | BAGHERA | | | | |
|---|---|---|---|---|---|---|---|
| ICD10 | $h^2(se)$ | Significant loci | $h^2(sd)$ | Significant loci | Common loci | $\rho$ | p-value |
| C50 | $0.0111$ $(0.00316)$ | 2 | $0.0149$ $(0.0018)$ | 119 | 2 | $0.78$ | $\leq 10^{-6}$ |
| C61 | $0.00896$ $(0.00316)$ | 1 | $0.0098$ $(0.0017)$ | 116 | 1 | $0.76$ | $\leq 10^{-6}$ |

Experimental results showed a strong consensus between the genome-wide heritability estimates of both methods, whereas BAGHERA the largest number of heritability loci, including the two found by HESS. In Supplementary Figure 7 and 8, we show the results of our analysis in detail; for each figure, the first panel shows $\ddot{h}^2_k$ estimates for HESS and BAGHERA, while the second one is limited to the significant regions defined by BAGHERA and overlapping HESS estimates, and the last panel, instead, rescales HESS $\ddot{h}^2_k$ estimates to BAGHERA's $h^2_k$, as $\ddot{h}^2_k/M_k \times M$. It is straightforward to note that BAGHERA provides more robust local heritability estimates, since the number of negative estimates is significantly lower than HESS, as clearly

shown when rescaling the results. While BAGHERA might still return negative local heritability estimates, in practice, this phenomenon is well controlled compared to HESS.

## 1.3 Analysis of 38 UK Biobank cancer datasets

### 1.3.1 Data processing and curation

We downloaded the metadata tables associated with the UK Biobank summary statistics for cancer on $30/07/2019$ from `http://www.nealelab.is/uk-biobank`. From the list of all phenotypes, we selected those corresponding to malignant neoplasms, which are identified by ICD10 codes C00-C97 (see `http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202`), and removed the benign neoplasms and in situ carcinoma/melanoma and the secondary neoplasms (C77,C78,C79). With these parameters, we identified 38 different types of cancers.

LD-score data was downloaded from `https://data.broadinstitute.org/alkesgroup/LDSCORE/` on 15/03/2018-15/04/2018 and used Gencode version 31 available at `https://www.gencodegenes.org/`). The Gene Ontology (GO) slim dataset was generated using the MAP2SLIM utility of the OWL tools on 16/10/2019. We also report enrichment results for the entire Gene Ontology dataset downloaded from the MSigDB, (`http://software.broadinstitute.org/gsea/msigdb`). The Precision Oncology Knowledge Base (OncoKB) dataset, alongside the MSK and Vogelstein data, were downloaded on 01/10/2018, while the Cancer Gene Census data was downloaded from `https://cancer.sanger.ac.uk/census` on 17/07/2019. The DNA repair gene list has been downloaded from `https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html` on 25/02/2019. The PCAWG compendium of mutational driver elements was downloaded on 24/04/2020 from `https://dcc.icgc.org/pcawg/`. All dates are reported as dd/mm/yyy.

### 1.3.2 Relationship between genome-wide significant SNPs and local heritability

We tested whether higher levels of heritability could be explained by the presence of genome-wide significant SNPs ($P < 5 \times 10^{-8}$) in or nearby protein-coding regions.

For each cancer, we identified loci harbouring at least $1$ genome-wide significant SNP, and denoted these as minSNPs. We found $119$ minSNPs in total, with at least $1$ minSNP in $18$ of the $38$ cancers (Supplementary Table 5). This is a striking difference compared to the $1523$ heritability loci found in total for all $38$ malignancies; interestingly, our method was able to recover $98$ $(82\%)$ of the minSNSP suggesting that it can detect heritability genes regardless of the association strength of their SNPs.

We then proceeded to analyse whether there is a correlation between minSNP p-values and heritability estimates. Interestingly, while we found many minSNPs to be also heritability loci, we do not observed a linear relationship between BAGHERA $\eta$ estimates and GWAS p-values (see Supplementary Figure 17 and 18). However, as expected, there is a correlation between each gene average statistics and local heritability (see Supplementary Figure 17).

### 1.3.3 Comparison with self-reported tumors

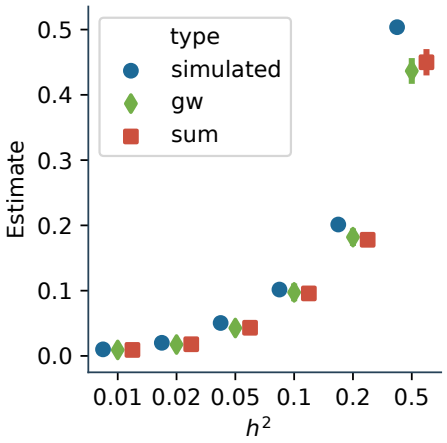
The UK Biobank provides GWAS results for multiple malignancies classified by patient self-reported cancer type at time of assessment. Here we show the results for this dataset using summary statistics computed by B. Neale et al. We found only 11 datasets with $\hat{\chi^2} > 1.01$ compared to the 17 found using the histologically classified tumors (see Supplementary Table 4), along with higher prevalence for the latter $(0.0029)$ compared to the average of self-reported tumors $(0.0023)$.

We then proceeded with the analysis of the self-reported dataset, similarly to what shown for the histologically characterized tumours. Breast and prostate cancer show high values of heritability, with both breast and testicular cancer have more than $30\%$ of their heritability explained by heritability loci (see Supplementary Figure 15A). As expected, these datasets, whose signal is lower compared to the histologically classified malignancies, have a higher heritability enrichment, consistent with results on simulated data (Supplementary Figure 15B). CHGs occurring in multiple malignancies are consistent both in number (see Supplementary Figure 15C) and identity with those found in the 38 cancers identified using the histological classification (Supplementary Figure 15D and 12D).

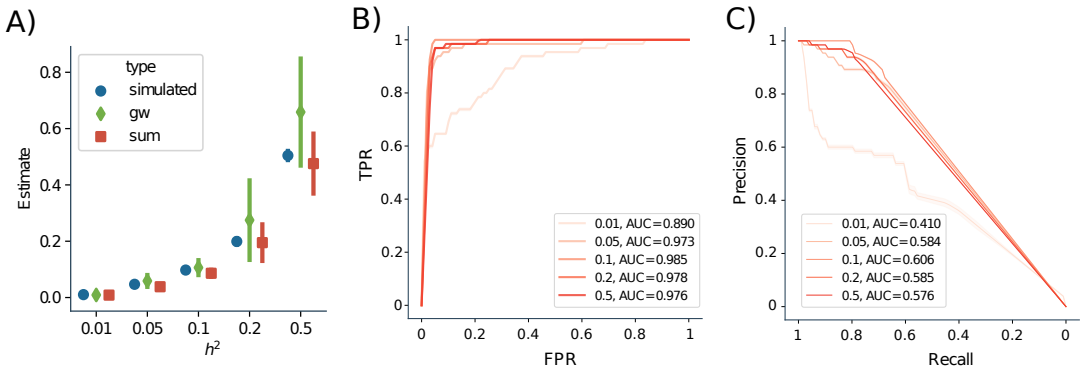
Overall, we find that quantitatively comparing the heritability loci results for self-reported and histologically classified cancers might be difficult. We then considered the Jaccard similarity coefficient computed between heritability genes for each pair of cancers (see Supplementary Figure 14). Here we used the Gencode v27 annotation, which might have resulted in a slightly different mapping of the genes; thus, for the Jaccard coefficient, we directly compared the genes rather than loci. As expected, in some cases, there is consensus between same cancers, although the great differences in signal and the different mapping might decrease the power of detecting similarities, especially for tumours with fewer heritability loci.

Interestingly, when characterizing the CHGs for the self-reported cancer types, we find the overall results to be highly consistent with those of the histologically characterized datasets (see Supplementary Figure 16). We would also like to point out that $90\%$ of the significant GO terms in this analysis are also significant in the same analysis for the histologically characterized cancers; moreover, we also found a significant enrichment for tumour suppressors genes over oncogenes.

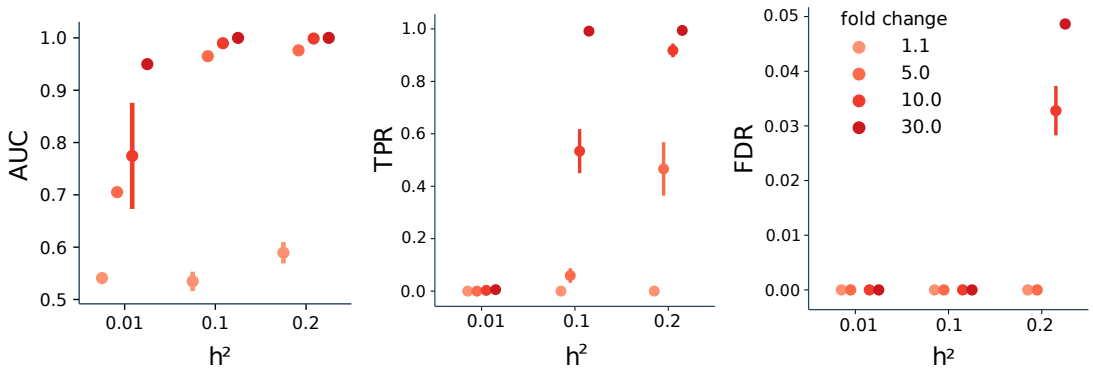
## 2 Supplementary Figures



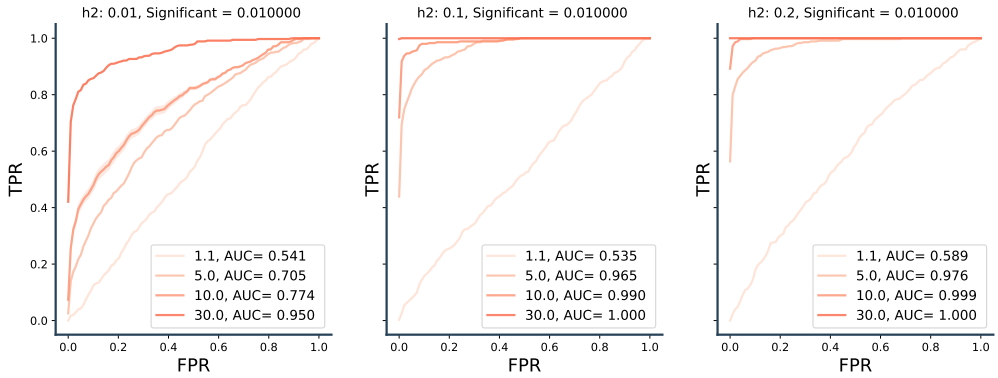

Supplementary Figure 1: **Performance on genome-wide heritability estimation for simulated dense effect datasets.** Genome-wide heritability estimates for dense effects. For each value of $h^2$, we plot the simulated heritability level, the genome-wide (gw) estimate, which is the median of the posterior of genome-wide heritability term, and the gene-level estimate which is the sum of all median gene heritability estimates (sum). For each parameter setting, we simulated 5 datasets, where error bars represent the standard deviation of the estimates. Genotype data has been simulated only for chromosome 1.



