Machine learning and large scale cancer omic data: decoding the biological mechanisms underpinning cancer.

Viola Fanfani

THE UNIVERSITY
of EDINBURGH

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

to the

University of Edinburgh

2021

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

December 30, 2021

Ai miei genitori. A loro devo tutto.

"I should dearly like to resuscitate one or two of the rascals,
just to know what they would think when they saw all going on as before,
in spite of the disappearance of the human race.
Would they then imagine that everything was made
and maintained solely for them?"

Giacomo Leopardi
"Dialogue between a goblin and a gnome"

Abstract

Many of the mechanisms underpinning cancer risk and tumorigenesis are still not fully understood. However, the next-generation sequencing revolution and the rapid advances in big data analytics allow us to study cells and complex phenotypes at unprecedented depth and breadth. While experimental and clinical data are still fundamental to validate findings and confirm hypotheses, computational biology is key for the analysis of system- and population-level data for detection of hidden patterns and the generation of testable hypotheses.

In this work, I tackle two main questions regarding cancer risk and tumorigenesis that require novel computational methods for the analysis of system-level omic data. First, I focused on how frequent, low-penetrance inherited variants modulate cancer risk in the broader population. Genome-Wide Association Studies (GWAS) have shown that Single Nucleotide Polymorphisms (SNP) contribute to cancer risk with multiple subtle effects, but they are still failing to give further insight into their synergistic effects. I developed a novel hierarchical Bayesian regression model, BAGHERA, to estimate heritability at the gene-level from GWAS summary statistics. I then used BAGHERA to analyse data from 38 malignancies in the UK Biobank. I showed that genes with high heritable risk are involved in key processes associated with cancer and are often localised in genes that are somatically mutated drivers.

Heritability, like many other omics analysis methods, study the effects of DNA variants on single genes in isolation. However, we know that most biological processes require the interplay of multiple genes and we often lack a broad perspective on them. For the second part of this thesis, I then worked on the integration of Protein-Protein Interaction (PPI) graphs and omics data, which bridges this gap and recapitulates these interactions at a system level. First, I developed a modular and scalable Python package, PyGNA, that enables robust statistical testing of genesets' topological properties. PyGNA complements the literature with a tool that can be routinely introduced in bioinformatics automated pipelines. With PyGNA I processed multiple genesets obtained from genomics and transcriptomics data. However, topological properties alone have proven to be insufficient to fully characterise complex phenotypes.

Therefore, I focused on a model that allows to combine topological and functional data to detect multiple communities associated with a phenotype. Detecting cancer-specific submodules is still an open problem, but it has the potential to elucidate mechanisms detectable only by integrating multi-omics data. Building on the recent advances in Graph Neural Networks (GNN), I present a supervised geometric deep learning model that combines GNNs and Stochastic Block Models (SBM). The model is able to learn multiple graph-aware representations, as multiple joint SBMs, of the attributed network, accounting for nodes participating in multiple processes. The simultaneous estimation of structure and function provides an interpretable picture of how genes interact in specific conditions and it allows to detect novel putative pathways associated with cancer.

Lay Summary

Research has shown that tumors are caused by changes in the DNA sequence that result in cells that grow uncontrollably. Indeed, the DNA, a long sequence of nucleotides, stores and maintains the genetic content of an organism, which encodes all information required for the activities of the cell. The flow of information that allows cells, and organisms, to function, is simplified by the central dogma of biology: DNA makes RNA, and RNA makes proteins. Peculiarities, modifications and errors in the DNA sequence can affect the whole cascade of processes that regulate the machinery of an organism by changing the structure and function of RNA and proteins, the actual effector molecules. While we know that some of these events lead to tumorigenesis, the exact mechanisms that drive the formation of tumors are still not fully understood.

Thanks to the dramatic technological improvements of the last 20 years, we are now able to study the DNA of thousands of cancer patients and develop computational methods instrumental to gain insights into genetic variability and its impact on tumor susceptibility. In this work, I tackle two main questions regarding cancer risk and tumorigenesis that required the development of novel computational methods for the analysis of cancer data.

First I focused on how germline variants, variations of the DNA sequence inherited from parents, influence the risk of developing cancer. Indeed, we know that some rare inherited pathogenic DNA variants lead to cancers that recurrently occur in families, for example, the BRCA1 mutations associated with a higher risk of breast cancer. However, studies carried out on thousands of people have shown that DNA variants frequently observed in the population can also modulate cancer risk. Thus, I developed a novel statistical method, called BAGHERA (Bayesian Gene Heritability Analysis) that estimates where in the DNA (i.e., into which genes) the variants increasing cancer risk are more likely to be found. Then, knowing which genes are often affected, we can investigate what are the systems they disrupt and what other factors might concur to increase the risk of a tumor.

BAGHERA studies the location and impact of DNA variability in isolation. Nonetheless, most biological processes require the interplay of multiple proteins and regulatory elements where multiple smaller disruptions are more likely to go unnoticed and generate a systemic problem, the tumor. Detecting the combinations of issues that lead to cancer is non-trivial, but understanding them is key to learning how tumors arise and how we can treat them.

In the second part of this thesis, I illustrate how novel computational methods can be used to map DNA variations associated with tumors onto protein interaction networks. Indeed, biological experiments have reconstructed the networks describing how proteins interact with each other and we can use them in order to identify the groups, or modules, of genes that are responsible for tumorigenesis. I developed a software, PyGNA, that enables researchers to systematically map their experimental results on protein networks and analyse their connectivity properties; knowing the structure and communication patterns of the network helps investigate how cancer cells rewire. Moreover, I propose

a novel artificial intelligence method, that leverages data from multiple experiments, to predict which genes are more likely to drive tumors and how they are organised in the networks. Combining the connectivity information of the network with functional information of each gene improves our ability to detect the modules of proteins associated with cancer.

Taken together, in this thesis I show that new computational methods can elucidate some of the mechanisms underpinning cancer by aggregating multiple lines of evidence and considering the cell system with its disruptions as a whole.

Acknowledgements

First, I would like to acknowledge the people who made this thesis possible.

Giovanni Stracquadanio, for the constant supervision, support, and guidance in this new world of doctoral studies. I owe him a lot.

Luca Citi, for helping me with this thesis, but to whom I am immensely grateful for allowing me to study abroad and believing in me when I really needed it.

Guido Sanguinetti, for his guidance and advice in Edinburgh.

Adrian L. Harris and Francesco Pezzella for the incredibly helpful insights into cancer biology.

Pietro Liò, for his help, I really value the time he dedicated to me.

Angelo Gaeta, for patiently explaining biology to me and for the fruitful conversations in front of our screens; but also just Angelo, my brother-in-arms, for making life bearable in some really unfun times.

I would also like to thank those who helped me along the way, we all need more help to navigate life than books.

Ana, a true friend and the biggest gift of this PhD. This page is not enough to thank her.

Ash, Davide, Dimitri, Maria, Miguel who were of great company and made my life in Essex much more enjoyable. David, for some really unexpected fun times.

All my ‘italian’ friends. They are my anchor and safe harbor.

My family. I would be lost without them.

Contents

Acronyms

BAGHERA Bayesian gene-level heritability analysis.

CSG Cancer Susceptibility Gene.

GAT Graph Attention Network.

GCN Graph Convolutional Network.

GNN Graph Neural Network.

GWAS Genome Wide Association Studies.

MLP Multi Layer Perceptron.

NGS Next Generation Sequencing.

NN Neural Network.

PPI Protein-Protein Interaction.

PRS Poligenic Risk Score.

SBM Stochastic Block Model.

SNP Single Nucleotide Polymorphism.

UKBB UK Biobank.

1 Introduction

1.1 Background

Estimates of 2020 forecast that around one in five individuals worldwide, and one in three in Europe, will develop a neoplasm during their lifetime [Sung et al. 2021; Ferlay et al. 2020]. Similarly, the World Health Organization (WHO) has estimated an incidence of around 19.3 million of newly developed tumors in 2020, which is expected to rise to 30.2 million in 2040. In regards to mortality, the WHO reports that out of a total of 50 million deaths worldwide, around 9 million were due to a neoplasm [WHO 2020].

It is undebatable that cancer poses a considerable burden onto single individuals and health systems, and it comes with no surprise that a large part of the research worldwide has been focused on cancer. Today, hundreds of studies a year produce a deluge of data on tumors, with cancer being the main topic of about 16% of PUBMED entries [Reyes-Aldasoro 2017]. Nonetheless, many aspects of cancer risk, development, and progression are still far from being fully understood. Fortunately, in this somber landscape, both biological and computational research are undergoing fast-paced improvements.

The Next Generation Sequencing (NGS) revolution has allowed studying cells at an unprecedented level of depth and breadth. Since the release of the first human genome draft [Craig Venter et al. 2001; Lander et al. 2001], up until the recently published Telomere-to-Telomere human reference genome [Nurk et al. 2021], sequencing has become increasingly less expensive, allowing to collect population-level data, and novel experimental protocols and technologies have flourished [Reuter, Spacek, and Snyder 2015; Aslam et al. 2017; Lowe et al. 2017]. Last year, for example, the PCAWG consortium has released whole-genome data for more than 2,658 cancer samples [Campbell et al. 2020] and one of the largest consortium of public health, the UK Biobank [Sudlow et al. 2015], is on track to release health records and sequencing data for 500,000 individuals. At the same time, we are now able to perform sequencing experiments at the single-cell level [Stuart and Satija 2019; Abascal et al. 2021], capturing not only the properties of cancerous tissues but also the detail and heterogeneity of cells within them.

On the other side, statistical and computational sciences have undergone dramatic advances as well. Technological improvements on the hardware, with faster, more powerful, local and distributed machines, have powered the 'big data' innovation process. Moreover, computational power has translated into a deluge of novel methods, tailored to tackle high-dimensional data. Specifically, machine learning [Alpaydin 2014; Hastie, Tibshirani, and Friedman 2009], which we broadly use here to encompass all statistical learning, artificial intelligence, and data science fields, harnesses large amounts of data, that can now be stored and processed, to optimise the underlying statistical model. Indeed, a common task of data analysis is to understand, and describe, the processes behind the observed examples. Complex, stochastic, non-fully characterised processes, are hard to approximate with mechanistic models. Conversely, machine learning benefits from 'big data' by using it to automatically generalize the properties of the datasets, being then able to detect hidden patterns that would otherwise go undetected. By now, machine learning is being ubiquitously applied to all fields, and biology is no exception [Ching et al. 2018; Zitnik et al. 2019].

While this can be sometimes overlooked, projects that harness and analyse population-level data would not be possible without the dramatic improvements of technologies such as high-performance computing and the methodological advances of machine learning [Berger, Peng, and Singh 2013; Stephens et al. 2015]. Hence we find ourselves at a peculiar, and exciting, intersection between the complexity that we face studying cancer and the potential of newly developed technologies.

Unsurprisingly, the combination of high-throughput experiments and computational sciences has led to novel insights into many aspects of cancer. The classical paradigm of tumorigenesis, see Fig. 1.1, describes it as a sequence of aberrations conferring selective advantage to cells that eventually undergo immortalization and grow uncontrolled [Vogelstein et al. 2013; Stratton, Campbell, and Futreal 2009]. A lot of effort has indeed been placed into detecting the events driving tumorigenesis and their functional effects. DNA sequencing has allowed identifying hundreds of germline and somatic aberrations that are associated with cancer phenotypes [Pleasance et al. 2010; Hoadley et al. 2018; Campbell et al. 2020; Sondka et al. 2018; Sud, Kinnersley, and Houlston 2017] distinguishing driver events from passengers and validating their tumorigenic effect [Bailey et al. 2018; Martínez-Jiménez et al. 2020]. Moreover, functional genomics studies have revealed the impact of aberrations by identifying their effects on gene expression [Cieślik and Chinnaiyan 2018], cell trajectories and tissue heterogeneity [Rozenblatt-Rosen et al. 2020], and the epigenetic marks regulating tumor expression [Dawson 2017].

Most of the experiments described above are using single-modality snapshots of tumors and normal cells. While extremely fruitful, these studies have also revealed the polygenicity and heterogeneity of cancer; we are still unable to detect driver events for all patients or to fully understand the effects of known aberrations and their combinations since they can be peculiar to the environmental stimuli, tissue, and genetic background [Campbell et al. 2020; Martínez-Jiménez et al. 2020; Abascal et al. 2021]. However, while the specific aberrations might be tumor-dependent, driver events tend to recurrently hit genomic elements involved in hallmark processes of cancer [Hanahan and Weinberg 2011]. In this context, system biology, that is the comprehensive and integrative study of complex biological systems, has the potential to reveal the processes underpinning cancer. Integrative studies have already led to novel insights into cancer biology; among others, whole-genome co-essentiality maps, that capture gene-gene funtional interdependency and actual pathway-level gene cooperation [Wainberg et al. 2021], numerous therapeutic opportunities directly targeting disrupted proteins or the processes they control [Hahn et al. 2021], novel cancer drivers prioritised by functional annotation [Reyna et al. 2020].

In this context, the goal of this thesis is to develop novel methods to elucidate the mechanisms underpinning cancer; by using large-scale omic data and novel machine learning methods, this work focuses on the interplay between different tumorigenic events. Indeed novel computational methods for the analysis of high-dimensional multi-modal datasets are instrumental to gather a three-dimensional picture of tumorigenesis and generate novel, testable, hypotheses, see Fig. 1.1. Specifically, we first address the question of cancer risk in the broader population. We then focus on the interplay between genes that participate in cancer-associated phenotypes by integrating omics data onto biological networks. In the next section, we briefly introduce and discuss the outline of the thesis, highlighting the motivation and the background of the main topics we tackled.