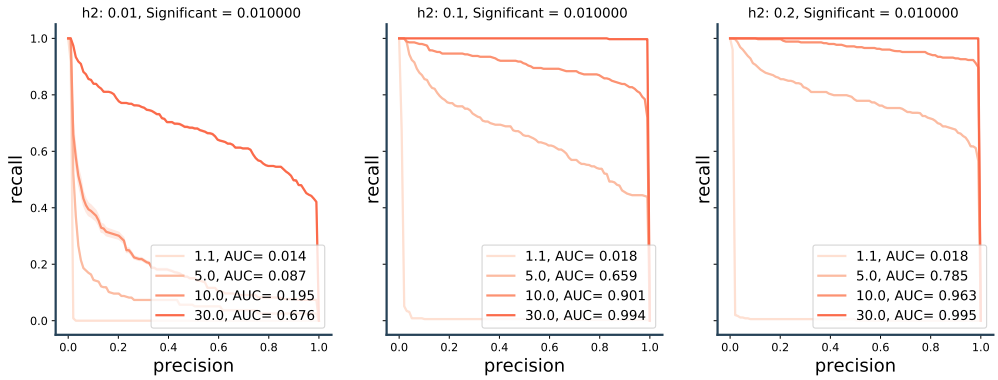

Supplementary Figure 2: **Performance on gene-level heritability estimation for simulated datasets.** A) Genome-wide heritability estimates for datasets with varying gene-level heritability. For each value of $h^2$, we plot the simulated heritability level, the genome-wide (gw) estimate, which is the median of the prior heritability term, and the gene-level estimate which is the sum of all median gene heritability estimates (sum). For each parameter setting, we have simulated 5 datasets, error bars represent the standard deviation of the estimates across different datasets. Genotype data has been simulated only for chromosome 1. B-C) Receiver Operator Characteristic curves and Precision Recall curves for the performance of BAGHERA in discovering significant loci for different levels of genome-wide heritability $h^2$. For each parameter setting, we simulated 5 datasets.



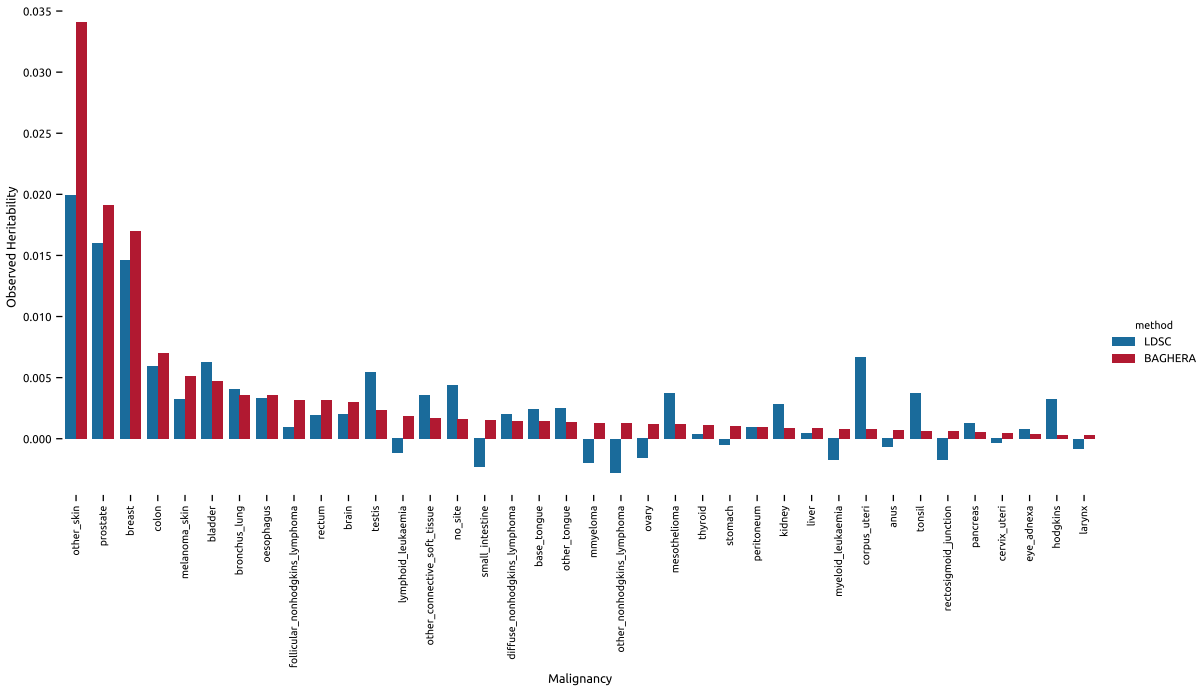

Supplementary Figure 3: **Performance on whole-genome simulated data.** Performance of BAGHERA for different levels of heritability $h^2$ (x-axes) and gene-level heritability enrichment (color coded). Here we show the AUCs of the ROC curves, the True Positive Rate (TPR) and False Discovery Rate (FDR) for $\eta > 0.99$. Datasets have been simulated from summary statistics for 22 chromosomes.



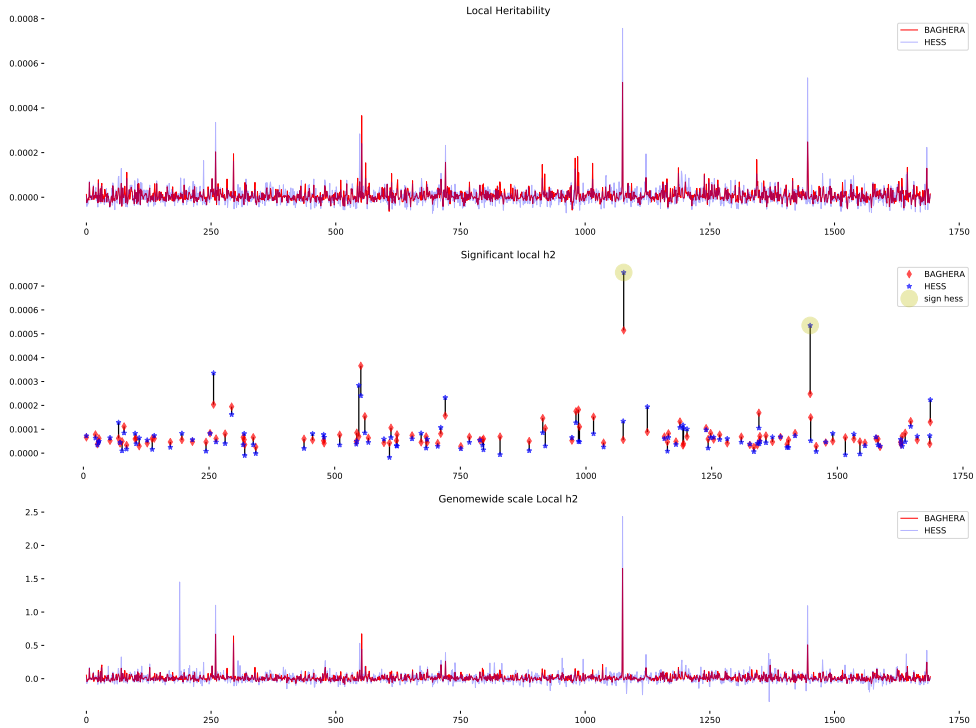

Supplementary Figure 4: **ROC curves for summary statistics simulations.** Receiver Operating Characteristic curve for data simulated from summary statistics. Fold changes, $f_c = \{1.1, 5, 10, 30\}$, are color-coded, while each column corresponds to different values of $h^2 = \{0.01, 0.1, 0.2\}$.



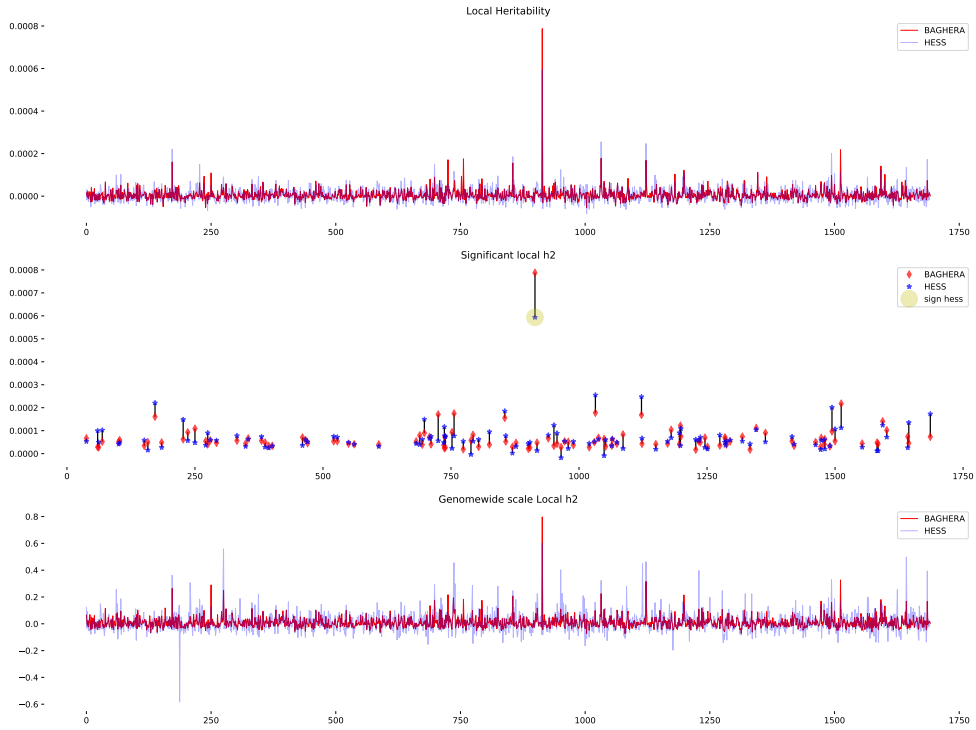

Supplementary Figure 5: **PR curves for summary statistics simulations.** Precision Recall curves for the data simulated from summary statistics. Fold changes, $f_c = \{1.1, 5, 10, 30\}$, are color-coded, while each column corresponds to different values of $h^2 = \{0.01, 0.1, 0.2\}$.



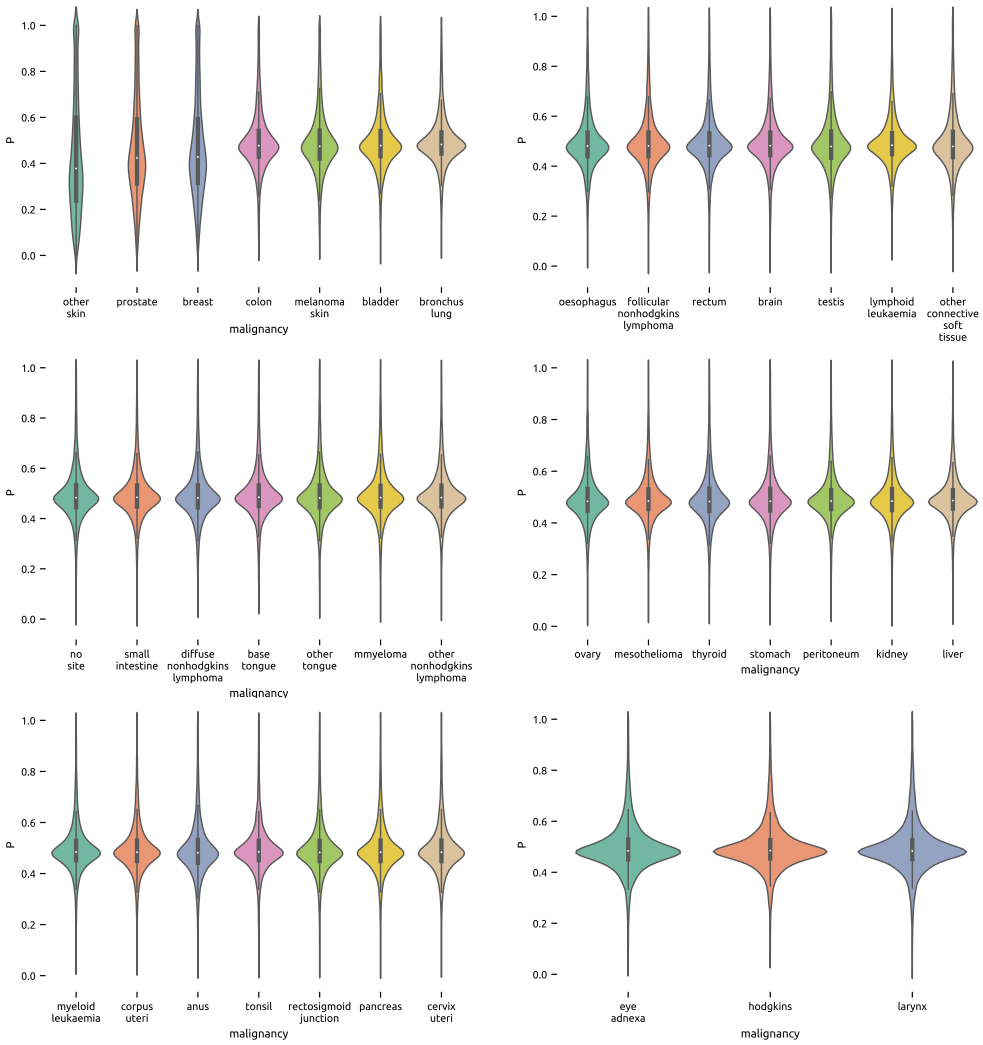

Supplementary Figure 6: **Comparison between LDSC and BAGHERA heritability estimates.** For each of the 38 malignancies (x-axis), we show the observed $h^2$ estimate (y-axis) for LDSC (blue) and BAGHERA (red).



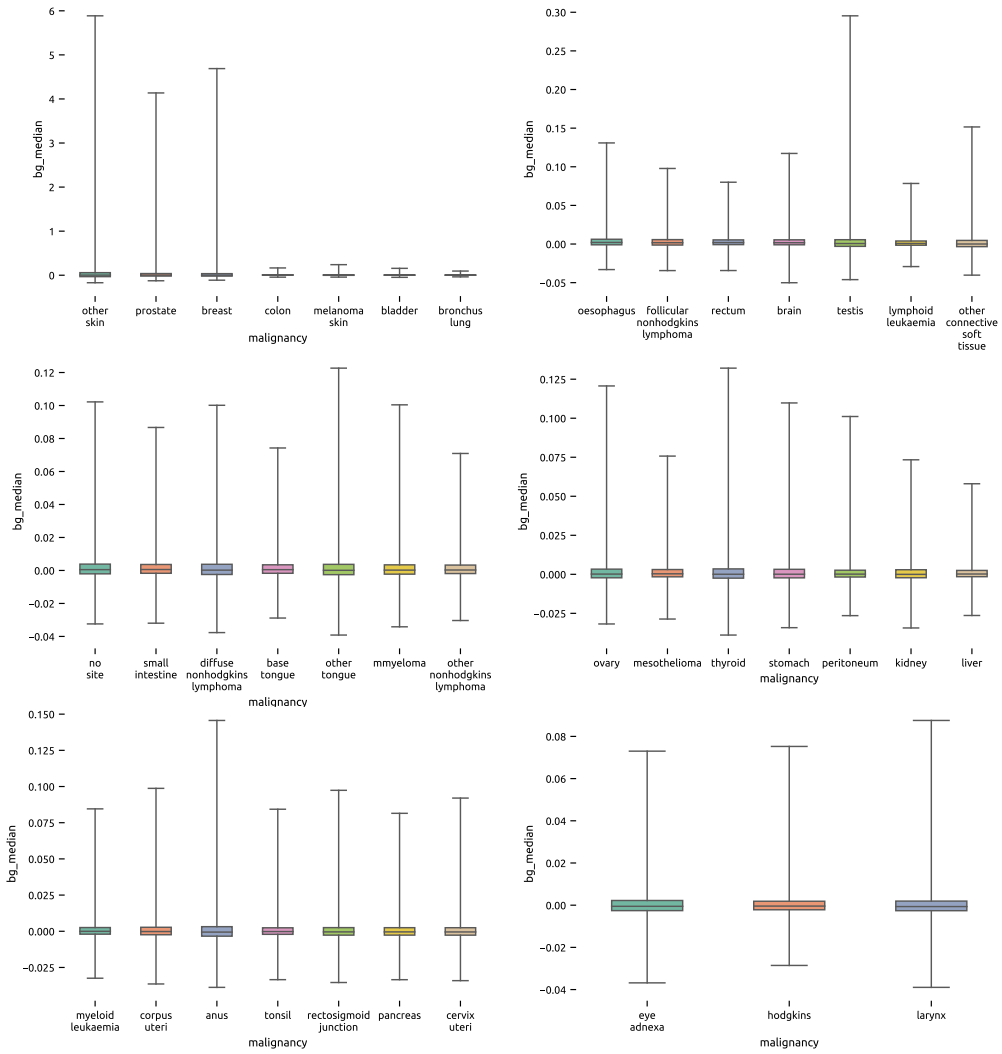

Supplementary Figure 7: **Comparison between BAGHERA and HESS local heritability estimates for breast cancer (C50).** The first panel shows HESS and BAGHERA values of local heritability $\ddot{h}^2_k$. The second panel reports the values of $\ddot{h}^2_k$, but it is limited the regions that are reported as significant by BAGHERA and HESS. The last panel, instead, shows HESS estimates rescaled to be comparable with BAGHERA, as $\ddot{h}^2_k / M_k \times M$.



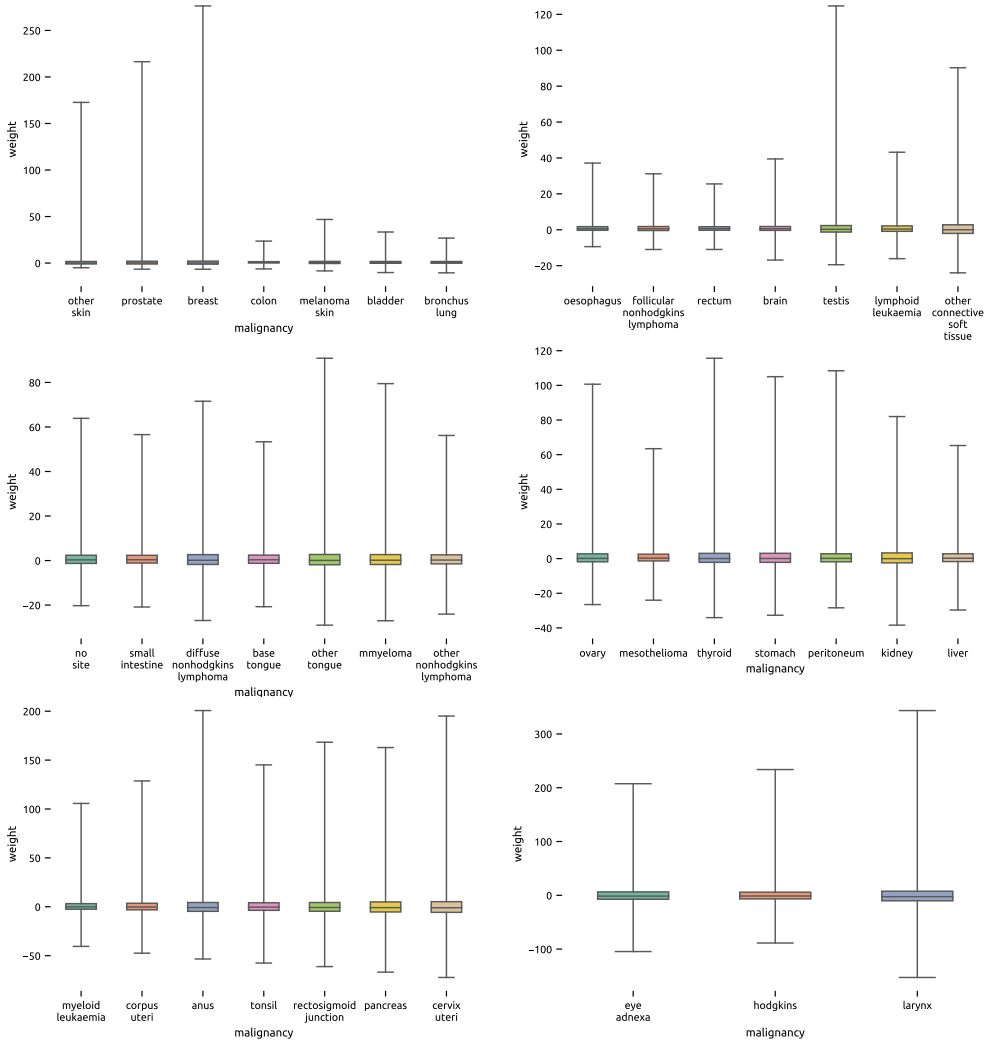

Supplementary Figure 8: **Comparison between BAGHERA and HESS local heritability estimates for prostate cancer (C61).** The first panel shows HESS and BAGHERA values of local heritability $\ddot{h}^2_k$. The second panel reports the values of $\ddot{h}^2_k$, but it is limited the regions that are deemed as significant by BAGHERA and HESS. The last panel, instead, shows HESS estimates rescaled to be comparable with BAGHERA, as $\ddot{h}^2_k / M_k \times M$.