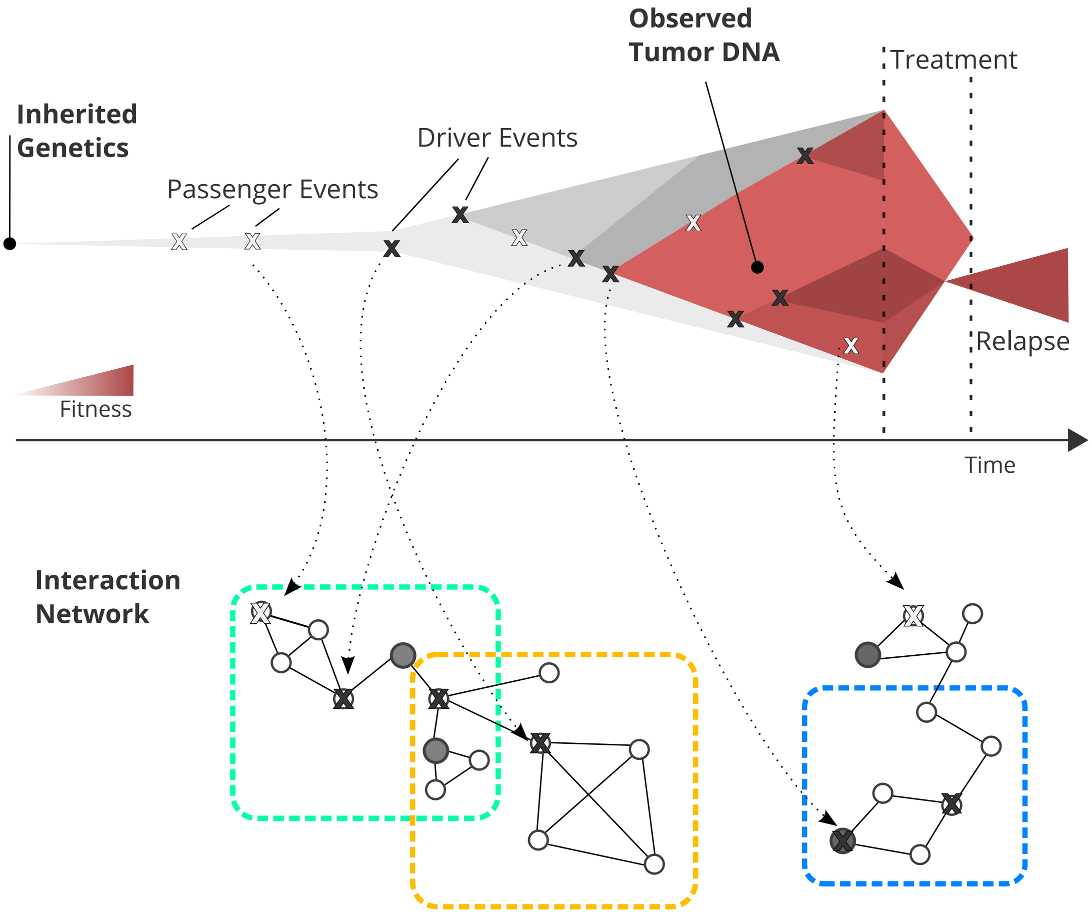


Figure 1.1: *The timeline of tumorigenesis. Inherited mutations shape the genetic background of all tissues. Driver events (dark crosses) confer selective advantage to the cell that then generates clonal populations (expanding grey areas) that have metabolisms enhancing replication and are more likely to escape cell death. At some point, a cancer-driving event, a point mutation or a structural variation, occurs and triggers cancer (expanding red areas), which leads to cell immortalization and uncontrolled proliferation. Available treatments, either drugs or physical procedures might be able to attack the tumor, causing it to disappear in some cases, or to regress before a relapse. All these events, however, are not functionally isolated as they occur in genes -or any functional genomic element- that interact with each other. By mapping (dashed arrows) the driver events (crosses) and the underlying genetics (darker network nodes) to the network we can explore the disrupted pathways (colored dashed boxes) and their functional relevance. Figure adapted from [Campbell et al. 2010] and Prof. Getz's lectures*

1.2 Thesis Outline

Familial and targeted sequencing studies have been able to detect and characterise the effects of high-penetrance germline variants for cancer risk [Miki et al. 1994; Wooster et al. 1994; Anderson 1974; Lynch and Chapelle 2003]. These works have shed light on the role of germline mutations for cancer predisposition, and have been instrumental in finding genetic markers for hereditary cancer syndromes [Foulkes, Knoppers, and Turnbull 2016; Foulkes 2008; Southey et al. 2016]. However, the rare or uncommon mutations responsible for early-onset hereditary malignancies do not explain how the genetic background mediates cancer risk in the broader population. Genome Wide Association Studies (GWAS) have been able to detect many frequent germline variants associated with cancer, nonetheless, the functional characterisation of their effects is still inadequate and often proves challenging [Sud, Kinnersley, and Houlston 2017; Lawrenson et al. 2015].

The observation that cancer is a polygenic disease [Boyle, Li, and Pritchard 2017; Mavaddat et al. 2019], that is it requires multiple aberrations targeting different biological processes to develop, and the mounting evidence of subtle effects of germline variants [Zhang et al. 2020; Whitington et al. 2016; Dimitrakopoulos et al. 2019], sustains our interest in studying how multiple low-penetrance variants might be mediating cancer risk alongside somatic mutations. In chapter 2 we present the state-of-the-art of breast cancer GWAS and describe results and limitations of current literature. Moreover, we review the main methodological approaches to the study of GWAS heritability, the additive effects of all inherited Single Nucleotide Polymorphisms (SNP) to the phenotypic variance. Indeed, we find that heritability is an appropriate measure of inherited cancer risk as it aggregates subtle germline effects and can be applied to link the genetic variations to their functional effects.

In chapter 3 we present BAGHERA, a novel statistical learning method to estimate heritability at the gene level. We believe that apportioning inherited cancer risk to functional loci, such as genes and cis-regulatory regions, might effectively capture the biological mechanisms underpinning cancer. Indeed, we found that the genes whose heritability is higher than expected by chance are preferentially involved in the hallmark processes of cancer and we hypothesise an interplay between germline and somatic mutations for tumorigenesis.

Our analysis of gene heritability studies the combined effects of multiple variants on the same gene, however, it does not explicitly take into account the interaction between genes. Pathway analysis methods [Ma, Shojaie, and Michailidis 2019; Khatri, Sirota, and Butte 2012], those that statistically test for an overrepresentation of a candidate set of genes into known pathways and biological processes, allow mapping aberrations onto a functional representation of cellular machinery. While extremely insightful, these methods are applied downstream of the analysis and are biased towards known pathways, hindering the chances of revealing novel ones.

Conversely, biological networks describe the interplay between genomic elements and can be used to detect novel interaction patterns [Kuenzi and Ideker 2020]. Graph-structured data have been used in a variety of system biology applications, leading to some pivotal observations about cancer drivers [Reyna et al. 2020], druggable master regulators driving tumor progression [Hahn et al. 2021], drug repositioning [Gysi et al. 2021].

Chapters 4 and 5 discuss how experimental data can be mapped onto Protein-Protein Interaction (PPI) networks that are large-scale graphs recapitulating known physical interactions between proteins. Although incomplete and lacking tissue-specificity, PPI networks can be combined with other datasets to infer pathways, communities, or subgraphs, of genes that are synergistically affecting the phenotype. In chapter 4 we describe the background literature on PPIs and the state-of-the-art methods for the integration of omics data onto the graph structure. We focus on graph topology, that is the connectivity properties of the genes in the network, as a way to summarize and detect the interplay within and between genesets. As we identified the lack of scalable computational tools to analyse the topological properties of genesets obtained from high-throughput experiments, we present PyGNA, a modular Python package, integrable into existing bioinformatics pipeline, that implements multiple statistical tests for graph topology.

Topological properties alone are insufficient to fully capture the pathways underlying cancer connectivity: genes involved in complex phenotypes are targeting multiple processes and might not be all closely linked to each other [Agrawal, Zitnik, and Leskovec 2018]. Moreover, most state-of-the-art methods are better suited for attributed networks with a single attribute, one score per node that summarises the observed data. In chapter 5 we present a deep learning method that uses both graph structure and multiple gene features to integrate topology and experimental evidence. We use Graph Neural Networks (GNN) to simultaneously learn gene properties from graph structure and experimental observations. Within the deep learning architecture, we also use the Stochastic Block Model to infer communities in the network, which can be then readily used to describe the main pathways involved in cancer processes. Our model, namely SBM-GNN, predicts cancer driver genes while organising them into communities of both drivers and non-drivers and it is a promising stepping stone for the identification of novel putative cancer-implicated pathways.

2 Cancer risk in the broader population

2.1 Introduction

Tumors have very complex genetic patterns arising from both inherited and acquired mutations. Once we analyse the tumor DNA, we are observing the results of the tumor clonal evolution, the DNA damage subsequent to cancer-driving aberrations, on top of the genetic background and passenger events [Gerstung et al. 2020]. Discriminating the driving, mediating, or innocuous effects is a difficult task, nonetheless, it is critical to understand cancer risk and tumorigenesis.

Research has, so far, been able to decode some of these events underpinning cancer. For instance, the most frequent mechanism of tumorigenesis in adults is an accumulation of somatic mutations, due to environmental exposure to cancer risk factors, which culminates in a catastrophic event that triggers cancer. In many cases, a coding mutation occurs within a master regulator and it impairs critical cellular functions, like those controlled by the p53 tumor suppressor pathway or those that regulate cell survival like *KRAS* or *MYC*

[Vogelstein et al. 2013].

Insights into environmental risk factors [U.S. Department of Health and Human Services 2016] or into mutational and transcriptional signatures [Alexandrov et al. 2013; Bernard et al. 2009] have been instrumental for cancer prevention and patient treatment. However, these somatic events do not recapitulate the whole landscape of tumorigenesis. Evidence for the causal role of low frequency highly penetrant inherited variants in familial and cancer syndromes has been identified more than 30 years ago [Anderson 1974; Miki et al. 1994; Wooster et al. 1995]. These are rare inherited mutations in cancer susceptibility genes (CSG) [Rahman 2014] that directly increase the risk of cancer in first-degree relatives, but they do not explain cancer risk in the broader population. In other cases, we have evidence that mutations in cancer driver genes did not trigger tumorigenesis, reinforcing the hypothesis that further cancer risk factors might explain the complexity and heterogeneity of tumorigenesis events [Martincorena et al. 2015; Moore et al. 2020].

Frequent germline variants contribute to the genetic background of all tumors, but much of their contribution to cancer risk is still unexplored. Single Nucleotide Polymorphisms (SNP) are inherited point variants frequently found in the population (MAF $> 1\%$). Consortia are now able to genotype, relatively inexpensively, SNPs in large cohorts and to carry out Genome Wide Association Studies (GWAS) that look for associations between frequent genetic variants and phenotypes.

In the last 20 years, GWAS have been carried out on a broad spectrum of traits [Visscher et al. 2017] and they have led to the identification of many SNPs associated with increased risk of cancer [Sud, Kinnersley, and Houlston 2017]. SNPs with a significant association with a trait or disease are usually reported in the GWAS catalog [MacArthur et al. 2017]. Currently, around 167 studies,

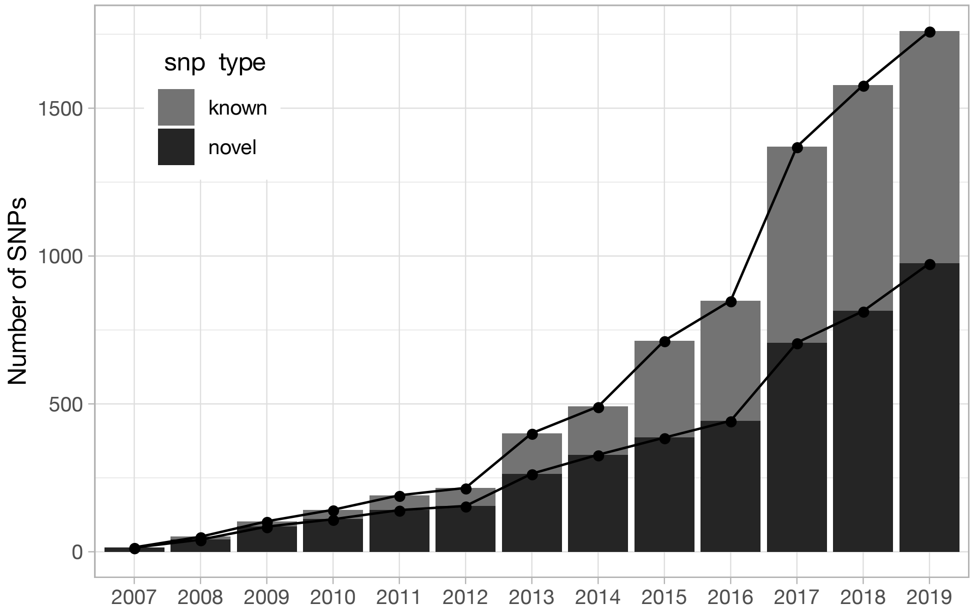


Figure 2.1: *Cumulative number of cancer GWAS hits, SNPs, reported in the GWAS catalog [MacArthur et al. 2017] each year. Reported SNPs are either novel findings, included in the catalog for the first time (dark bar), or known risk loci, already in the catalog at the time of publication (light bar). We do not report here data for 2020 and 2021, as they are incomplete, but the number of SNPs in the catalog has reached 1225.*

carried out since 2007, have reported 1225 SNPs significantly associated (p-value $< 5\times10^{-8}$) with malignancies in European populations, see Fig. 2.1. While these loci are significant genome-wide, their effects are subtle, with average odds ratio (OR) 1.66, see Fig. 2.2, ranging from 1.02 to 2.69 for rs995030 which is a well-known association locus for testicular germ cell tumors [Rapley et al. 2009; Ruark et al. 2013]. Moreover, the vast majority of them reside in non-coding regions, with almost 80% of them located in either intron or intergenic regions, see Fig. 2.2.