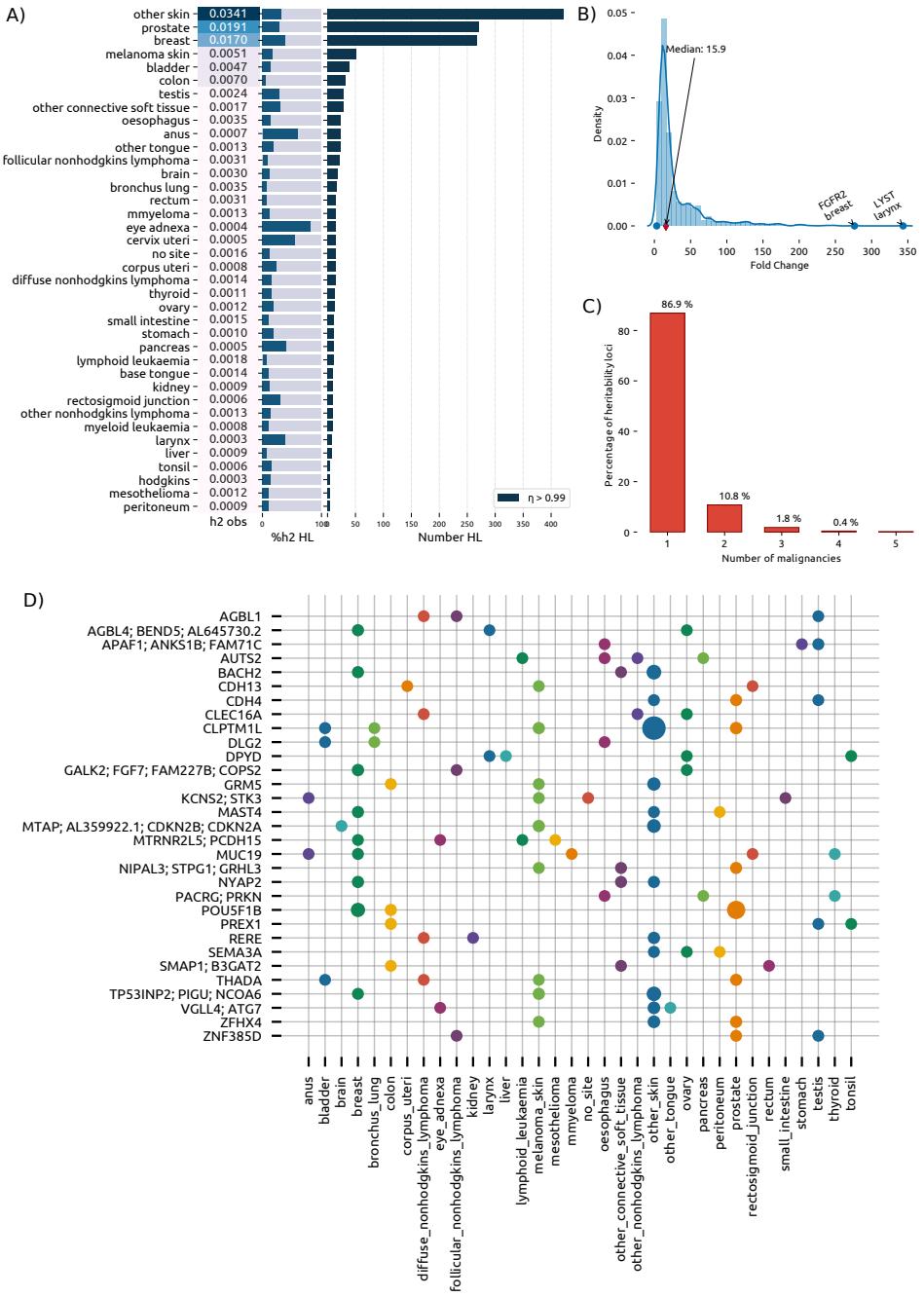

Supplementary Figure 9: **BAGHERA results - $\eta$ distribution across $38$ cancers in the UK Biobank.** For each dataset (x-axes), a violin plot shows the mass distribution of the indicator function $\eta$, which in the software implementation is named P (y-axes).



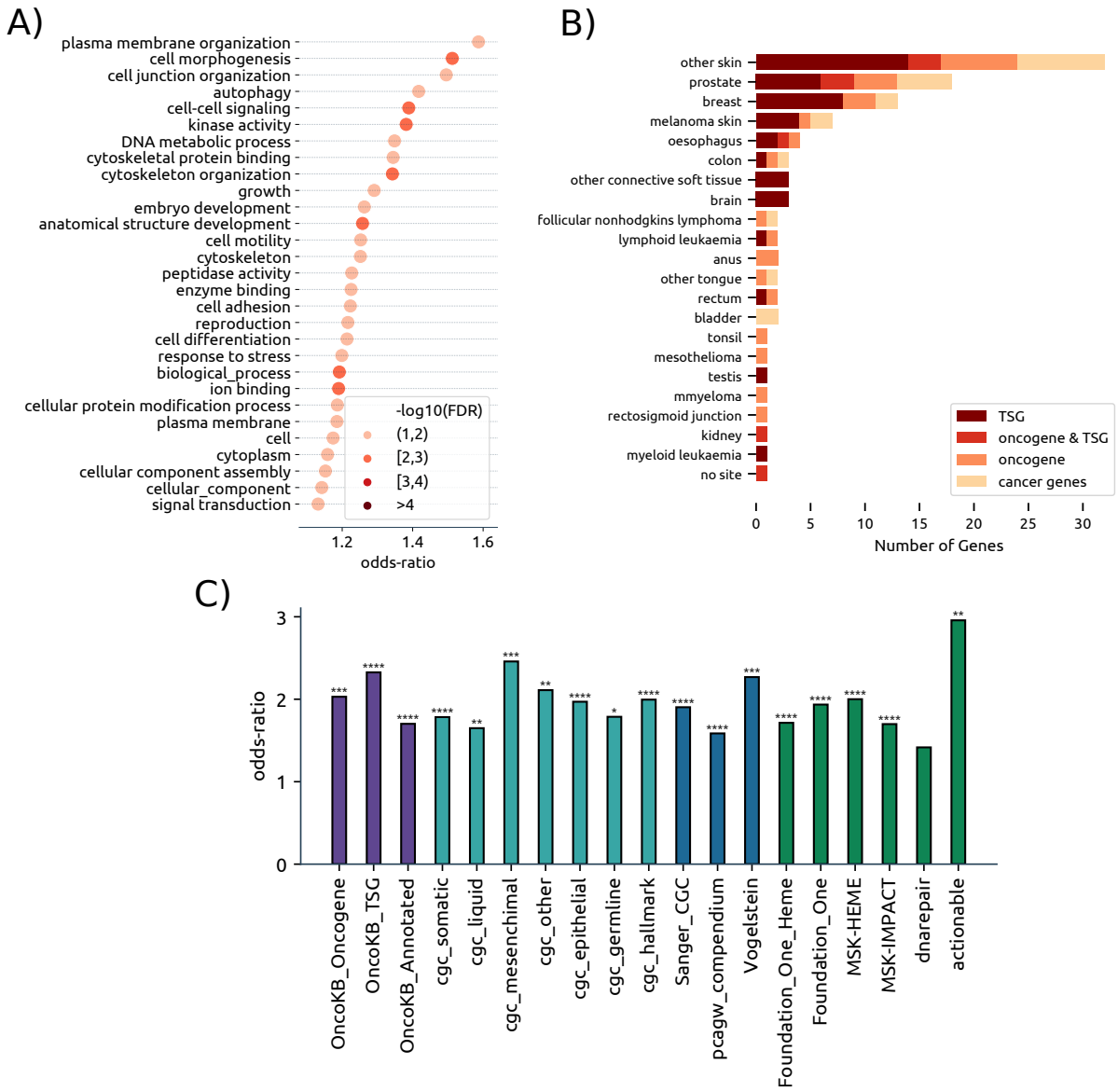

Supplementary Figure 10: **BAGHERA results - local heritability distribution across $38$ cancers in the UK Biobank.** For each dataset (x-axes), we show the boxplot of the median $h^2_k$ for each gene, which in the software implementation is named bg median (y-axes).



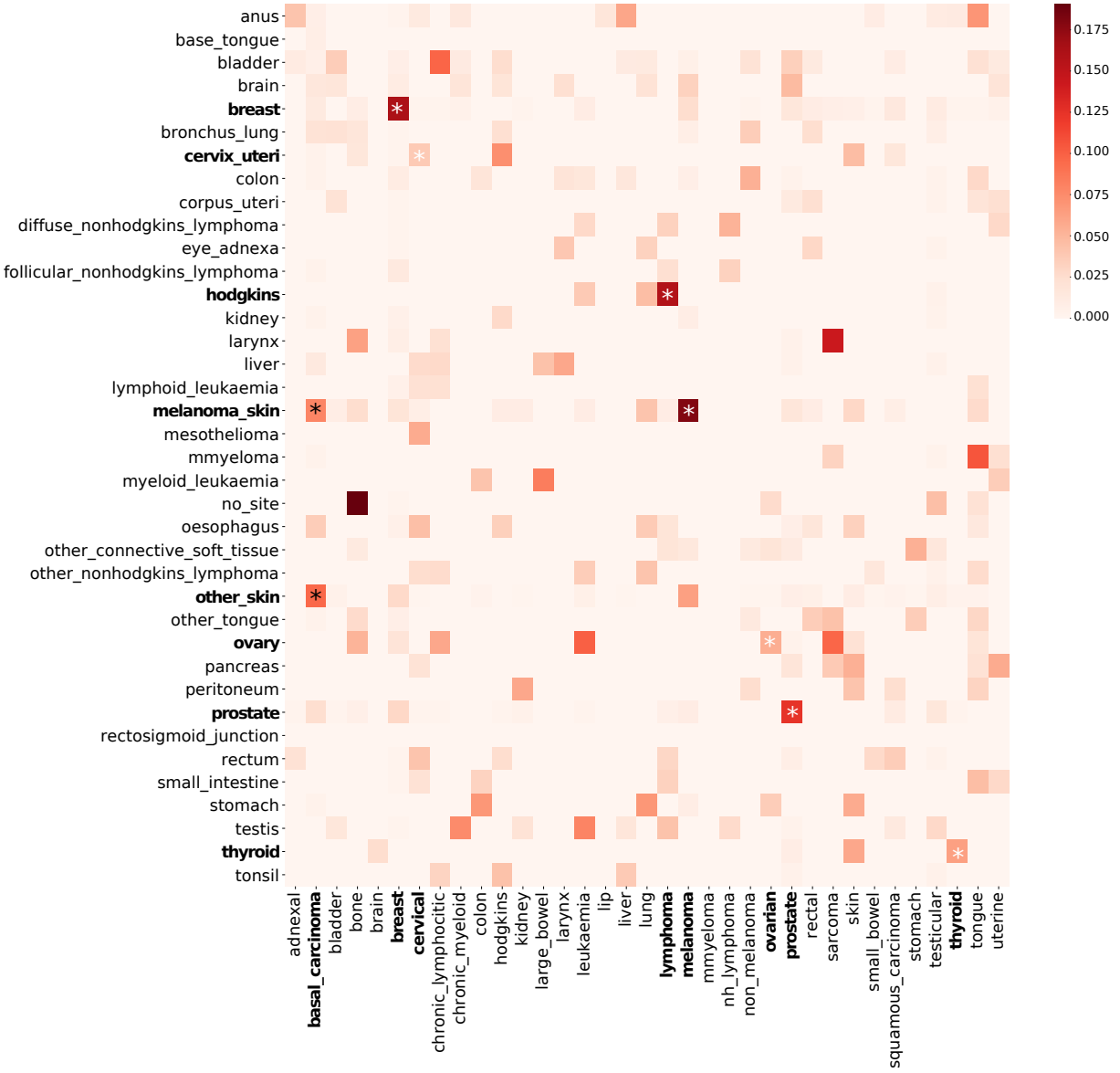

Supplementary Figure 11: **BAGHERA results overview: local heritability weights across** $38$ **cancers in the UK Biobank**. For each analysed dataset (x-axes), we show the boxplot of the local heritability weights $w_k = (h_k^2 - h^2)/h^2$ for each gene. Please note that the fold change has the following relationship with the weights: $f_{c_k} = w_k + 1$.



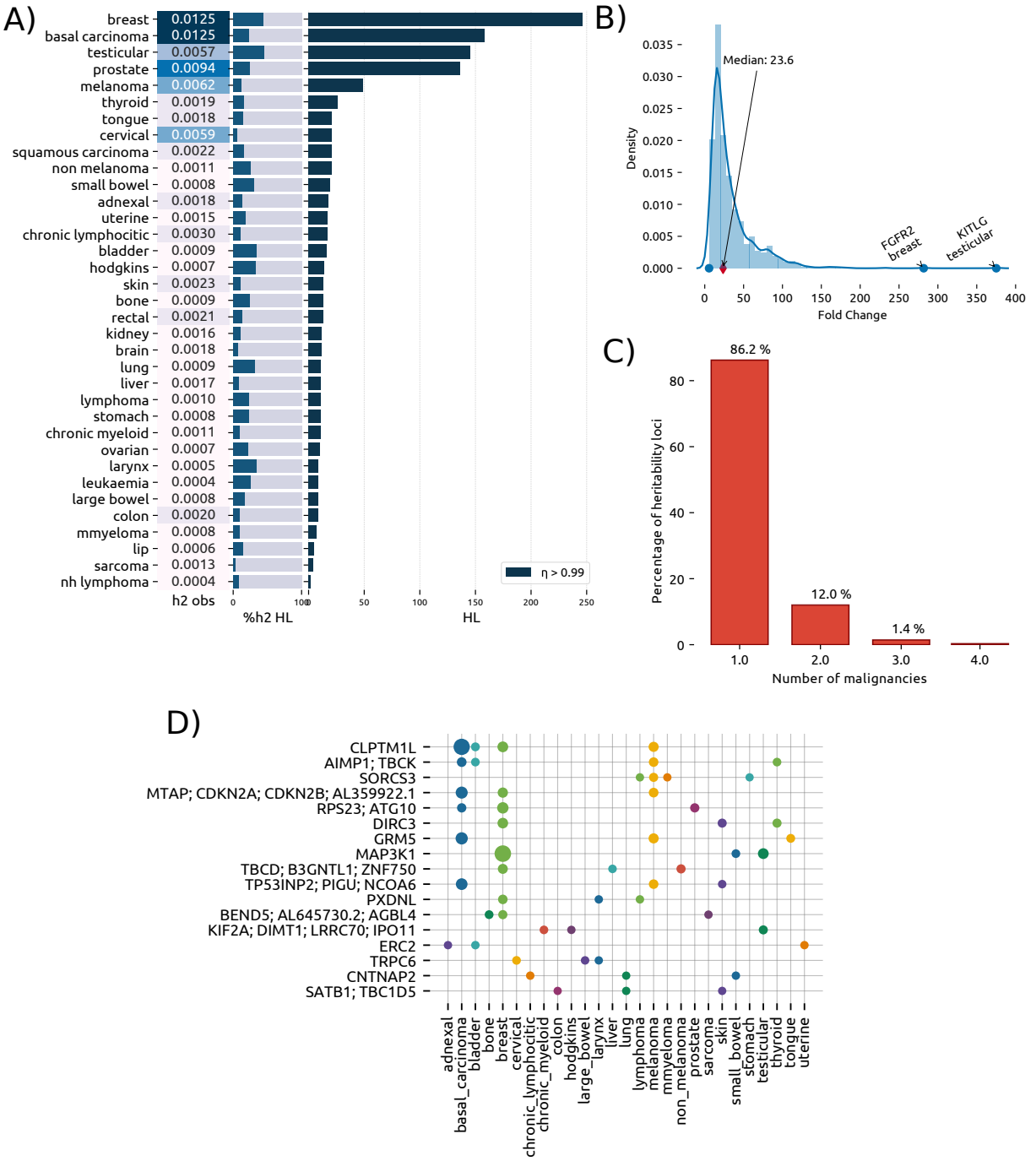

Supplementary Figure 12: **Heritability loci across $38$ cancers in the UK Biobank.** A) For each malignancy we report the observed heritability ($h^2_{SNP}$, left box), the percentage of $h^2_{SNP}$ explained by heritability loci (central barplot, dark blue is the percentage explained by HLs) and the number of heritability loci (right barplot). B) Gene-level heritability density distribution across heritability loci, expressed as fold-change with respect to the genome-wide estimate. Highlighted are the top loci and the median fold-change across all cancers. C) Percentage of cancer heritability loci associated with multiple cancers. Less than $13\%$ of heritability loci are common to multiple malignancies. D) Cancer heritability loci associated with multiple cancers. We report the loci common to at least 3 malignancies sorted by name, for example we can notice that CLPTM1L is common to 5 cancer types. Here the size of the dot is proportional to the fold-change of the locus in the specific cancer.



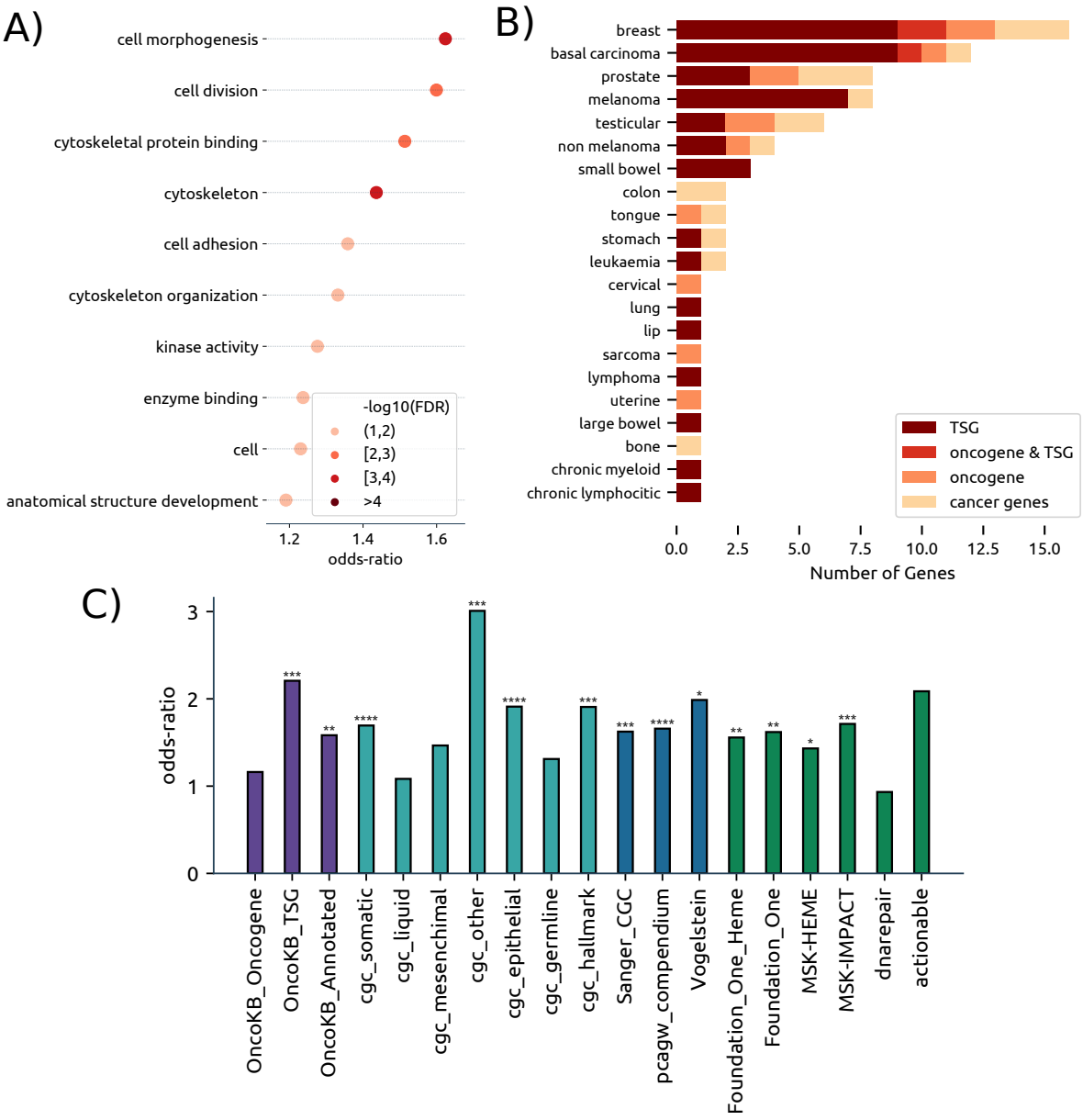

Supplementary Figure 13: **Functional characterization of cancer heritability genes across** $38$ **cancers in the UK Biobank**. A) Gene Ontology enrichment analysis using Fisher's exact test. For each significant term, we report the odds-ratio (x-axis) and $-log_{10}$(FDR) (color gradients). B)Tumour 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, cancer genes are known to be drivers, but their specific role is not reported). C) 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 $p < 10^{-4}$.