The subtle effect sizes and the unclear functional effects render testing the impact of cancer SNPs very challenging. Indeed, complex phenotypes are polygenic [Boyle, Li, and Pritchard 2017], that is they arise from the combination of multiple molecular events that are encoded by many loci in the genome. This

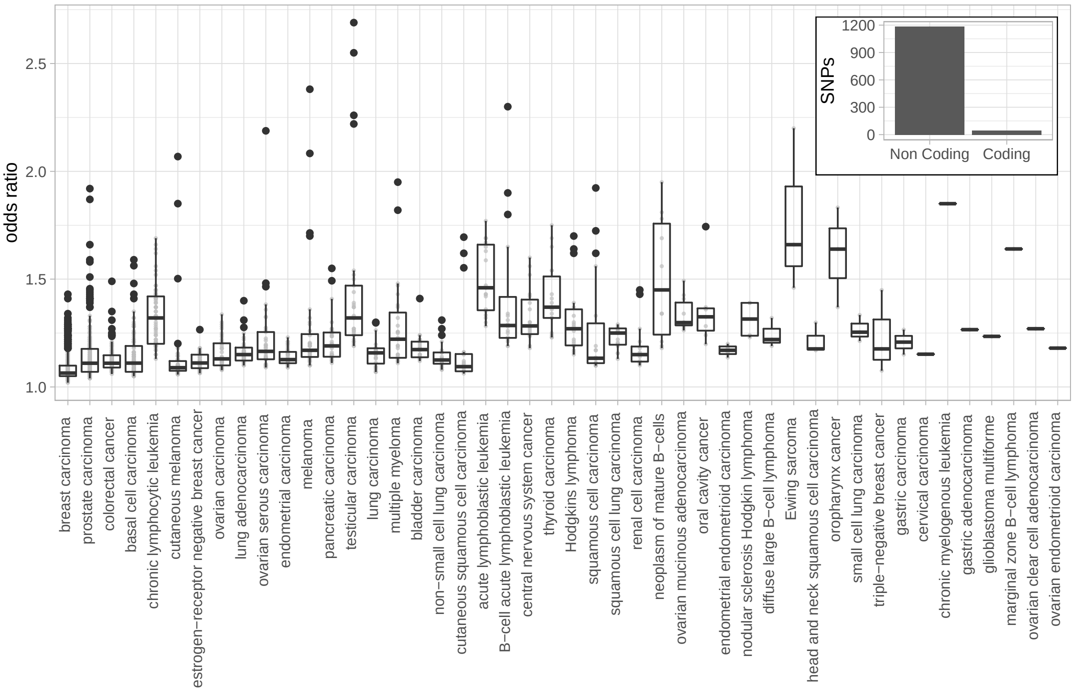


Figure 2.2: *Odds Ratios of all reported SNPs in the GWAS catalog [MacArthur et al. 2017] for each malignancy, sorted by the number of reported hits. The vast majority of SNPs have OR* < 2*, with the most of median values below* 1.5*. In the inset (top right) we show the reported functional annotation with more than* 80% *of them being in non-coding regions.*

is often due to the effects that the variants have on regulatory elements [Li et al. 2016; Lawrenson et al. 2015], and the combination of GWAS and functional annotation has shown interesting results for the prioritization of variants [Pickrell 2014].

In the paper below, we provide a detailed account of the state-of-the-art cancer risk SNPs. We focus on breast cancer, which is one of the most frequent malignancies, to gather a detailed map of known cancer risk loci in the broader population. We explore the potential functional effects of significant SNPs and pinpoint multiple genes whose effect could be mediated by inherited germline variants.

We also introduce GWAS heritability, the amount of risk due to genetic effects, as a method to account for the effects of all SNPs, regardless of their statistical significance. While GWAS detect statistically significant associations between single SNPs and cancer, it is reasonable to expect a non-null contribution to cancer risk of all the other variants. We give a detailed overview of the state-of-the-art methods to study cancer heritability and the available estimates of heritability for breast cancer.

2.2 Dissecting the heritable risk of breast cancer: from statistical methods to susceptibility genes

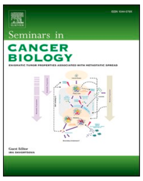
The whole manuscript has been drafted and revised by V. Fanfani, with the supervision of G. Stracquadanio. V. Fanfani reviewed the literature on GWAS and heritability and produced all the figures. The details on gene function and putative mechanisms for breast cancer tumorigenesis were revised by M. Zatopkova, A.L. Harris, and F. Pezzella.

Seminars in Cancer Biology 72 (2021) 175–184

Contents lists available at ScienceDirect

Seminars in Cancer Biology

journal homepage: www.elsevier.com/locate/semcancer



Review

Dissecting the heritable risk of breast cancer: From statistical methods to susceptibility genes

Viola Fanfani [a], Martina Zatopkova [b], Adrian L. Harris [c], Francesco Pezzella [d], Giovanni Stracquadanio [e],*

[a] Institute of Quantitative Biology, Biochemistry, and Biotechnology, SynthSys, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, UK
[b] Department of Clinical Studies, Faculty of Medicine, University of Ostrava, Ostrava, Czech Republic
[c] Molecular Oncology Laboratories, Department of Oncology, The Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK
[d] Nuffield Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Oxford, UK
[e] Institute of Quantitative Biology, Biochemistry, and Biotechnology, SynthSys, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, UK

ARTICLE INFO

Keywords:
Cancer
Heritability
GWAS
SNP
Cancer risk

ABSTRACT

Decades of research have shown that rare highly penetrant mutations can promote tumorigenesis, but it is still unclear whether variants observed at high-frequency in the broader population could modulate the risk of developing cancer. Genome-wide Association Studies (GWAS) have generated a wealth of data linking single nucleotide polymorphisms (SNPs) to increased cancer risk, but the effect of these mutations are usually subtle, leaving most of cancer heritability unexplained. Understanding the role of high-frequency mutations in cancer can provide new intervention points for early diagnostics, patient stratification and treatment in malignancies with high prevalence, such as breast cancer.

Here we review state-of-the-art methods to study cancer heritability using GWAS data and provide an updated map of breast cancer susceptibility loci at the SNP and gene level.

1. Introduction

Breast cancer is the most frequent cancer among women worldwide, representing approximately one third of all diagnosed malignancies. Breast cancer has a cumulative risk of 5%, that is 5 in 100 newborns are expected to develop this malignancy during their lifetime. While the survival in first-world countries is usually very high, about 70% of all cases, breast cancer was still responsible for more than 600,000 deaths in 2018 [1,2].

The mechanisms affecting cancer predisposition, tumorigenesis and progression are still unclear; in the majority of cases, tumors are triggered by the accumulation of somatic mutations, which impair critical cellular functions, like those controlled by the p53 tumor suppressor pathway [3]. While the causal role of somatic mutations has been confirmed by in-vitro and in-vivo models, there is limited understanding of whether inherited mutations mediate the risk of developing cancer. Familial and cancer syndrome studies have shown a causal role of inherited variants; usually, low frequency highly penetrant variants in cancer susceptibility genes (CSG), directly increase the risk of cancer in first-degree relatives. In particular, breast cancer has been one of the first malignancies for which evidence of inheritance has been found and whose CSGs have been identified [4], including the well known BRCA1/2 genes [5,6].

However, rare mutations explain only a small fraction of the risk of cancer in the broader population, suggesting that cancer risk could be somehow mediated by high-frequency low-penetrance mutations, such as single nucleotide polymorphisms (SNPs). Recent advances in high-density genotyping arrays and DNA sequencing technologies allow genotyping SNPs in large cohorts, paving the way to population-scale Genome Wide Association Studies (GWAS). Currently, more than 100,000 SNP alleles have been associated with various traits and diseases, of those around 5000 variants are associated with various tumor types, including breast cancer [7].

The contribution of germline mutations to the inherited risk of cancer is estimated through heritability analysis. Heritability estimates for cancer have been usually obtained through familial studies; however, these estimates have not been replicated when analysing inherited mutations in the broader population, thus leading to the concept of missing cancer heritability [8]. Missing heritability could be apportioned to a number of factors, including structural variants, gene–gene

* Corresponding author.
E-mail address: giovanni.stracquadanio@ed.ac.uk (G. Stracquadanio).

https://doi.org/10.1016/j.semcancer.2020.06.001
Received 15 November 2019; Received in revised form 1 June 2020; Accepted 2 June 2020
Available online 20 June 2020

and gene–environment interactions, as well as rare highly penetrant variants [9,10]. Ultra rare variants, which are difficult to detect with current technologies, have also shown to have a significant role in complex diseases [11]. However, even accounting for rare highly penetrant variants and genome-wide significant SNPs, the difference in risk between individuals is not completely explained [12].

There is strong evidence suggesting that the risk of complex diseases, such as cancer, can be explained by the co-inheritance of a large number of frequent variants with subtle effects [13]. In this case, we consider a disease to be polygenic [14], thus we are interested in quantifying the contribution of low-penetrance inherited mutations to cancer risk. Here, we focus on narrow sense heritability, h^2, that is the cumulative effect of all loci on the phenotype variance [15]. Interestingly, using GWAS data, we can estimate the heritability explained by SNPs regardless of their statistical significance. Heritability analysis is becoming a crucial step in recent cancer GWAS analyses, providing insights on the inherited risk of many malignancies, including prostate [16,17], cervical [18], testicular germ cell tumor [19], and breast cancer [20].

Here we aim at providing an overview of state-of-the-art methods to estimate the amount of heritability explained by SNPs and an updated reference of the genetic architecture of breast cancer at the SNP and gene level. We organised this review as follows; in Section 2, we introduce common notation and standard statistical analyses performed in GWAS, and we then present state-of-the-art methods for the estimation of heritability. Finally, in Section 3, we systematically characterize current GWAS data available for breast cancer, and propose a curated resource of SNPs and genes that can be used for further investigations.

2. Estimating the risk of cancer explained by high-frequency inherited mutations

DNA sequencing technologies have enabled the discovery of thousands of rare and common variants that are associated with complex traits and diseases. While high-throughput whole genome sequencing is now routinely used to detect both common and low-frequency mutations across relatively small cohorts (<10,000 individuals), cost-effective genotyping arrays allow to carry out genetic studies at a population scale, albeit limited to only known loci.

Population scale genotyping is pivotal to understand the role of high-frequency low penetrant inherited mutations as genomic modifiers controlling quantitative traits and disease risk in the broader population. While highly penetrant mutations are often identified in relatively small cohorts [21], quantifying the contribution of high-frequency but low penetrance mutations requires genotyping large number of individuals.

In the last 30 years, genome-wide association studies (GWAS) have identified thousands of SNPs associated with increased risk of many diseases. In this context, cancer is not an exception; GWAS have been carried out on a broad spectrum of malignancies leading to the identification of a plethora of SNPs associated with increased risk of cancer [22]. However, experimental and analytical challenges have limited GWAS contribution in understanding the mechanisms underpinning cancer heritability.

Since the focus of this review is on computational methods for cancer GWAS analysis, we will focus on the methodological limits of SNP association tests, rather then issues arising from different experimental designs. GWAS have also complex interpretability limits; in particular, since variants often reside in non coding genomic regions, associations between SNP genotype and a trait provides limited mechanistic insights.

Here, we will introduce methods for heritability analysis as a framework to dissect the contribution of SNPs to the heritable risk of a disease, focusing on how to use these methods to study cancer heritability.

2.1. Tests of association

We refer to a single nucleotide polymorphism (SNP), as a locus where two or more distinct nucleotides are observed in a given population. Hereby, we assume SNPs to be bi-allelic, that is only 2 nucleotides are observed or considered at a given locus; this is a reasonable assumption for the vast majority of loci in the human genome.

We denote the most frequent nucleotide, as the major allele B, and the other as the minor allele, b. Since human cells are diploid, there are three possible genotypes, namely homozygous major (BB), heterozygous (Bb), and homozygous minor (bb).

For a binary phenotype, such as case–control studies, the association between the genotype and the disease status (e.g. 0: normal, 1: affected) can then be tested using a χ^2 test with 2 degrees of freedom. For each SNP, the test is carried out by comparing genotype counts in cases and controls, g_{ij}, with their expected value, $\widehat{g}_{\mathrm{ij}}$, as follows:

$$\chi^2 = \sum_i \sum_j \left[\frac{(g_{\mathrm{ij}} - \widehat{g}_{\mathrm{ij}})^2}{\widehat{g}_{\mathrm{ij}}} \right] \tag{1}$$

where i is the disease status, j is one of the three possible genotypes and $\widehat{g}_{\mathrm{ij}} = f_j N_i$, with f_j being the genotype frequency. While the above is the general formulation, theχ^2 association test can be adapted to different hypotheses and data [23].

Logistic regression can instead be used to account for confounders, like age or sex. For a GWAS with N individuals and M SNPS, a logistic regression model can be defined as follows:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{Y} : N \times 1$ is a binary vector encoding the disease status, $\mathrm{X} : N \times M$ is the genotype matrix, with x_{ij} being the number of minor alleles for the ith locus of jth individual.