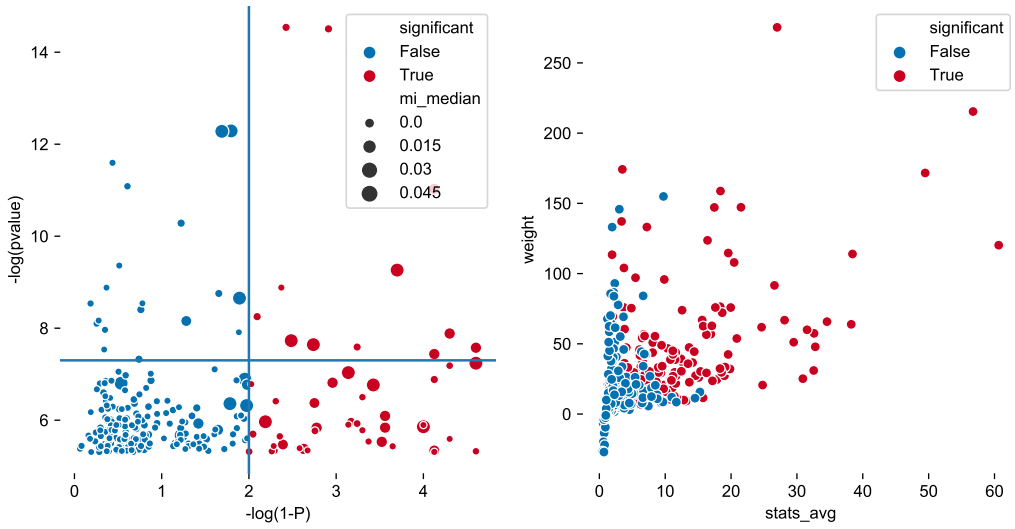

Supplementary Figure 14: **Jaccard similarity coefficient of heritability loci obtained from the 38 ICD10-classified datasets and the 35 self-reported cancers in the UKBB**. The heatmap shows the Jaccard similarity coefficient between significant genes of the histologically characterized dataset, y-axis, and the self-reported ones, x-axis, with darker colours corresponding to higher similarity. In bold and with white stars we have highlighted high similarities for the same tumour type, while with the dark stars we have highlighted the similarity between different skin-cancer types.



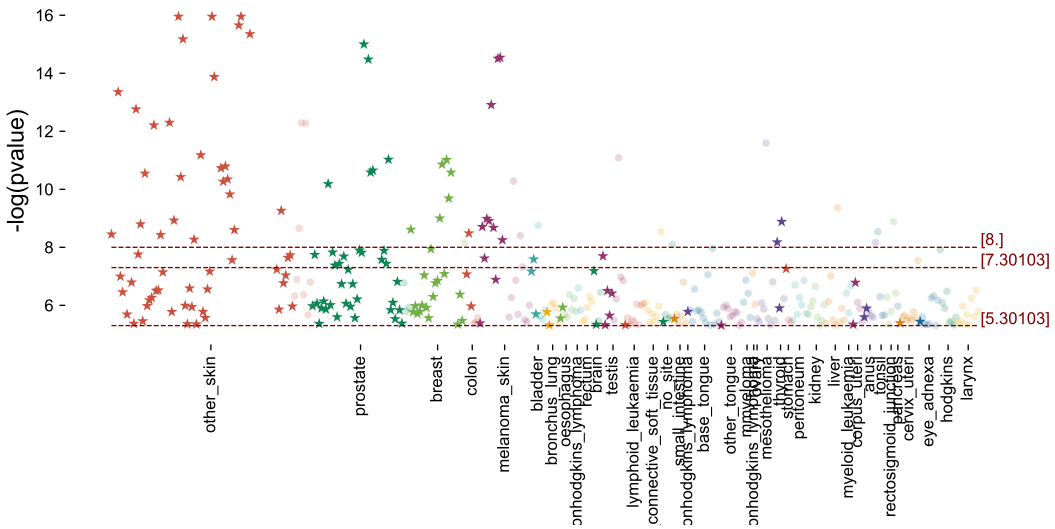

Supplementary Figure 15: **Heritability loci across $35$ self-reported cancers in the UK Biobank** A) For each malignancy, we report the observed heritability ($h^2_{SNP}$, left box), the percentage of $h^2_{SNP}$ explained by heritability loci (central barplot, dark blue is the percentage explained by HLs) and the number of heritability loci (right barplot). B) Gene-level heritability density distribution across heritability loci, expressed as fold-change with respect to the genome-wide estimate. Highlighted are the top loci and the median fold-change across all cancers. C) Percentage of cancer heritability loci associated with multiple cancers. More than $10\%$ of loci are common to multiple malignancies. D) Cancer heritability loci associated with multiple cancers. We report the HLs common to at least 3 cancers; here the size of the dot is proportional to the heritability enrichment of the locus in the specific cancer.




Supplementary Figure 16: **Functional characterization of cancer heritability genes for the** $35$ **self-reported cancers**. A) 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). B)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, cancer genes are known to be drivers, but their specific role is not reported). C) 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 $P < 10^{-4}$.




Supplementary Figure 17: **Relationship between genome-wide significant SNPs and local heritability across the $38$ cancers in the UK Biobank.** On the left panel, we show the correlation between GWAS pvalues (x-axis, we consider only loci with p:$< 10^{-5}$) and BAGHERA $\eta$ (x-axis; in the software implementation $\eta$ is named P, and it is here transformed to $1 - \eta$ to be comparable to pvalues). For each locus analysed by BAGHERA, we selected the smallest p-value of its SNPs. Horizontal line is the GWAS significance threshold (p: $5 \times 10^{-8}$), vertical line is for $\eta = 0.99$. Size of the marker is proportional to the genome-wide $h^2_{SNP}$ estimate (which in the software implementation is denoted as mi median). It is worth noting that there is no linear relation between BAGHERA $\eta$ and GWAS pvalues. In some cases, see top left quadrant, there are locus harboring SNPs with very small p-values, that are not significant for the heritability analysis. On the right panel, instead, we show the correlation between each locus average $\chi^2$ and local heritability (y-axis, to make results from different cancer types comparable we show the locus weight as $w_k = (h^2_k - h^2)/h^2$). Significant loci are color coded in red. As expected, there is correlation between the average value of the test statistics of a locus and its local heritability.




Supplementary Figure 18: **Single malignancy genome-wide significant SNPs.** For each cancer type, color coded, we selected loci harbouring SNPs with p:$< 10^{-5}$. On the x-axis, for each malignancy, we sorted the loci by their $\eta$, from the largest to the smallest. Loci that are significant for BAGHERA are dark stars, while those that are not significant are represented with dots. Horizontal lines are different p-value significance thresholds. This figure details the results in Supplementary Figure 17

.

# 3 Supplementary Tables

| Genes | chrom | SNPs | cancers | Cancer types |
|---|---|---|---|---|
| CLPTM1L | 5 | 27 | 5 | melanoma skin, prostate, other skin, bronchus lung, bladder |
| MUC19 | 12 | 183 | 5 | thyroid, myeloma, breast, anus, rectosigmoid junction |
| MTRNR2L5; PCDH15 | 10 | 978 | 4 | lymphoid leukaemia, mesothelioma, eye adnexa, breast |
| AUTS2 | 7 | 489 | 4 | oesophagus, lymphoid leukaemia, other nonhodgkins lymphoma, pancreas |
| DPYD | 1 | 574 | 4 | liver, ovary, tonsil, larynx |
| THADA | 2 | 165 | 4 | melanoma skin, prostate, diffuse nonhodgkins lymphoma, bladder |
| KCNS2; STK3 | 8 | 188 | 4 | melanoma skin, small intestine, no site, anus |
| CDH13 | 16 | 1502 | 3 | corpus uteri, melanoma skin, rectosigmoid junction |
| PACRG; PRKN | 6 | 1353 | 3 | thyroid, oesophagus, pancreas |
| NIPAL3; STPG1; GRHL3 | 1 | 136 | 3 | melanoma skin, prostate, other connective soft tissue |
| CLEC16A | 16 | 170 | 3 | other nonhodgkins lymphoma, ovary, diffuse nonhodgkins lymphoma |
| MAST4 | 5 | 383 | 3 | peritoneum, other skin, breast |
| DLG2 | 11 | 1014 | 3 | oesophagus, bronchus lung, bladder |
| APAF1; ANKS1B; FAM71C | 12 | 582 | 3 | testis, oesophagus, stomach |
| SMAP1; B3GAT2 | 6 | 162 | 3 | rectum, other connective soft tissue, colon |
| AGBL1 | 15 | 698 | 3 | testis, diffuse nonhodgkins lymphoma, follicular nonhodgkins lymphoma |
| AGBL4; BEND5; AL645730.2 | 1 | 475 | 3 | ovary, larynx, breast |
| TP53INP2; PIGU; NCOA6 | 20 | 106 | 3 | melanoma skin, other skin, breast |
| GRM5 | 11 | 313 | 3 | melanoma skin, other skin, colon |
| ZFHX4 | 8 | 116 | 3 | melanoma skin, prostate, other skin |
| RERE | 1 | 141 | 3 | kidney, other skin, diffuse nonhodgkins lymphoma |
| CDH4 | 20 | 540 | 3 | testis, prostate, other skin |
| VGLL4; ATG7 | 3 | 245 | 3 | other skin, eye adnexa, other tongue |
| NYAP2 | 2 | 154 | 3 | other skin, other connective soft tissue, breast |
| MTAP; AL359922.1; CDKN2B; CDKN2A | 9 | 162 | 3 | melanoma skin, other skin, brain |
| BACH2 | 6 | 215 | 3 | other skin, other connective soft tissue, breast |
| PREX1 | 20 | 190 | 3 | testis, tonsil, colon |
| GALK2; FGF7; FAM227B; COPS2 | 15 | 157 | 3 | ovary, follicular nonhodgkins lymphoma, breast |
| SEMA3A | 7 | 287 | 3 | peritoneum, other skin, ovary |
| ZNF385D | 3 | 862 | 3 | testis, prostate, follicular nonhodgkins lymphoma |
| POU5F1B | 8 | 137 | 3 | prostate, breast, colon |

Supplementary Table 1: **Heritability loci common to more than 2 malignancies among the** $38$ **cancers in the UK Biobank**. For each locus, we report the gene names, the chromosome, the number of SNPs in the locus, and the cancers for which the locus shows significant heritability enrichment.