Under the model in Eq. (2), $\boldsymbol{\beta}$ represents the effect-size of all SNPs and $\boldsymbol{\varepsilon}$ is the error introduced by confounders. In presence of other covariates, C_i, the regression is extended to include those terms, such that $Y = X\boldsymbol{\beta} + C_1\boldsymbol{\beta}_1 + C_2\boldsymbol{\beta}_2 + \ldots + \boldsymbol{\varepsilon}$. Under the null hypothesis of no association between the SNP and the disease, $\beta_j \sim N(0, \sigma^2)$; thus, the statistical significance of each effect-size can be tested using a Wald-test or a likelihood-ratio test. While the above formulations are useful to understand the idea behind association testing, in practice, these analyses require more complex models, which account for population biases, such as structure and relatedness, and genotype uncertainty.

While association analysis provides a mathematically tractable framework for testing whether a SNP genotype is associated with a trait, it is prone to false discoveries. This is largely due to SNP co-inheritance, a phenomenon usually referred to as linkage disequilibrium (LD); during meiotic crossing-over, proximal SNPs are more likely to be inherited together, resulting in a non-independence of their occurrence. From a statistical point of view, LD inflates the test statistic of the variants co-inherited with true causal SNPs, ultimately hindering the discovery of causal variants. LD for all genotyped SNPs in a GWAS can be represented as a lower-triangular matrix, $R : M \times M$, where r^2_{ij} is the LD between the alleles in SNP i and j. However, it is important to note that LD estimates are population-dependent and are biased by non-genotyped variants. Ultimately, finding causal variants usually requires integration of functional data to prioritize alleles within a given set of SNPs in LD [24].

To limit the number of false positives, GWAS studies usually apply a stringent family-wise error correction; in this context, empirical studies have concluded that 5×10^{-8} is a reasonable threshold to filter false positives out [13]. While for large populations and easily measurable phenotypes, such as height or blood pressure, it is possible to identify robust associations for a large number of loci, in cancer studies only a handful of SNPs pass correction for multiple hypotheses testing, resulting in the contribution of other loci with subtle effects to be neglected.

Thus, it is becoming apparent that methods able to estimate the cumulative contribution of multiple SNPs will be pivotal to maximize the information gained from GWAS. The rationale behind grouping SNPs together is based on the hypothesis that multiple variants in the same gene or pathway are more likely to have a stronger association with the phenotype regardless of their individual statistical significance. This is particularly true for cancer, whose inherited risk is thought to be mediated by a polygenic genetic architecture.

2.2. Estimating the heritable risk

Whenever referring to heritability, clarity is paramount; in GWAS analysis, inheritance does not refer to the amount of familial resemblance, rather to the effects of all inherited genomic loci to the phenotypic variance [25]. While broad sense heritability encompasses the effects of all genetic factors, narrow sense heritability accounts only for additive genetic effects. Thus, narrow sense heritability can be estimated from GWAS data as the cumulative contribution of all SNPs to the inherited risk.

Heritability in the narrow sense is defined as the portion of variance explained by the variance of the additive genetic effects, $h^2 = \frac{\sigma^2_{\mathrm{Add}}}{\sigma^2_P}$ [15]. The phenotype Y can be partitioned into two terms: a genotype term G and an environmental term E. The genotype contribution can be further partitioned into an additive genetic effect (add), a dominant genetic effect (dom) and an epistatic genetic effect (epi). Thus, a phenotype, Y, can be expressed as $Y = G + E = (\mathrm{add} + \mathrm{dom} + \mathrm{epi}) + E$.

By estimating heritability from GWAS data, we are assuming the variance of the phenotype P to be $\mathrm{Var}(P) = \mathrm{Var}(G+E)$. Assuming independence between the terms, the overall phenotype variance is explained by narrow sense heritability, other genetic factors, and environmental effects as follows:

$$\frac{\sigma^2_P}{\sigma^2_P} = h^2 + \frac{\sigma^2_{\mathrm{epi}} + \sigma^2_{\mathrm{dom}}}{\sigma^2_P} + \frac{\sigma^2_E}{\sigma^2_P} \tag{3}$$

GWAS can be used to estimate narrow sense heritability, since germline variants are accounting for additive genetic effects. However, the estimate obtained from the genotyped SNPs, h^2_{SNP}, is a lower bound of the narrow sense heritability, $h^2_{\mathrm{SNP}} \le h^2$, since the genotyped loci are usually a subset of all the variants in the genome. Hereby, we will refer to the term heritability as a synonym of narrow-sense heritability, which we will denote as h^2.

Advances in statistical genetics are leading to an increasing number of methods to estimate the heritability explained by all genotyped SNPs, a quantity we will refer to as genome-wide heritability. However, these methods provide limited insights into the genetic architecture of a disease. While the reasons for this stall are probably multifaceted, there are many challenges that affect the accuracy of heritability estimation methods. In general, we would like to measure the contribution of the SNPs to a binary trait, that is the disease status. However, many popular methods to estimate h^2 are working under the assumption of continuous traits. This problem is overcome by introducing the concept of liability [26]. Since most continuous traits can be approximated by a normal distribution, binary traits have been modelled by a liability threshold model; thus, there is the underlying assumption that disease risk follows a normal distribution, which represents the sum of many independent and normally distributed genetic and environmental effects. Thus, the binary phenotype represents whether the liability score exceeds a certain threshold t. Hence, in a normally distributed population, the quantile distribution function at t is the probability of the disease and is usually set from the observed prevalence in the population. In this framework, the observed value of heritability, h^2_{observed}, can be easily translated on the liability scale, $h^2_{\mathrm{liability}}$, as follows:

$$h^2_{\mathrm{observed}} = \frac{z(t)^2 h^2_{\mathrm{liability}}}{K(1-K)} \tag{4}$$

where K is the incidence, z is the standard Gaussian density.

Although mapping h^2 from the observed to the liability scale is straightforward, it is important to check whether the assumptions made by a method hold for the study under consideration. In particular, for many cancer types, the incidence can be extremely low and so are the values of h^2_{observed}; in both cases, the case-control ratio of the GWAS is incremented by design. While this procedure increments h^2_{observed}, thus making heritability detectable, it introduces a bias due to the difference between the real prevalence of the disease and the one in the cohort.

We now move forward describing methods to estimate heritability from GWAS data, highlighting their strength and weaknesses in the context of cancer GWAS analysis.

2.3. Methods for the estimation of genome-wide heritability

Estimates of genome-wide heritability can be obtained using a plethora of methods, each working under specific hypotheses, using different estimators, and requiring different input data. However, these methods estimate the heritability explained by genotyped SNPs and it is common to refer to this quantity as array-heritability, h^2_{array}, or SNP heritability, h^2_{SNP}.

Here we present state-of-the-art methods classified based on the required input, that is either genotype data or SNP summary statistics. Methods using raw data require genotype and covariates for each patient. Conversely, methods using summary statistics require only SNP test statistics and standard errors, along with population-level parameters that can be estimated from reference panels.

Here we describe methods using genotype data first as they are regarded as the gold-standard in the field; we then introduce those using summary statistics highlighting differences and advantages between the other class.

2.3.1. Estimating heritability from genotype data

Heritability is obtained by regressing the variance of the phenotype against the variance of the genotype as defined in Eq. (2).

To do that, the vast majority of methods regress h^2 using linear mixed models (LMM) [27–31]. The genomic-relatedness-based restricted maximum-likelihood approach (GREML, [27]), was the first to be introduced and it is routinely used for heritability studies. GREML uses genotype data with allele frequency as input and regress h^2 using restricted maximum-likelihood. GREML assumes that effect sizes β and errors ε in Eq. (2) are normally distributed with variance σ^2_g and σ^2_e, respectively. The variance of the phenotype then becomes:

$$\mathrm{var}[\boldsymbol{Y}] = \boldsymbol{G}\sigma^2_g + \boldsymbol{I}\sigma^2_e \tag{5}$$

where $G = XX^T/M$ is the genetic relationship matrix (GRM) between pairs of individuals at M loci.

This method has been extended to account for differences in allele frequencies and relatedness. GREML has also been applied to binary traits [29], transforming the observed heritability estimates on the liability scale h^2_l, following the procedure outlined in Eq. (4). However, this procedure should be used with caution when analysing cancer data, since GREML works under the assumption that the phenotype is normally distributed. While the liability model is a good approximation for diseases with high prevalence, REML assumptions do not hold when study prevalence does not match the true population prevalence; this leads to consistently biased estimates [32], thus suggesting that GREML-like approaches are not appropriate to analyse cancer data [33].

A second class of methods adapts the Haseman–Elston regression [34] to GWAS analysis, specifically focusing on case–control studies [35,32,36]. The Phenotype Correlation–Genotype Correlation method (PCGC) does not rely on normality assumptions, but instead obtains heritability estimates by considering the relationship between phenotypic and genotypic correlations between individual i and individual j. The phenotypic correlation, $\mathrm{E}(y_i y_j)$ can be written as a generic function of the heritabily and the genotypic correlation:

$$\mathrm{E}(y_i y_j) = f(h^2, G_{\mathrm{ij}}) \quad (6)$$

In its simplest formulation, considering only additive quantitative phenotypes and no specific study design confounders, $f(h^2, G_{\mathrm{ij}}) = h^2 G_{\mathrm{ij}}$ and h^2 can be estimated by least squares as follows:

$$h^2 = \mathrm{argmin} \sum_{i,j,i \neq j} [y_i y_j - h^2 G_{\mathrm{ij}}] \quad (7)$$

Case–control studies, extreme phenotypes, studies with related individuals are all modelled by using an appropriate $f(h^2, G_{\mathrm{ij}})$.

For binary phenotypes, the phenotypic correlation, $\mathrm{E}(y_i y_j)$, is accounted for to obtain estimates of heritability on the liability scale. The general consensus is that PCGC is better suited for binary phenotypes, being more robust to different covariates and cohort sizes.

Methods using genotype data are considered the gold-standard for heritability analysis and are readily available as part of many bioinformatics packages [37,38]. However, these methods require access to high-performance computing (HPC) infrastructures and genotype data; while HPC facilities are routinely found in academic and industrial environments, access to cancer patients' genotype is usually difficult, due to privacy concerns, thus limiting their use in practice.

2.3.2. Estimating heritability from summary statistics data

There has been an increasing interest in estimating heritability using GWAS summary statistics to overcome the limitations imposed by methods requiring genotype data [39–43]. Summary statistics are usually publicly available, since genotype information cannot be traced back from regression weights, and the analysis is not computationally taxing. Here we review how genome-wide heritability can be estimated from GWAS summary statistics.

The most widely used approach to estimate heritability from summary statistics is the LD score (LDSC) regression method [39,44]. LDSC computes heritability estimates by regressing h^2 as follows:

$$\mathrm{E}\left[\chi_j^2\right] = \frac{N}{M} h^2 l_j + \mathrm{N\ a} + 1 \quad (8)$$

where χ_j^2 is the summary statistic of the jth SNP for a GWAS with N individuals and M variants. Here l_j is a quantity called LD score, computed as $l_j = \sum_{i=0}^{K} r_{\mathrm{ij}}^2$, that is by summing up the correlation coefficients of all the SNPs in a window of prefixed size from the jth variant. Here, $\mathrm{Na} + 1$ is a term introduced to account for confounding bias, which can be estimated as the intercept of the linear regression between the LD score of each variant and its test statistic. The heritability is regressed using reweighted least squares, where the weights are adjusted to account for heteroscedasticity of the test statistic. LDSC is also implemented as part of the SumHer software, which improves the original LDSC model by taking into account allele frequency [42].

Recently, PCGC has also been extended to take summary statistics in input (s-PCGC, [41]). It has been shown that LDSC and s-PCGC are almost equivalent in absence of covariates with strong effects [45], although s-PCGC is recommended in presence of effects that could severely skew the liability distribution.

The methods discussed so far have been shown to be sensitive to the input data and trait properties, e.g. low or high heritability, low or high disease prevalence. This is due to the assumptions made by each model, which must be carefully considered as part of the analysis [46,47,45].

Although estimating heritability from genotype data usually leads to more accurate estimates, summary statistics proved to be sufficient to obtain accurate heritability estimates across a number of phenotypes [48]. Moreover, the negligible computational burden of summary statistics methods has made them the preferred approach for population scale studies [49] and the steppingstone to estimate the heritability of SNP groups.

Nonetheless, genome-wide heritability analyses have major interpretability limits. The estimate of h^2 gives a measure of the contribution of all genotyped SNPs to the heritable risk, which is usually an underestimation for cancer, in part due to the low prevalence of the disease. Importantly, current heritability studies do not provide insights into the mechanisms underpinning disease risk; thus, the focus has shifted on estimating the heritability explained by SNPs in functional genomic regions to provide a mechanistic interpretation of GWAS associations.

2.4. Methods for partitioning heritability

The vast majority of methods providing genome-wide heritability estimates usually assume that all genotyped SNPs have the same contribution to heritability. This assumption has already been questioned in literature [28], since it is more reasonable to assume that the amount heritability explained by a group of SNPs depends on the genomic region where they are located, e.g. promoter or coding regions. Thus, it is becoming apparent that estimating the heritability explained by SNPs residing in functional loci could give further insights in the genetic architecture of a disease.