| GO Term | GO id | CHGs | TP | OR | p-value | FDR |
|---|---|---|---|---|---|---|
| cell morphogenesis | GO:0000902 | 822 | 140 | 1.51249 | 0.00002 | 0.00145 |
| cell-cell signaling | GO:0007267 | 1364 | 215 | 1.38895 | 0.00003 | 0.00145 |
| anatomical structure development | GO:0048856 | 4094 | 576 | 1.25771 | 0.00002 | 0.00145 |
| kinase activity | GO:0016301 | 1291 | 203 | 1.38162 | 0.00006 | 0.00214 |
| cytoskeleton organization | GO:0007010 | 1260 | 194 | 1.34248 | 0.00029 | 0.00703 |
| biological process | GO:0008150 | 6375 | 848 | 1.19188 | 0.00030 | 0.00703 |
| ion binding | GO:0043167 | 5328 | 716 | 1.18984 | 0.00045 | 0.00900 |
| cell differentiation | GO:0030154 | 3263 | 454 | 1.21364 | 0.00058 | 0.01009 |
| plasma membrane | GO:0005886 | 4994 | 672 | 1.18500 | 0.00069 | 0.01068 |
| response to stress | GO:0006950 | 2975 | 412 | 1.19890 | 0.00164 | 0.02240 |
| cytoskeleton | GO:0005856 | 1597 | 232 | 1.25172 | 0.00210 | 0.02240 |
| cellular protein modification process | GO:0006464 | 3321 | 455 | 1.18618 | 0.00205 | 0.02240 |
| enzyme binding | GO:0019899 | 2076 | 295 | 1.22522 | 0.00199 | 0.02240 |
| DNA metabolic process | GO:0006259 | 789 | 123 | 1.34854 | 0.00244 | 0.02257 |
| cytoskeletal protein binding | GO:0008092 | 817 | 127 | 1.34455 | 0.00231 | 0.02257 |
| cytoplasm | GO:0005737 | 4713 | 628 | 1.15849 | 0.00318 | 0.02763 |
| cell motility | GO:0048870 | 1274 | 186 | 1.25239 | 0.00470 | 0.03845 |
| cellular component | GO:0005575 | 5314 | 699 | 1.14209 | 0.00581 | 0.04488 |
| growth | GO:0040007 | 797 | 120 | 1.29075 | 0.00852 | 0.06231 |
| signal transduction | GO:0007165 | 5214 | 683 | 1.13158 | 0.00974 | 0.06681 |
| autophagy | GO:0006914 | 379 | 62 | 1.41713 | 0.01009 | 0.06681 |
| cell | GO:0005623 | 2157 | 297 | 1.17421 | 0.01101 | 0.06958 |
| cell adhesion | GO:0007155 | 1149 | 165 | 1.22322 | 0.01378 | 0.07801 |
| peptidase activity | GO:0008233 | 1118 | 161 | 1.22701 | 0.01351 | 0.07801 |
| embryo development | GO:0009790 | 818 | 121 | 1.26283 | 0.01409 | 0.07801 |
| cell junction organization | GO:0034330 | 245 | 42 | 1.49541 | 0.01459 | 0.07801 |
| cellular component assembly | GO:0022607 | 2556 | 346 | 1.15242 | 0.01559 | 0.08027 |
| plasma membrane organization | GO:0007009 | 172 | 31 | 1.58688 | 0.01700 | 0.08150 |
| reproduction | GO:0000003 | 1133 | 162 | 1.21614 | 0.01689 | 0.08150 |

Supplementary Table 2: **Statistically significant Gene Ontology terms for the 38 cancers in the UK Biobank.** We report the gene ontology terms significantly associated with cancer heritability genes of all $38$ cancers in the UKBB, at $10\%$FDR. For each term, we report the GO id term, the number of annotated CHGs, the number of CHGs shared with the GO term, the odds ratio, the p-value from the Fisher's Exact test and the adjusted p-value after applying the Benjamini-Hochberg procedure.

| Geneset | CHGs | OR | p-value |
|---|---|---|---|
| actionable | 12 | 2.95704402853006 | 0.003010513617533 |
| OncoKB Annotated | 82 | 1.70182693656355 | 3.45E-05 |
| OncoKB Oncogene | 30 | 2.03015313527443 | 0.000989619358728 |
| OncoKB TSG | 41 | 2.32559883961873 | 1.10E-05 |
| MSK-IMPACT | 74 | 1.69855042892001 | 8.20E-05 |
| MSK-HEME | 72 | 2.00040589657017 | 1.04E-06 |
| Foundation One | 60 | 1.93523581681476 | 1.70E-05 |
| Foundation One Heme | 93 | 1.71410442349529 | 8.99E-06 |
| Vogelstein | 25 | 2.26853809360218 | 0.000688378926034 |
| Sanger CGC | 105 | 1.90314876984706 | 4.42E-08 |
| cgc hallmark | 52 | 1.99517925729025 | 2.98E-05 |
| cgc somatic | 114 | 1.78333561882259 | 2.14E-07 |
| cgc germline | 19 | 1.7869406867846 | 0.021626151797484 |
| cgc epithelial | 68 | 1.96978537106247 | 3.10E-06 |
| cgc other | 18 | 2.11038080867497 | 0.006672381597831 |
| cgc mesenchimal | 24 | 2.45812653699978 | 0.000340751626412 |
| cgc liquid | 50 | 1.64904739495146 | 0.001689100231574 |
| dnarepair | 23 | 1.41604940491173 | 0.085540295201593 |
| pcagw compendium | 111 | 1.58551000032207 | 2.59E-05 |

Supplementary Table 3: **Cancer genesets enrichment analysis for the 38 cancers in the UK Biobank.** Results of the enrichments analysis between the Curated cancer dataset terms and the heritability genes of all datasets.

| code | Malignancy | cases | prevalence | $\hat{\chi^2}$ | $h^2_{SNP}$ | $h^2_{SNP_L}$ | HL |
|---|---|---|---|---|---|---|---|
| 1002 | **breast cancer** | 7480 | 0.02219 | 1.08192 | 0.01245 | 0.09668 | 246 |
| 1061 | **basal cell carcinoma** | 3156 | 0.00936 | 1.06533 | 0.01250 | 0.18314 | 158 |
| 1044 | **prostate cancer** | 2495 | 0.00740 | 1.05405 | 0.00939 | 0.16460 | 136 |
| 1045 | **testicular cancer** | 614 | 0.00182 | 1.03105 | 0.00567 | 0.30420 | 145 |
| 1059 | **malignant melanoma** | 2677 | 0.00794 | 1.02615 | 0.00622 | 0.10342 | 49 |
| 1041 | **cervical cancer** | 1347 | 0.00400 | 1.02078 | 0.00590 | 0.16776 | 21 |
| 1022 | **colon cancer/sigmoid cancer** | 1134 | 0.00336 | 1.01659 | 0.00196 | 0.06403 | 9 |
| 1040 | **uterine/endometrial cancer** | 843 | 0.00250 | 1.01499 | 0.00148 | 0.06127 | 17 |
| 1062 | **squamous cell carcinoma** | 404 | 0.00120 | 1.01276 | 0.00225 | 0.17012 | 21 |
| 1065 | **thyroid cancer** | 317 | 0.00094 | 1.01245 | 0.00195 | 0.18077 | 26 |
| 1023 | **rectal cancer** | 253 | 0.00075 | 1.01187 | 0.00213 | 0.23923 | 13 |
| 1034 | kidney/renal cell cancer | 436 | 0.00129 | 1.00968 | 0.00156 | 0.11121 | 12 |
| 1035 | bladder cancer | 799 | 0.00237 | 1.00685 | 0.00091 | 0.03954 | 16 |
| 1003 | skin cancer | 1046 | 0.00310 | 1.00679 | 0.00226 | 0.07854 | 13 |
| 1019 | small intestine/small bowel cancer | 156 | 0.00046 | 1.00618 | 0.00076 | 0.12919 | 19 |
| 1030 | eye and/or adnexal cancer | 102 | 0.00030 | 1.00408 | 0.00184 | 0.44827 | 18 |
| 1052 | hodgkins lymphoma / hodgkins disease | 331 | 0.00098 | 1.00324 | 0.00067 | 0.06010 | 14 |
| 1047 | lymphoma | 92 | 0.00027 | 1.00229 | 0.00101 | 0.26830 | 11 |
| 1063 | primary bone cancer | 105 | 0.00031 | 1.00193 | 0.00090 | 0.21425 | 13 |
| 1053 | non-hodgkins lymphoma | 631 | 0.00187 | 1.00082 | 0.00043 | 0.02267 | 2 |
| 1060 | non-melanoma skin cancer | 507 | 0.00150 | 1.00076 | 0.00109 | 0.06863 | 21 |
| 1018 | stomach cancer | 121 | 0.00036 | 0.99947 | 0.00079 | 0.16616 | 11 |
| 1068 | sarcoma/fibrosarcoma | 181 | 0.00054 | 0.99930 | 0.00126 | 0.18758 | 4 |
| 1011 | tongue cancer | 115 | 0.00034 | 0.99905 | 0.00181 | 0.39809 | 21 |
| 1006 | larynx/throat cancer | 250 | 0.00074 | 0.99786 | 0.00052 | 0.05865 | 9 |
| 1004 | cancer of lip/mouth/pharynx/oral cavity | 78 | 0.00023 | 0.99756 | 0.00060 | 0.18505 | 5 |
| 1039 | ovarian cancer | 579 | 0.00172 | 0.99745 | 0.00069 | 0.03903 | 10 |
| 1056 | chronic myeloid | 85 | 0.00025 | 0.99734 | 0.00112 | 0.32044 | 11 |
| 1032 | brain cancer / primary malignant brain tumour | 155 | 0.00046 | 0.99648 | 0.00177 | 0.30057 | 12 |
| 1048 | leukaemia | 158 | 0.00047 | 0.99611 | 0.00045 | 0.07506 | 9 |
| 1024 | liver/hepatocellular cancer | 125 | 0.00037 | 0.99530 | 0.00168 | 0.34389 | 11 |
| 1020 | large bowel cancer/colorectal cancer | 475 | 0.00141 | 0.99524 | 0.00077 | 0.05125 | 9 |
| 1001 | lung cancer | 190 | 0.00056 | 0.99519 | 0.00091 | 0.13020 | 11 |
| 1050 | multiple myeloma | 115 | 0.00034 | 0.99491 | 0.00083 | 0.18195 | 7 |

Supplementary Table 4: **Self reported cancers in the UK Biobank.** We report summary informations of of the $35$ self-reported cancer types analysed in the first round of the GWAS analysis on the UK Biobank. For each cancer, we report the number of cases out of the $337,159$ total samples, the prevalence in the cohort, the average $\chi^2$ 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 $\eta > 0.99$. Both prevalence and $\hat{\chi^2}$ are lower than the data used in the main study; in particular, there are only $11$ tumours with $\hat{\chi^2} > 1.01$.

| ICD10 | Cancer | Significant SNPs | minSNPs | minSNP ∩ HL | HL |
|---|---|---|---|---|---|
| C44 | Other malignant neoplasms of skin | 580 | 58 | 55 | 422 |
| C50 | Malignant neoplasm of breast | 178 | 10 | 9 | 267 |
| C61 | Malignant neoplasm of prostate | 203 | 20 | 20 | 271 |
| C18 | Malignant neoplasm of colon | 4 | 1 | 1 | 33 |
| C43 | Malignant melanoma of skin | 42 | 14 | 9 | 52 |
| C15 | Malignant neoplasm of oesophagus | 0 | 0 | 0 | 24 |
| C67 | Malignant neoplasm of bladder | 11 | 2 | 1 | 39 |
| C34 | Malignant neoplasm of bronchus and lung | 0 | 0 | 0 | 17 |
| C20 | Malignant neoplasm of rectum | 0 | 0 | 0 | 15 |
| C62 | Malignant neoplasm of testis | 19 | 2 | 1 | 29 |
| C71 | Malignant neoplasm of brain | 0 | 0 | 0 | 19 |
| C45 | Mesothelioma | 1 | 1 | 0 | 5 |
| C91 | Lymphoid leukaemia | 0 | 0 | 0 | 11 |
| C02 | Malignant neoplasm of other and unspecified parts of tongue | 0 | 0 | 0 | 23 |
| C16 | Malignant neoplasm of stomach | 0 | 0 | 0 | 12 |
| C83 | Diffuse non-Hodgkin's lymphoma | 1 | 0 | 0 | 14 |
| C82 | Follicular non-Hodgkin's lymphoma | 0 | 0 | 0 | 21 |
| C90 | Multiple myeloma and malignant plasma cell neoplasms | 0 | 0 | 0 | 15 |
| C56 | Malignant neoplasm of ovary | 0 | 0 | 0 | 13 |
| C54 | Malignant neoplasm of corpus uteri | 0 | 0 | 0 | 14 |
| C48 | Malignant neoplasm of retroperitoneum and peritoneum | 0 | 0 | 0 | 5 |
| C64 | Malignant neoplasm of kidney except renal pelvis | 0 | 0 | 0 | 10 |
| C01 | Malignant neoplasm of base of tongue | 1 | 1 | 0 | 10 |
| C73 | Malignant neoplasm of thyroid gland | 23 | 2 | 2 | 13 |
| C49 | Malignant neoplasm of other connective and soft tissue | 1 | 1 | 0 | 28 |
| C80 | Malignant neoplasm without specification of site | 1 | 1 | 0 | 14 |
| C53 | Malignant neoplasm of cervix uteri | 1 | 1 | 0 | 14 |
| C22 | Malignant neoplasm of liver and intrahepatic bile ducts | 5 | 1 | 0 | 7 |
| C21 | Malignant neoplasm of anus and anal canal | 1 | 1 | 0 | 23 |
| C85 | Other and unspecified types of non-Hodgkin's lymphoma | 0 | 0 | 0 | 9 |
| C09 | Malignant neoplasm of tonsil | 1 | 1 | 0 | 5 |
| C92 | Myeloid leukaemia | 0 | 0 | 0 | 9 |
| C17 | Malignant neoplasm of small intestine | 0 | 0 | 0 | 12 |
| C19 | Malignant neoplasm of rectosigmoid junction | 1 | 1 | 0 | 10 |
| C25 | Malignant neoplasm of pancreas | 0 | 0 | 0 | 12 |
| C81 | Hodgkin's disease | 6 | 1 | 0 | 5 |
| C69 | Malignant neoplasm of eye and adnexa | 0 | 0 | 0 | 14 |
| C32 | Malignant neoplasm of larynx | 1 | 0 | 0 | 7 |

Supplementary Table 5: **Comparison between GWAS results and gene-level heritability analysis for the $38$ cancers in the UK Biobank.** For each cancer type, we report the number of significant SNPs found by the GWAS analysis, the number of genes that harbor at least a genome-wide significant SNP (minSNSP), the number of heritability loci (HL), and the overlap between minSNP and HL.

| Genes | chrom | SNPs | cancers | Cancer types |
|---|---|---|---|---|
| CLPTM1L | 5 | 27 | 4 | prostate, melanoma skin, bladder,bronchus lung |
| THADA | 2 | 165 | 4 | prostate, melanoma skin,bladder, diffuse nonhodgkins lymphoma |
| APAF1; ANKS1B; FAM71C | 12 | 582 | 3 | oesophagus, testis, stomach |
| MTRNR2L5; PCDH15 | 10 | 978 | 3 | breast, mesothelioma, lymphoid_leukaemia |
| AGBL1 | 15 | 698 | 3 | testis, diffuse nonhodgkins lymphoma, follicular nonhodgkins lymphoma |
| POU5F1B | 8 | 137 | 3 | breast, prostate, colon |
| ZNF385D | 3 | 862 | 3 | prostate, testis, follicular nonhodgkins lymphoma |
| DLG2 | 11 | 1014 | 3 | oesophagus, bladder, bronchus lung |

Supplementary Table 6: **Heritability loci common to more than 2 malignancies among the** $16$ **cancers in the UK Biobank**. The table refers to the top hits of Figure 3D. For each locus, we report the gene names, the chromosome, the number of SNPs, and the cancers for which the locus shows significant heritability enrichment.

| Geneset | OR | CHG in dataset | p-value |
|---|---|---|---|
| actionable | 2.63453493776791 | 7 | 0.026951610993734 |
| OncoKB_TSG | 2.4758427927671 | 27 | 7.90E-05 |
| cgc_mesenchimal | 2.24609098939929 | 14 | 0.007835265509946 |
| MSK-HEME | 2.19714313105167 | 48 | 3.93E-06 |
| cgc_other | 2.07244104690334 | 11 | 0.027306118314416 |
| cgc_hallmark | 2.06286703907705 | 33 | 0.00030018959537 |
| Foundation_One | 1.93993932601498 | 37 | 0.000393887476418 |
| Foundation_One_Heme | 1.83497871569604 | 60 | 3.91E-05 |
| OncoKB_Oncogene | 1.83348095659876 | 17 | 0.019840274395826 |
| Vogelstein | 1.83053839364519 | 13 | 0.038349307393117 |
| OncoKB_Annotated | 1.78464447477968 | 52 | 0.000213156056084 |
| MSK-IMPACT | 1.7840487630967 | 47 | 0.000407877043307 |
| cgc_epithelial | 1.75509927797834 | 38 | 0.001711381100092 |
| Sanger_CGC | 1.74637430939227 | 60 | 0.000130077696757 |
| cgc_somatic | 1.70276736998878 | 67 | 0.000110483300951 |
| pcagw_compendium | 1.55039109506619 | 66 | 0.001117467439307 |
| dnarepair | 1.54926413964234 | 15 | 0.080471017287826 |
| cgc_germline | 1.50414250207125 | 10 | 0.151946556078459 |
| cgc_liquid | 1.49944841979726 | 28 | 0.033703719324945 |

Supplementary Table 7: **Cancer geneset enrichment analysis for the $16$ cancers in the UK Biobank.** Results of the enrichment analysis between the curated cancer genesets and the heritability genes of the $16$ datasets with sufficient power in the UK Biobank. The table refers to the results in Figure 4 in the main text.

| gene | PS | SG | EIR | CRI | TPI | IM | A | GIM | EPCD | CCE | tsg | og | fusion |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XPO1 | P | | | | | | | | P | | 0 | 1 | 0 |
| TP63 | P | | | | | P | | | P | | 1 | 1 | 0 |
| SMAD2 | | P | | P | | S | | | S | | 1 | 0 | 0 |
| ROS1 | P | | | | | | | | | | 0 | 1 | 1 |
| RAP1GDS1 | P | | | | | P | | | | | 0 | 1 | 1 |
| RABEP1 | | P | | | | | | | | | 0 | 0 | 1 |
| PPARG | P | P | | | | | | | | P | 1 | 0 | 0 |
| POT1 | | | | | | | | S | | | | 0 | 0 |
| PIK3R1 | | P | | S | | S | | | | | 1 | 0 | 0 |
| PBX1 | | | | | | | P | | P | P | 0 | 1 | 1 |
| PBRM1 | | P | P | | S | S | | S | S | P | 1 | 0 | 0 |
| NT5C2 | P | | | | | | | | P | | 0 | 1 | 0 |
| NCOR2 | | P | | | | | | S | P,S | | 1 | 0 | 0 |
| NAB2 | | | | | | | S | | | | 1 | 0 | 1 |
| MTOR | P | | | | | P | P | | P | P | 0 | 1 | 0 |
| MLLT10 | | | | P | | | | | | | 0 | 1 | 1 |
| LRP1B | | P | | | | S | | | | | 1 | 0 | 0 |
| JAK2 | P | | | | P,S | | | | P | P | 0 | 0 | 0 |
| FOXA1 | P | | | | | S | | | | | 0 | 1 | 0 |
| FGFR2 | P | | | | | | | | P | | 1 | 1 | 0 |
| FAT4 | | P | | | | S | | | | | 1 | 0 | 0 |
| ESR1 | P | P | P | | | P,S | | | | | 1 | 1 | 1 |
| ERBB4 | P | P | | | | | | | P,S | | 1 | 1 | 0 |
| EBF1 | | P | | | | | | | | | 1 | 0 | 1 |
| CTNNB1 | P | P | P | P | | P | P | S | P | P | 0 | 1 | 1 |
| CLIP1 | | | | | | | | | | | 0 | 0 | 1 |
| CIITA | | | S | | | | | | | | 1 | 0 | 1 |
| CDKN2A | | P | | | | S | S | | S | | 1 | 0 | 0 |
| CDH11 | | | | | | S | | | S | | 1 | 0 | 0 |
| CCDC6 | | P | | | | | | S | S | P | 1 | 0 | 1 |
| CBFA2T3 | | P | | | | | | | | P | 1 | 0 | 1 |
| ALK | P | | | | | P | | | P,S | | 0 | 1 | 1 |
| LATS2 | | P | | | | P,S | | S | S | | 1 | 0 | 0 |

Supplementary Table 8: **Cancer heritability genes associated with the hallmark of cancers across** $16$ **cancers in the UK Biobank**. Each column corresponds to one of the hallmarks. P stands for promotes, S stands for suppresses. We also report whether the gene is known to be a tumor suppressor, TSG, and oncogene or fusion gene. This table corresponds to the results in Figure 4 in the main text. **PS**: proliferative signalling, **SG**: suppression of growth, **EIR**: escaping immunic response to cancer, **CRI**: cell replicative immortality, **TPI**: tumour promoting inflammation, **IM**: invasion and metastasis, **A**: angiogenesis, **GIM**: genome instability and mutations, **EPCD**: escaping programmed cell death, **CCE**: change of cellular energetics, **tsg**: tumor suppressor gene, **og**: oncogene.

## References


[1] 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.

[2] Tian Ge et al. “Phenome-wide heritability analysis of the UK Biobank”. In: *PLoS Genetics* 13.4 (2017), pp. 1–21. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006711.

[3] 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.

[4] Zhan Su, Jonathan Marchini, and Peter Donnelly. “HAPGEN2: Simulation of multiple disease SNPs”. In: *Bioinformatics* 27.16 (2011), pp. 2304–2305. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr341. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150040/.