Finucance et al. proposed a stratified LD score regression method (s-LDSC, [50]), which has been used to study the UK Biobank cohort. The method computes the heritability explained by SNPs belonging to a list of 53 functional binary classes, such as coding regions or histone marks. To do that, s-LDSC estimates the heritability explained by C functional categories, as follows:

$$\mathrm{E}\left[\chi_j^2\right] = N \sum_{c \in C} \tau_c l(j,c) + \mathrm{N\ a} + 1 \quad (9)$$

where the LD-score is computed only over the SNPs within the cth class and τ_c is the per-SNP heritability contribution of the cth class. Thus, the portion of heritability of one class with L_c^{SNP} variants is: $h_i^2 = L_c^{\mathrm{SNP}} \tau_c$. Recently, the model has been extended to account for continuous annotations, such as GC content or recombination rate [51].

An alternative approach uses the heritability estimator from summary statistics, HESS [52], to partition the genome in 1703 independent loci [53] and to then estimate the explained heritability as follows:

$$h_{\mathrm{local}}^2 = \frac{N\beta R^{-1}\beta - M}{N - M}$$

where β are the summary statistics for a GWAS with N individuals and M SNPS. R^{-1} is the inverse of the LD matrix approximated by a singular value decomposition, since the inverse usually does not exist due to linkage disequilibrium between SNPs. While for each category, s-LDSC partitions the whole genome in just two classes, HESS divides the genome in multiple regions (see supplementary figure S1). The scope of partitioning is to test whether a category has an heritability enrichment, that is the SNPs in the category explains a larger amount of h^2 compared to the genome-wide estimate. If h_k^2 is the heritability explained by the M_k SNPs belonging to annotation k, the quantity $(h_k^2/M_k)M$ is on the same scale of the genome-wide estimate; thus, in absence of any enrichment, the heritability for the single SNP h_k^2/M_k should be approximately equal to the genome-wide estimate h^2/M.

While partitioning methods could provide insights into genomic regions explaining a large proportion of heritability, there are still limits to use partitioned heritability to study cancer GWAS. Both HESS and LDSC are not robust for small sample sizes and low heritability diseases; this usually has the effect of providing erroneous negative local heritability estimates, suggesting that new robust estimators are needed to maximize the utility of these analyses.

3. The genetic landscape of breast cancer

The genetics of breast cancer has been extensively studied due to its relatively high prevalence and incidence in the broader population. The

Table 1
Breast cancer heritability estimates in European populations. For each study, we report the heritability estimate on the liability scale (h_l^2), the reported standard error or the 95% confidence intervals (CI) and the disease prevalence.

Cancer (subtype)	$h2_l$	Cases/controls
Breast (ER negative) [57]	0.096 (CI = [0, 0.199])	1998/3263
Breast (Self-reported) [49]	0.1104 (s.e. = 0.0221)	7480/329,679
Breast [58]	0.13 (s.e. = 0.011)	122,977/105,974

first three GWAS on breast cancer were published in 2007 and new targeted studies have been conducted in different populations. To date, the Breast Cancer Association Consortium (BCAC) is the largest breast cancer GWAS in Europeans, including more than 120,000 cases [54]; moreover, new genome-wide significant SNPs have been recently found in the same cohort using imputation [55]. Conversely, the UK Biobank (UKBB, [56]) represents the study with the largest total number of individuals ($N > 300,000$) and unbiased disease prevalence.

In this section we review the main results on breast cancer heritability, and then summarise and characterise susceptibility loci and genes for this malignancy.

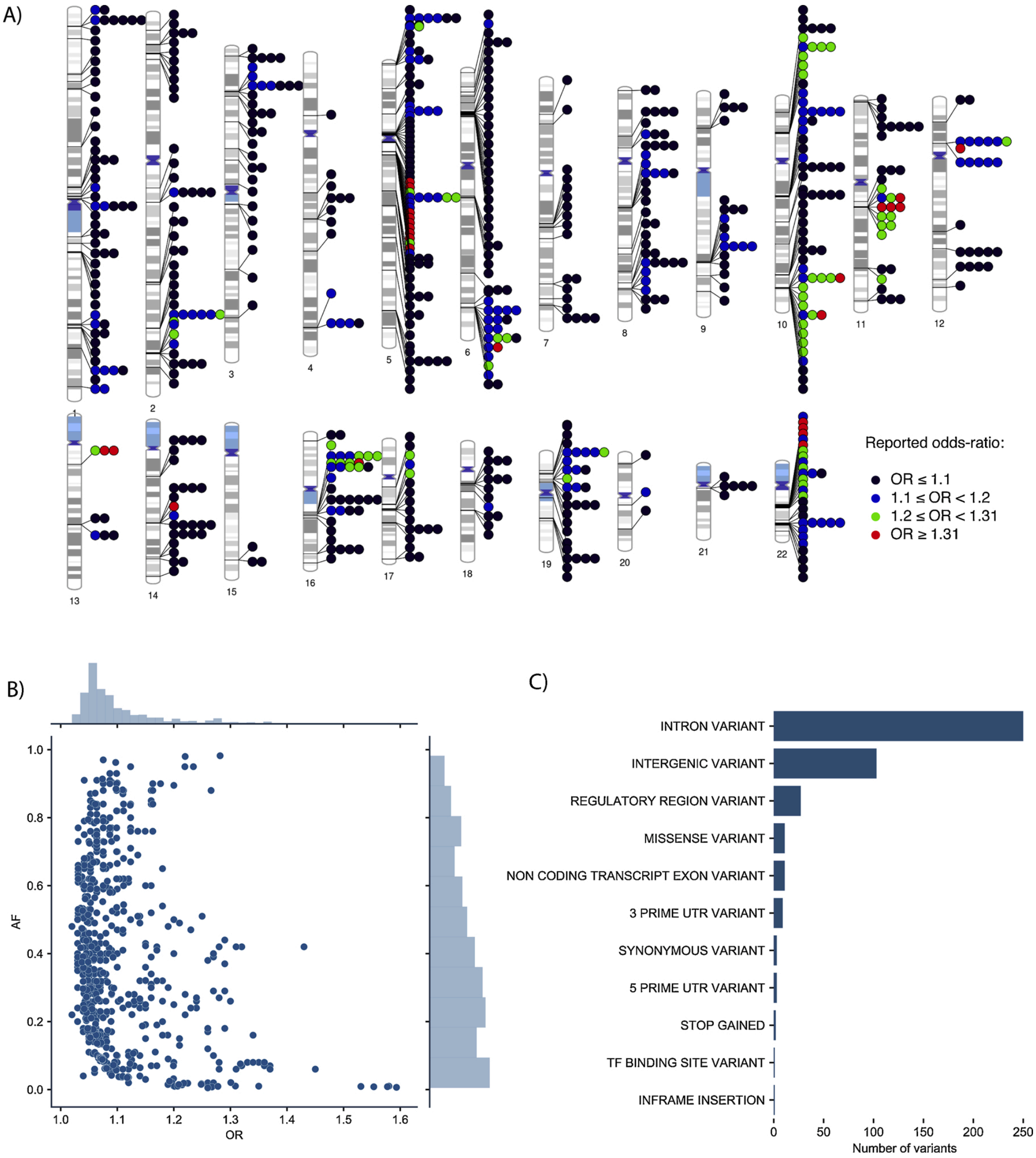


Fig. 1. Breast cancer susceptibility loci across the human genome. (A) Phenogram [96] of the 719 reported SNPs associated with breast cancer. Each SNP is represented by circles, and stacked symbols represent a locus for which multiple studies have reported an association. The color codes distinguish the reported odds-ratio (OR). Red circles denote those with stronger effect, OR ≥1.31, that are only 5% of the total. (B) Distribution of the odds-ratios (OR) and risk allele frequencies (AF). The central scatter plot shows the ORs and AFs for each SNP, where the top and right side are the corresponding histograms of OR and AF, respectively. For SNPs reporting only regression coefficients, β, we transformed these values in odds-ratios as follows $OR = exp(\beta)$. ORs are charactecterized by a long-tail distribution, whereas AF seems uniformly distributed. It is important to note the correlation between OR and AF, with rare variants have consistently stronger effects. (C) Functional classification of the variants reported by the GWAS catalog.

Table 2
Breast cancer susceptibility loci in European populations. We report SNPs associated with increased risk of breast cancer, whose odds-ratios (OR) are in the 95th percentile among all those reported in the GWAS catalog for this malignancy. For each SNP, we report the rsid, the cytogenic region, the reported odds ratio (OR), the functional consequence as sequence ontology term, the nearest gene, the reported risk allele frequency and the PUBMED id of the study.

SNPS	Region	OR	Context	Genes	Risk allele frequency	Pubmedid
rs62235635	22q12.1	1.59	Intron variant	PITPNB	0.0065	29059683
rs11571833	13q13.1	1.58	Stop gained	BRCA2	0.01	29058716
rs62235681	22q12.1	1.58	Intergenic variant	CHEK2	0.0085	29059683
rs1314913	14q24.1	1.57	Intron variant	RAD51B		23001122
rs62237615	22q12.1	1.55	Intron variant	TTC28	0.0082	29059683
rs62237573	22q12.1	1.53	Intron variant	TTC28	0.0092	29059683
rs3803662	16q12.1	1.5	Non coding transcript exon variant	CASC16		23001122
rs2229882	5q11.2	1.45	Synonymous variant	MAP3K1	0.06	24493630
rs2981579	10q26.13	1.43	Intron variant	FGFR2	0.42	20453838
rs10771399	12p11.22	1.39	Intergenic variant	PTHLH		24325915
rs16886448	5q11.2	1.37	Intron variant	MAP3K1	0.07	24493630
rs7726354	5q11.2	1.37	Intron variant	MIER3	0.06	24493630
rs16886034	5q11.2	1.36	Intergenic variant		0.08	24493630
rs16886364	5q11.2	1.36	Intron variant	MAP3K1	0.07	24493630
rs3822625	5q11.2	1.36	Synonymous variant	MAP3K1	0.07	24493630
rs16886397	5q11.2	1.36	Intron variant	MAP3K1	0.07	24493630
rs16886113	5q11.2	1.35	Regulatory region variant		0.08	24493630
rs614367	11q13.3	1.34	Intergenic variant	LINC01488	0.16	24493630
rs78540526	11q13.3	1.34	Intergenic variant	LINC01488	0.08	25751625
rs1017226	5q11.2	1.33	Intron variant	AC008937.2;MAP3K1	0.08	24493630
rs9397437	6q25.1	1.32	Intergenic variant	CCDC170	0.07	29058716
rs1219648	10q26.13	1.32	Intron variant	FGFR2	0.42	20872241
rs75915166	11q13.3	1.31	Regulatory region variant		0.06	25751625

3.1. Heritability estimates

The estimation of heritability from high-frequency variants for cancer presents multiple challenges and the results are highly dependent on the cohort and downstream processing. However, as novel studies with large cohorts are released and targeted GWAS are carried out, it is reasonable to expect that understanding cancer risk in the broader population will be possible.

While the exact heritability estimate varies across GWAS studies, there is a consensus estimate of breast cancer heritability being $h^2 \sim 0.1$ on the liability scale (see Table 1). This value is significantly smaller than previous familial estimates, $h^2 \sim 0.3$, although there is mounting evidence that this value could be an overestimation [31,33]. Sampson et al. report values of heritability, estimated via GREML, between 0.092 and 0.25, after adjusting for age, minor allele frequency and gender [57]. While the authors analysed GWAS data calibrated for cancer studies, the cohort is considerably smaller than the UKBB and BCAC cohorts. Jiang et al. analysed the BCAC cohort using LDSC regression [58], finding an heritability estimate $h^2 \sim 0.13$; interestingly, when excluding genome-wide significant SNPs and their linked loci, the heritability estimate is significantly smaller, suggesting that up to 45% of the total heritability is explained by genome-wide significant variants. Estimates obtained by LDSC on the UKBB cohort show remarkably coherent estimates, despite the prevalence of the malignancy being significantly smaller than other studies [49].

Recently, there has been increasing interest in identifying functional elements, such as histone mark or DNA I hypersensitive regions, explaining breast cancer heritability. However, analyses performed using stratified LDSC regression on the UKBB and BCAC cohorts were inconclusive [49]. Nonetheless, there is evidence suggesting that taking into account SNP location and functional effects in the analysis could provide useful insights on the role of inherited variants for cancer [59–61]. On this point, using local co-heritability between breast, lung, and prostate cancer [62,58], a pattern of local risk inheritance has been found. This result provides preliminary evidence that improvements in the analysis of partitioned heritability could be useful to discover loci across the human genome mediating the risk of multiple cancers.

3.2. Breast cancer risk loci across the human genome

Heritability studies have shown that high-frequency inherited mutations explain a significant proportion of breast cancer risk. We then move forward to identify SNPs and genes that are associated with increased risk of breast cancer in the broader population; ultimately, we aim at providing an updated map of breast cancer susceptibility genes across the human genome.

We obtained SNPs data from the GWAS Catalog [7], which reports more than 143,000 SNPs across 3522 studies. We then retrieved SNPs associated with breast cancer in European populations and mapped SNPs to genes, after applying quality control filters (see Supplementary Methods and Supplementary Figure S2). We also discarded SNPs, approximately half of the total reported, that did not reach genome-wide statistical significance set at $p < 5 \times 10^{-8}$; usually, p-values above this threshold are indicative of a small population size or old genotyping arrays, thus we preferred to filter those out as a conservative approach for our downstream analysis.

We found 719 significant variants (see Fig. 1A) reported by 26 different studies, which are within 50kb from 311 genes (see Supplementary Table 1, we consider a 50kb window to include regulatory regions in the analysis). Interestingly, of those 719 reported variants, 108 are reported in more than one study, while 311 are reported only once; while this provides preliminary evidence to support the robustness of a reported association, differences in tag SNP selection and reporting criteria across studies will likely result in different SNPs being reported for the same susceptibility haplotype (see Supplementary Methods and Supplementary Figure S2).

We observed that most variants account for limited increase in risk, with average odds ratio OR:1.11, and ranging from 1.02 for rs17529111 to 1.59 for rs62235635 (See Supplementary Figure S3,S4); moreover, the odds ratio for rs62235635 is still well below the strongest reported cancer association, that is for SNP rs995030-G in testicular germ cell tumors (OR: 2.26) [63] (Table 2).

The risk allele frequency for breast cancer is 0.37 on average, ranging from 0.005 to 0.98 (Supplementary Figure S5). Unsurprisingly, the data suggests a negative-correlation between cancer risk and allele frequency (see Fig. 1B). In particular, SNP rs62235635 in PITPNB, which is the variant with the lowest frequency, is also the one with the highest odds ratio $OR : 1.589$. This is consistent with other studies, which have shown that SNPs with detrimental impact are less frequently observed in the broader population because are likely to be subject to negative selection [51,64].

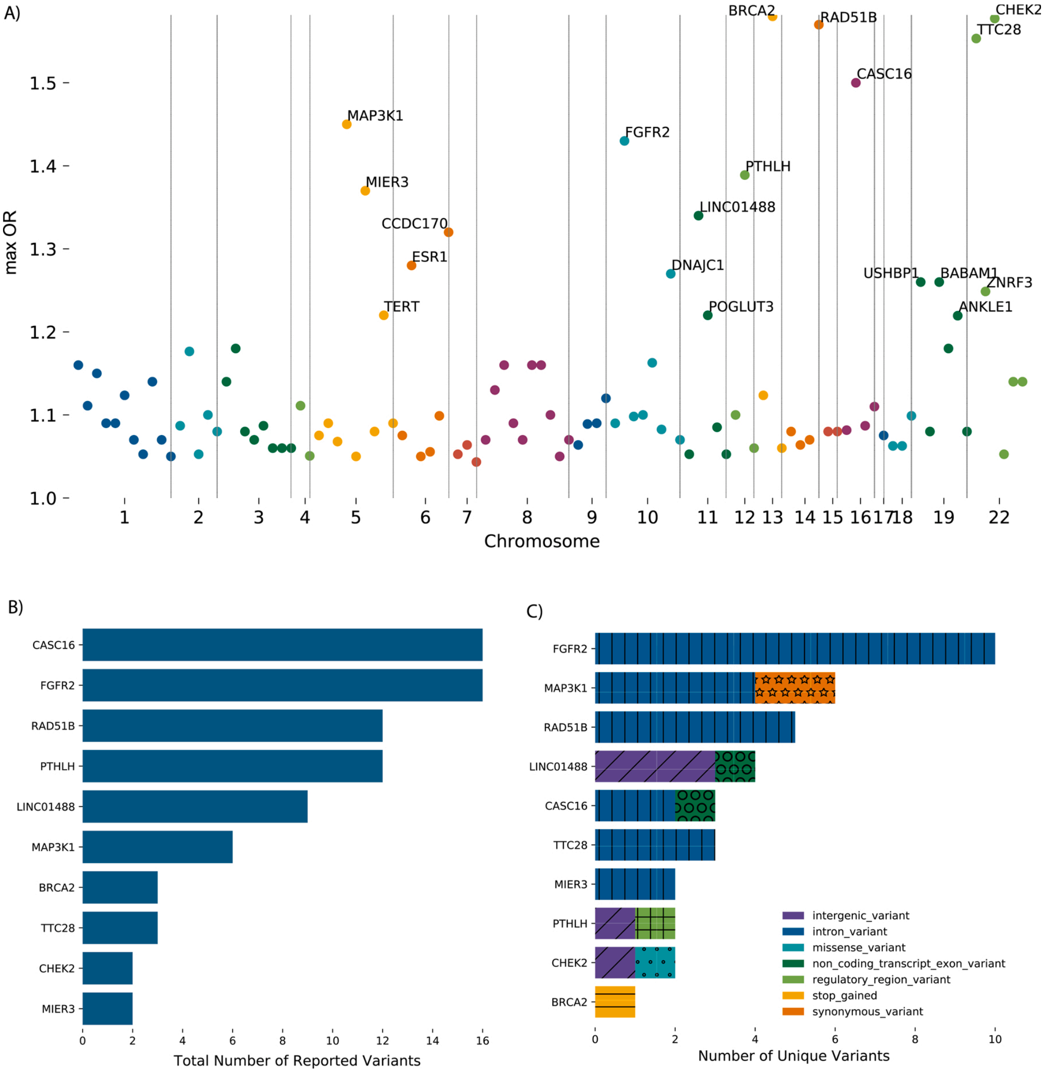


Fig. 2. Breast cancer susceptibility genes. (A) For each gene, we report the variants that are mapped within 50Kb of the gene body and the corresponding odds-ratios (ORs); variants reporting only regression coefficients were transformed into ORs by computing $OR = exp(\beta)$. The 10 genes with highest OR were further characterized below. (B) Number of reported variants for each gene. It is important to note that the same gene could harbor different variants, or the same variant could have been reported in multiple studies. (C) Number of unique variants grouped by gene and mutation effect. Only BRCA2, CHEK2 and MAP3K1 harbor exon variants.

We then analyzed the functional impact of each SNP associated with breast cancer (see Fig. 1C) and found that the vast majority of SNPs reside in introns or intergenic regions, with only a negligible fraction located in coding regions and possibly causing detrimental changes, such as missense variations or stop codon gain. While functional genomics techniques are continuously improving, testing functional effects of cancer SNPs will likely remain challenging, since phenotypic changes are going to be subtle and difficult to detect (Supplementary Figure S6). Nonetheless, we found that 89% of breast cancer SNPs are in or around a coding region, suggesting that most of them could act as cis-regulator of an upstream or downstream gene. We then used this information to compile a draft panel of genes associated with breast cancer heritability.

3.3. Genes associated with breast cancer susceptibility

We analysed 104 genes, out of the 311 in total, reported in at least 2 studies and associated with a Hugo symbol (Supplementary figure S7). It is worth noting that a gene can be reported multiple times because the same variant might have been reported in multiple studies or because different variants are mapped to the same genes.

We assigned the highest reported odds ratio, OR_{max}, and focused on those with the highest effect-size (see Fig. 2A). There are 20 genes with an $OR_{max} > 1.2$, with the top 10 genes having $OR_{max} > 1.28$; we hereby refers to these genes as breast cancer susceptibility genes (BCSGs, see Fig. 2B and C).

We then analyzed the functional role of BCSGs to identify possible mechanisms mediating breast cancer heritability. After performing literature curation, we found that 4 BCSGs control cell cycle, whereas 5 others are involved in DNA repair and invasion (see Fig. 3), which are fundamental processes underpinning all cancers [65,66]. It is important to note that CASC16 has been reported as a cancer susceptibility gene, but its functional role remains unclear.

We identified 4 BCSGs, namely CHEK2, FGFR2, MAP3K1 and TTC28, which control critical steps of the cell cycle. CHEK2 is a tumour sup-

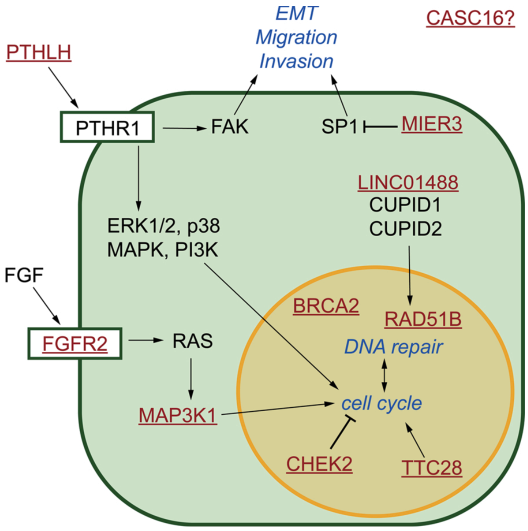


Fig. 3. Function and location of the breast cancer susceptibility genes. Breast cancer susceptibility genes (BCSGs, in red) are linked to three main biological processes (italic blue), namely cell cycle, DNA repair and invasion. When coupled to its ligand, FGFR2 triggers the RAS pathway, which activates downstream MAP3K1, thus promoting cell cycle. TTC28 and the hormone PTHLH also promote cell cycle while CHEK2 inhibits it. PTHLH induces FAK phosphorylation, leading to increased invasion, which is in turn inhibited by MIER3. Finally, both RAD51B and BRCA2 are active in DNA repair, whereas LINC01488 (CUPID2) mediates this process by impairing RAD51 recruitment.

pressor gene activated upon DNA damage, which activates genes controlling basic cellular activities, such as apoptosis, DNA repair, and cell cycle arrest. The mechanism is triggered via activation of TP53, BRCA1 or BRCA2 proteins [67]. Mutations in this gene are known to lead to the dysregulation of cell cycle and thus facilitate malignant transformation of the cell, and development of various types cancer, including breast cancer [68]. Mutations in CHEK2 gene mediate response to anthracycline based chemotherapy in breast cancer patients [69]. FGFR2 (Fibroblast growth factor receptor 2) negatively modulates activity of ESR1 and can inhibit estrogen signalling [70]. It has been clearly shown that FGFR2 mediates cancer susceptibility and mutations at this locus can account for an increase in the risk of breast cancer of up to 16% [71]. FGFR2 is also a member of the fibroblast growth factor receptor (FGFR) family, which controls upregulation of MAPK, PI3K/AKT, STAT and PLCγ signaling pathways. These pathways are involved in cancer mediating processes, such as cancer cell proliferation, differentiation, invasion, survival and carcinogenesis [72–74]. The mitogen-activated protein kinase kinase kinase 1 (MAP3K1) is a serine/threonine kinase having a role in signal transduction cascades, like MAPK, ERK, NF-κB, JNK or JUN pathway, which control critical cellular processes, including apoptosis, proliferation and differentiation [75]. Mutations in this gene affect kinase activity and are identified as oncogenic drivers [76]. TTC28 is a gene with oncogenic activity required during the cell cycle for condensation of spindle midzone microtubules, formation of the midbody, and completion of cytokinesis [77]. The gene resides in the proximity of the CHEK2 gene, thus suggesting a possible pattern of co-inheritance.

A second group includes 3 genes, namely BRCA2, RAD51B and LINC01488, which mediates repair mechanisms upon double-strand DNA breaks. BRCA2 is a well known cancer susceptibility gene, whose mutations are associated with 69% increase in risk of breast cancer and 17% increase in risk of ovarian cancer [78]. Mutations in this gene are also linked to other malignancies, including stomach, pancreatic and prostate cancer [79]. BRCA2 is also a therapeutic target of the FDA approved PARP inhibitors Rucaparib [80] and Niraparib [81]. For RAD51B there is evidence of association with familial breast cancer due to common variations [82]. In detail, RAD51B (RAD51 paralog B) encodes a protein which creates a complex with other RAD51 paralogs promoting binding of RAD51 upon DNA damage [83,84]. Damaged DNA prevents successful replication and cause a cell cycle arrest and apoptosis. Overexpression of RAD51 is usually found in tumors and mediates drug resistance [85]. Haploinsufficiency of RAD51B causes mild hypersensitivity to DNA-damaging agents favoring chromosome aberrations and aneuploidy in human cells by impairing RAD51 function [86]. LINC01488, also known as CUPID1, is a long non coding RNA regulated by estrogen and located in the 11q13 cytogenic band, which is associated with increased risk of breast cancer [87]. CUPID1, and the neighboring lncRNA CUPID2, have been shown to affect homologous-repair (HR) and non-homologous end joining (NHEJ) DNA repair mechanisms by impairing RAD51 recruitment.

We finally report 2 BCSGs, namely MIER3 and PTHLH, that are known to control invasion. MIER3 (MIER family member 3) together with MIER1/2 and BAHD1 (vertebrate protein that promotes heterochromatin formation and gene repression) repress expression of the steroid hormone receptor gene ESR1 [88]. MIER3 is reported to act as tumor suppressor [89] and is a known cancer susceptibility gene [90]. The Parathyroid Hormone Like Hormone (PTHLH), which encodes the Parathyroid hormone-related protein (PTHrP), is a gene responsible for the humoral hypercalcemia of the malignancy, mammary development and lactation [91,92]. During lactation it facilitates delivery of maternal calcium to milk and thus play a role in regulation of bone and mineral metabolism. By action through PTH1 receptors, PTHrP contributes to formation of bone metastasis through promotion of osteoclast formation and bone resorption [93]. It is important to note that FGFR2, MIER3 and LINC01488 are also involved in estrogen signaling, which regulates mammary gland development and is one of the main risk factors for breast cancer.

Taken together, the BCSGs identified in our analysis directly mediates cancer phenotypes and co-morbidities related to breast cancer. Upon further investigation, we also found these genes to be reported in many cancer panels (see Supplementary Figure 8) [94,95], thus suggesting also a possible link between somatic and inherited mutations.

4. Future directions

Decades of familial cancer studies provide evidence for a causal role of inherited genomic mutations, but these results have not been replicated by GWAS, when analyzing high-frequency mutations in the broader population. However, recent advances in sequencing and genotyping technologies, combined with accurate statistical methods, are enabling the identification of variants and quantify the heritable risk of many common malignancies, including breast cancer.

Here we provided an updated overview of SNPs and genes associated with breast cancer susceptibility, showing how variants in genes controlling cell cycle, DNA repair and invasion could modulate the risk of developing this disease. Since breast cancer susceptibility genes are often mutated in breast tumors, we speculate that a possible link between inherited and somatic mutations might exist and could provide new targets for clinical applications, including treatment and patients stratification. In particular, it is still difficult to dissect the functional role of the polymorphisms and how they may interact on a common mechanism, such as RAD51 regulation.

It will be of interest in long term follow up studies e.g. 'Generations' study, to see whether the type of breast cancer that develops is related to these polymorphisms, and to understand prevention studies e.g. hormone suppression in those with estrogen regulated polymorphic genes.

However, current experimental and analytical limitations lead us to believe that identifying the biological components modulating the risk of breast cancer and other oncological diseases will require substantial advances in statistical genetics. Moreover, experimental systems should be put in place to systematically validate the findings, and update and improve models. Taken together, heritability analysis is emerging as a powerful tool to quantify the effect of variants with subtle effects, but new robust methods able to identify biological units, such as genes or pathways, are needed to translate analytical results into biological and clinical findings.

5. Conflict of interest

The authors declare there are no conflicts of interest.

Acknowledgments

This work has been supported by the Wellcome Trust Seed Award in Science (207769/A/17/Z) to G.S. M.Z was supported by the European Regional Development Fund, Project ENOCH (No. CZ.02.1.010.00.016_0190000868). ALH was supported by the Breast Cancer Research Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.semcancer.2020.06.001.

References

[1] F. Bray, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. (2018), https://doi.org/10.3322/caac.21492. ISSN: 1542-4863. http://www.ncbi.nlm.nih.gov/pubmed/30207593.

[2] J. Ferlay, et al., Global Cancer Observatory: Cancer Today, International Agency for Research on Cancer, Lyon, France, 2018. https://gco.iarc.fr/today.

[3] B. Vogelstein, et al., Cancer genome landscapes, Science 340 (6127) (2013) 1546–1558, https://doi.org/10.1126/science.1235122. ISSN: 10959203. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749880/.

[4] D.E. Anderson, Genetic study of breast cancer: identification of a high risk group, Cancer 34 (4) (1974) 1090–1097. ISSN: 0008-543X.

[5] Y. Miki, et al., A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1, Science 266 (5182) (1994) 66–71, https://doi.org/10.1126/science.7545954. ISSN: 0036-8075.

[6] R. Wooster, et al., Identification of the breast cancer susceptibility gene BRCA2, Nature 378 (6559) (1995) 789–792, https://doi.org/10.1038/378789a0. ISSN: 00280836.

[7] J. MacArthur, et al., The new NHGRI-EBI Catalog of published genomewide association studies (GWAS Catalog), Nucleic Acids Res. 45 (D1) (2017) D896–D901, https://doi.org/10.1093/nar/gkw1133. ISSN: 13624962.

[8] A. Galvan, J.P.A. Ioannidis, T.A. Dragani, Beyond genomewide association studies: genetic heterogeneity and individual predisposition to cancer, Trends Genet. 26 (3) (2010) 132–141, https://doi.org/10.1016/J.TIG.2009.12.008. ISSN: 0168-9525. https://www.sciencedirect.com/science/article/pii/S0168952509002662.

[9] O. Zuk, et al., The mystery of missing heritability: genetic interactions create phantom heritability, Proc. Natl. Acad. Sci. U. S. A. 109 (4) (2012) 1193–1198, https://doi.org/10.1073/pnas.1119675109. ISSN: 1091-6490. http://www.ncbi.nlm.nih.gov/pubmed/22223662%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3268279.

[10] N.R. Wray, R. Maier, Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability, Curr. Epidemiol. Rep. 1 (4) (2014) 220–227, https://doi.org/10.1007/s40471-014-0023-3. ISSN: 2196-2995.

[11] R.D. Hernandez, et al., Ultrarare variants drive substantial cis heritability of human gene expression, Nat. Genet. 51 (9) (2019) 1349–1355, https://doi.org/10.1038/s41588-019-0487-7. ISSN: 1061-4036.

[12] T.A. Manolio, et al., Finding the missing heritability of complex diseases, Nature 461 (7265) (2009) 747–753, https://doi.org/10.1038/nature08494. ISSN: 00280836.

[13] P.M. Visscher, et al., 10 years of GWAS discovery: biology, function, and translation, Am. J. Hum. Genet. 101 (1) (2017) 5–22, https://doi.org/10.1016/j.ajhg.2017.06.005. ISSN: 00029297. https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409.

[14] E.A. Boyle, Y.I. Li, J.K. Pritchard, An expanded view of complex traits: from polygenic to omnigenic, Cell 169 (7) (2017) 1177–1186, https://doi.org/10.1016/j.cell.2017.05.038. ISSN: 10974172. http://linkinghub.elsevier.com/retrieve/pii/S0092867417306293.

[15] P.M. Visscher, W.G. Hill, N.R. Wray, Heritability in the genomics era concepts and misconceptions, Nat. Rev. Genet. 9 (4) (2008) 255–266, https://doi.org/10.1038/nrg2322. ISSN: 14710056.

[16] N. Mancuso, et al., The contribution of rare variation to prostate cancer heritability, Nat. Genet. 48 (1) (2016) 30–35, https://doi.org/10.1038/ng.3446. ISSN: 1061-4036.

[17] E.J. Saunders, et al., Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer, PLoS Genet. 10 (2) (2014) e1004129, https://doi.org/10.1371/journal.pgen.1004129. Ed. by Greg Gibson, ISSN: 1553-7404.

[18] D. Chen, et al., Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability, Carcinogenesis 36 (9) (2015) 992–998, https://doi.org/10.1093/carcin/bgv083. ISSN: 0143-3334.

[19] K. Litchfield, et al., Quantifying the heritability of testicular germ cell tumour using both population-based and genomic approaches, Sci. Rep. 5 (1) (2015) 13889, https://doi.org/10.1038/srep13889. ISSN: 2045-2322.

[20] Y. Sapkota, Germline DNA variations in breast cancer predisposition and prognosis: a systematic review of the literature, Cytogenet. Genome Res. 144 (2) (2014) 77–91, https://doi.org/10.1159/000369045. ISSN: 1424-859X. http://www.ncbi.nlm.nih.gov/pubmed/25401968.

[21] P.C. Sham, S.M. Purcell, Statistical power and significance testing in large-scale genetic studies, Nat. Rev. Genet. 15 (5) (2014) 335–346, https://doi.org/10.1038/nrg3706. ISSN: 14710064.

[22] A. Sud, B. Kinnersley, R.S. Houlston, Genome-wide association studies of cancer: current insights and future perspectives, Nat. Rev. Cancer 17 (11) (2017) 692–704, https://doi.org/10.1038/nrc.2017.82. ISSN: 14741768.

[23] G.M. Clarke, et al., Basic statistical analysis in genetic case–control studies, Nat. Protocols 6 (2) (2011) 121–133, https://doi.org/10.1038/nprot.2010.182. ISSN: 17542189.

[24] M.D. Gallagher, A.S. Chen-Plotkin, The Post-GWAS era: from association to function, Am. J. Hum. Genet. 102 (5) (2018) 717–730, https://doi.org/10.1016/j.ajhg.2018.04.002. ISSN: 15376605.

[25] A. Jacquard, Heritability: one word, three concepts, Biometrics 39 (2) (1983) 465, https://doi.org/10.2307/2531017. ISSN: 0006341X. https://www.jstor.org/stable/2531017?origin=crossref.

[26] A. Tenesa, C.S. Haley, The heritability of human disease: estimation, uses and abuses, 2013, https://doi.org/10.1038/nrg3377.

[27] J. Yang, et al., Common SNPs explain a large proportion of the heritability for human height, Nat. Genet. 42 (7) (2010) 565–569, https://doi.org/10.1038/ng.608. ISSN: 1061-4036.

[28] J. Yang, et al., Genome partitioning of genetic variation for complex traits using common SNPs, Nat. Genet. 43 (6) (2011) 519–525, https://doi.org/10.1038/ng.823. ISSN: 10614036. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4295936/.

[29] S.H. Lee, et al., Estimating missing heritability for disease from genomewide association studies, Am. J. Hum. Genet. 88 (3) (2011) 294–305, https://doi.org/10.1016/j.ajhg.2011.02.002. ISSN: 00029297. https://ac.elscdn.com/S0002929711000206/1s2.0S0002929711000206main.pdf?_tid=b134eef1-4774-41f6b049b67b4c6d1e3c&acdnat=1527155214_16ca4da6f407d1615fdb88c5b7fe7ba2.

[30] D. Speed, et al., Improved heritability estimation from genome-wide SNPs, Am. J. Hum. Genet. 91 (2012) 1011–1021, https://doi.org/10.1016/j.ajhg.2012.10.010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516604/pdf/main.pdf.

[31] N. Zaitlen, et al., Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits, PLoS Genet. 9 (5) (2013), https://doi.org/10.1371/journal.pgen.1003520. ISSN: 15537390. http://journals.plos.org/plosgenetics/article/file?id=10.1371/journal.pgen.1003520%7B%5C&%7Dtype=printable.

[32] D. Golan, E.S. Lander, S. Rosset, Measuring missing heritability: inferring the contribution of common variants, Proc. Natl. Acad. Sci. 111 (49) (2014) E5272–E5281, https://doi.org/10.1073/pnas.1419064111. ISSN: 0027-8424.

[33] J. Yang, et al., Concepts, estimation and interpretation of SNP-based heritability, Nat. Genet. (2017), https://doi.org/10.1038/ng.3941. ISSN: 15461718. arXiv: arXiv: 1011.1669v3.

[34] J.K. Haseman, R.C. Elston, The investigation of linkage between a quantitative trait and a marker locus, Behav. Genet. 2 (1) (1972) 3–19, https://doi.org/10.1007/BF01066731. ISSN: 0001-8244.

[35] G.B. Chen, Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression, Front. Genet. 5 (APR) (2014), https://doi.org/10.3389/fgene.2014.00107. ISSN: 16648021. www.frontiersin.org.

[36] Y. Wu, S. Sankararaman, A scalable estimator of SNP heritability for biobank-scale data, Bioinformatics 34 (13) (2018) i187–i194, https://doi.org/10.1093/bioinformatics/bty253. ISSN: 14602059. https://github.com/sriramlab/RHE-reg.

[37] J. Yang, et al., GCTA: a tool for genome-wide complex trait analysis, Am. J. Hum. Genet. 88 (2011) 76–82, https://doi.org/10.1016/j.ajhg.2010.11.011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014363/pdf/main.pdf.

[38] D. Speed, et al., Reevaluation of SNP heritability in complex human traits, Nat. Genet. 49 (7) (2017) 986–992, https://doi.org/10.1038/ng.3865. ISSN: 1061-4036.

[39] B. Bulik-Sullivan, et al., LD score regression distinguishes confounding from polygenicity in genome-wide association studies, Nat. Genet. 47 (3) (2015) 291–295, https://doi.org/10.1038/ng.3211. ISSN: 15461718.

[40] X. Zhou, P. Carbonetto, M. Stephens, Polygenic modeling with Bayesian sparse linear mixed models, PLoS Genet. 9 (2) (2013) e1003264, https://doi.org/10.1371/journal.pgen.1003264. ISSN: 15537390.

[41] O. Weissbrod, J. Flint, S. Rosset, Estimating SNP-based heritability and genetic correlation in case–control studies directly and with summary statistics, Am. J. Hum. Genet. 103 (1) (2018) 89–99, https://doi.org/10.1016/j.ajhg.2018.06.002. ISSN: 1537-6605. http://www.ncbi.nlm.nih.gov/pubmed/29979983%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6035374.

[42] D. Speed, D.J. Balding, SumHer better estimates the SNP heritability of complex traits from summary statistics, Nat. Genet. 51 (2) (2019) 277–284, https://doi.org/10.1038/s41588-018-0279-5. ISSN: 1061-4036.

[43] K. Hou, et al., Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture, Nat. Genet. 51 (8) (2019) 1244–1251, https://doi.org/10.1038/s41588-019-0465-0. ISSN: 1061-4036.

[44] J. Zheng, et al., LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis, Bioinformatics 33 (2) (2017) 272–279, https://doi.org/10.1093/bioinformatics/btw613. ISSN: 1367-4803.

[45] B. Bulik-Sullivan, Relationship between LD Score and Haseman–Elston regression, bioRxiv (2015) 018283, https://doi.org/10.1101/018283. https://www.biorxiv.org/content/10.1101/018283v1.full.

[46] L.M. Evans, et al., Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits, Nat. Genet. 50 (5) (2018) 737–745, https://doi.org/10.1038/s41588-018-0108x. ISSN: 15461718. URL: https://www.nature.com/articles/s41588-018-0108-x.pdf%20https://www.biorxiv.org/content/early/2017/03/10/115527.

[47] S. Gazal, et al., Reconciling S-LDSC and LDAK functional enrichment estimates, Nat. Genet. 51 (8) (2019) 1202–1204, https://doi.org/10.1038/s41588-019-0464-1. ISSN: 1061-4036.

[48] B. Pasaniuc, A.L. Price, Dissecting the genetics of complex traits using summary association statistics, Nat. Rev. Genet. 18 (2) (2017) 117–127, https://doi.org/10.1038/nrg.2016.142. ISSN: 14710064.

[49] T. Ge, et al., Phenome-wide heritability analysis of the UK Biobank, PLoS Genet. 13 (4) (2017) 1–21, https://doi.org/10.1371/journal.pgen.1006711. ISSN: 15537404.

[50] H.K. Finucane, et al., Partitioning heritability by functional annotation using genome-wide association summary statistics, Nat. Genet. 47 (11) (2015) 1228–1235, https://doi.org/10.1038/ng.3404. ISSN: 15461718.

[51] S. Gazal, et al., Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection, Nat. Genet. 49 (10) (2017) 1421–1427, https://doi.org/10.1038/ng.3954. ISSN: 1061-4036.

[52] H. Shi, G. Kichaev, B. Pasaniuc, Contrasting the genetic architecture of 30 complex traits from summary association data, Am. J. Hum. Genet. 99 (1) (2016) 139–153, https://doi.org/10.1016/j.ajhg.2016.05.013. ISSN: 15376605.

[53] T. Berisa, J.K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations, Bioinformatics 32 (2) (2015) btv546, https://doi.org/10.1093/bioinformatics/btv546. ISSN: 1367-4803.

[54] The Breast Association Consortium, http://bcac.ccge.medschl.cam.ac.uk/, 2019.

[55] K. Michailidou, et al., Association analysis identifies 65 new breast cancer risk loci, Nature 551 (7678) (2017) 92–94, https://doi.org/10.1038/nature24284. ISSN: 0028-0836.

[56] C. Sudlow, et al., UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, PLoS Med. 12 (3) (2015) e1001779, https://doi.org/10.1371/journal. pmed.1001779. ISSN: 1549-1676.

[57] J.N. Sampson, et al., Analysis of heritability and shared heritability based on genome-wide association studies for thirteen cancer types, J. Natl. Cancer Inst. 107 (12) (2015) djv279, https://doi.org/10.1093/jnci/djv279. ISSN: 1460-2105. URL: http://www.ncbi.nlm.nih.gov/pubmed/26464424%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4806328.

[58] X. Jiang, et al., Shared heritability and functional enrichment across six solid cancers, Nat. Commun. 10 (1) (2019) 431, https://doi.org/10.1038/s41467-018-08054-4. ISSN: 2041-1723.

[59] W. van Rheenen, et al., Genetic correlations of polygenic disease traits: from theory to practice, Nat. Rev. Genet. 20 (10) (2019) 567–581, https://doi.org/10.1038/s41576-019-0137-z. ISSN: 14710064. www.nature.com/nrg.

[60] M.A. Ferreira, et al., Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer, Nat. Commun. 10 (1) (2019), https://doi.org/10.1038/s41467-018-08053-5. ISSN: 20411723.

[61] X. Guo, et al., A comprehensive cis-eQTL analysis revealed target genes in breast cancer susceptibility loci identified in genome-wide association studies, Am. J. Hum. Genet. 102 (5) (2018) 890–903, https://doi.org/10.1016/j.ajhg.2018.03.016. ISSN: 15376605.

[62] H. Shi, et al., Local genetic correlation gives insights into the shared genetic architecture of complex traits, Am. J. Hum. Genet. 101 (5) (2017) 737–751, https://doi.org/10.1016/j.ajhg.2017.09.022. ISSN: 00029297. https://linkinghub.elsevier.com/retrieve/pii/S0002929717303919.

[63] E. Ruark, et al., Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14, Nat. Genet. 45 (6) (2013) 686–689, https://doi.org/10.1038/ng.2635. ISSN: 10614036.

[64] A.P. Schoech, et al., Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection, Nat. Commun. 10 (1) (2019) 790, https://doi.org/10.1038/s41467-019-08424-6. ISSN: 2041-1723.

[65] D. Hanahan, R.A. Weinberg, The hallmarks of cancer, 2000, https://doi.org/10.1016/S0092-8674(00)81683-9.

[66] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell 144 (5) (2011) 646–674, https://doi.org/10.1016/j. cell.2011.02.013. ISSN: 00928674.

[67] P. Apostolou, I. Papasotiriou, Current perspectives on CHEK2 mutations in breast cancer, Breast Cancer: Targets Ther. 9 (2017) 331–335, https://doi.org/10.2147/BCTT.S111394. ISSN: 11791314.

[68] C. Cybulski, et al., CHEK2 is a multiorgan cancer susceptibility gene, Am. J. Hum. Genet. 75 (6) (2004) 1131–1135, https://doi.org/10.1086/426403. ISSN: 0002-9297.

[69] S. Knappskog, et al., Low expression levels of ATM may substitute for CHEK2/TP53 mutations predicting resistance towards anthracycline and mitomycin chemotherapy in breast cancer, Breast Cancer Res. BCR 14 (2) (2012) R47, https://doi.org/10.1186/bcr3147. ISSN: 1465-542X. http://www.ncbi.nlm.nih.gov/pubmed/22420423%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3446381.

[70] T.M. Campbell, et al., Era binding by transcription factors NFIB and YBX1 enables FGFR2 signaling to modulate estrogen responsiveness in breast cancer, Cancer Res. 78 (2) (2018) 410–421, https://doi.org/10.1158/0008-5472.CAN-17-1153. ISSN: 15387445.

[71] D.J. Hunter, et al., A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer, Nat. Genet. 39 (7) (2007) 870–874, https://doi.org/10.1038/ng2075. ISSN: 10614036.

[72] E.M. Haugsten, et al., Roles of fibroblast growth factor receptors in carcinogenesis, Mol. Cancer Res. 8 (11) (2010) 1439–1452, https://doi.org/10.1158/1541-7786.MCR-10-0168. ISSN: 15417786.

[73] S. Wang, Z. Ding, Fibroblast growth factor receptors in breast cancer, Tumor Biol. 39 (5) (2017), https://doi.org/10.1177/1010428317698370. ISSN: 14230380.

[74] Y.K. Chae, et al., Inhibition of the fibroblast growth factor receptor (FGFR) pathway: the current landscape and barriers to clinical application, Oncotarget 8 (9) (2017) 16052–16074, https://doi.org/10.18632/oncotarget.14109. ISSN: 19492553.

[75] V. Sehgal, P.T. Ram, Network motifs in JNK signaling, Genes Cancer 4 (9-10) (2013) 409–413, https://doi.org/10.1177/1947601913507577. ISSN: 19476019.

[76] M. Michaut, et al., Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer, Sci. Rep. 6 (July 2015) (2016) 1–13, https://doi.org/10.1038/srep18517. ISSN: 20452322.

[77] T. Izumiyama, et al., A novel big protein TPRBK possessing 25 units of TPR motif is essential for the progress of mitosis and cytokinesis, Gene 511 (2) (2012) 202–217, https://doi.org/10.1016/j.gene.2012.09.061. ISSN: 03781119.

[78] K.B. Kuchenbaecker, et al., Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers, JAMA J. Am. Med. Assoc. 317 (23) (2017) 2402–2416, https://doi.org/10.1001/jama.2017.7112. ISSN: 15383598.

[79] H. Cavanagh, K.M.A. Rogers, The role of BRCA1 and BRCA2 mutations in prostate, pancreatic and stomach cancers, Hereditary Cancer Clin. Pract. 13 (1) (2015) 1–7, https://doi.org/10.1186/s13053-015-0038-x. ISSN: 18974287.

[80] A.M. Oza, et al., Antitumor activity and safety of the PARP inhibitor rucaparib in patients with high-grade ovarian carcinoma and a germline or somatic BRCA1 or BRCA2 mutation: integrated analysis of data from study 10 and ARIEL2, Gynecol. Oncol. 147 (2) (2017) 267–275, https://doi.org/10.1016/j.ygyno.2017.08.022. ISSN: 10956859.

[81] K.N. Moore, et al., Niraparib monotherapy for late-line treatment of ovarian cancer (QUADRA): a multicentre, open-label, single-arm, phase 2 trial, Lancet Oncol. 20 (5) (2019) 636–648, https://doi.org/10.1016/S1470-2045(19)30029-4. ISSN: 14745488.

[82] L.M. Pelttari, et al., RAD51B in familial breast cancer, PLoS ONE 11 (5) (2016) 1–18, https://doi.org/10.1371/journal.pone.0153788. ISSN: 19326203.

[83] N. Suwaki, K. Klare, M. Tarsounas, RAD51 paralogs: Roles in DNA damage signalling, recombinational repair and tumorigenesis, Semin. Cell Dev. Biol. 22 (8) (2011) 898–905, https://doi.org/10.1016/j.semcdb.2011.07.019. ISSN: 10963634.

[84] R. Prakash, et al., Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins, Cold Spring Harb. Perspect. Biol. 7 (4) (2015) 1–27, https://doi.org/10.1101/cshperspect.a016600. ISSN: 19430264.

[85] H.L. Klein, The consequences of Rad51 overexpression for normal and tumor cells, DNA Repair 7 (5) (2008) 686–693, https://doi.org/10.1016/j.dnarep.2007.12.008. ISSN: 15687864.

[86] O. Date, et al., Haploinsufficiency of RAD51B causes centrosome fragmentation and aneuploidy in human cells, Cancer Res. 66 (12) (2006) 6018–6024, https://doi.org/10.1158/0008-5472.CAN-05-2803. ISSN: 00085472.

[87] J.A. Betts, et al., Long noncoding RNAs CUPID1 and CUPID2 mediate breast cancer risk at 11q13 by modulating the response to DNA damage, Am. J. Hum. Genet. 101 (2) (2017) 255–266, https://doi.org/10.1016/j.ajhg.2017.07.007. ISSN: 15376605.

[88] G. Lakisic, et al., Role of the BAHD1 chromatin-repressive complex in placental development and regulation of steroid metabolism, PLoS Genet. 12 (3) (2016) 1–26, https://doi.org/10.1371/journal.pgen.1005898. ISSN: 15537404.

[89] M. Peng, et al., MIER3 suppresses colorectal cancer progression by down-regulating Sp1, inhibiting epithelial-mesenchymal transition, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/s41598-017-11374-y. ISSN: 20452322.

[90] A.D. DenDekker, et al., Rat Mcs1b is concordant to the genome-wide association identified breast cancer risk locus at human 5q11.2 and MIER3 is a candidate cancer susceptibility gene, Cancer Res. 72 (22) (2012) 6002–6012, https://doi.org/10.1158/0008-5472.CAN-12-0748. ISSN: 00085472.

[91] G.J. Strewler, The physiology of parathyroid hormone-related protein, N. Engl. J. Med. 342 (3) (2000) 177–185, https://doi.org/10.1056/NEJM200001203420306. ISSN: 00284793.

[92] W. Kim, J.J. Wysolmerski, Calcium-sensing receptor in breast physiology and cancer, Front. Physiol. 7 (SEP) (2016) 1–11, https://doi.org/10.3389/fphys.2016.00440. ISSN: 1664042X.

[93] T.J. Martin, R.W. Johnson, Multiple actions of parathyroid hormonerelated protein in breast cancer bone metastasis, Br. J. Pharmacol. (2019), https://doi.org/10.1111/bph.14709. ISSN: 14765381.

[94] D. Chakravarty, et al., OncoKB: a precision oncology knowledge base, JCO Precis. Oncol. 2017 (2017), https://doi.org/10.1200/PO.17.00011. ISSN: 2473-4284. URL: http://www.ncbi.nlm.nih.gov/pubmed/28890946%20http: //www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5586540.

[95] Z. Sondka, et al., The COSMIC cancer gene census: describing genetic dysfunction across all human cancers, Nat. Rev. Cancer 18 (11) (2018) 696–705, https://doi.org/10.1038/s41568-018-0060-1. ISSN: 1474-1768. http://www.ncbi.nlm.nih.gov/pubmed/30293088.

[96] Ritchie Lab Visualisation Tools, http://visualization.ritchielab.org/home/index.

2.3 Conclusions

Understanding cancer risk at the germline level could be beneficial for two main reasons. First, the identification of specific variants increasing cancer risk can lead towards targeted cancer surveillance. For instance, Poligenic Risk Scores (PRSs), which aim at providing a set of SNPs that correctly predict the risk of cancer for each individual [Mavaddat et al. 2019] , are used to inform public health policies, like cancer screening programs [Mars et al. 2020]. More interestingly for this work, understanding how the genetic background affects cancer risk would help to decode the pre-cancerous conditions that, alongside environmental exposure, lead to cancer [Ramroop, Gerber, and Toland 2019]. Integrating the germline signal into the existing knowledge of tumorigenesis has the potential to identify markers or secondary actionable targets and to stratify patients for treatment [Liu et al. 2021].

This review, consistent with previous accounts [Stracquadanio et al. 2016; Fagny et al. 2020], has found that significant GWAS SNPs often occur in genes involved in key functions, like signaling pathways, developmental processes, cell-cell adhesion processes. Furthermore, recent in-vitro models have shown that frequent germline variants have detectable regulatory effects [Lawrenson et al. 2015; Whitington et al. 2016], which might also affect cancer risk synergistically with environmental exposure [Jeffers et al. 2021; Surakhy et al. 2020].

Nonetheless, the analysis of GWAS significant hits has weaknesses that need to be addressed by future experimental designs and statistical methods. SNP prioritisation methods and functional interpretations of the results rely on mapping each locus to a functional category, e.g. whether the SNP is in a coding, regulatory, or intergenic region. Improved mappings [Yardımcı et al. 2019; Watanabe et al. 2017; Frankish et al. 2020] that integrate genomic and epigenomic annotations enable refined predictions of SNP function [Weissbrod et al. 2020; Hutchinson, Asimit, and Wallace 2020; Wen, Pique-Regi, and Luca 2017] and will be instrumental for the development of methods that correctly identify the effect of each locus.

Moreover, single hit GWAS analyses discard the contribution of the majority of SNPs and they have been shown to be insufficient to explain the whole heritability [Zhang et al. 2020]. We reason that accounting for all SNPs, regardless of significance, and linking them to their functional role might reveal subtle effects that would be otherwise undetected. State-of-the-art methods have tried to partition heritability to estimate the polygenic signal and attribute it to local and functional annotations. However, these methods lack sufficient resolution to generate testable hypotheses; they do not identify putative gene- or variant-level markers that can be investigated to translate analytical results into biological and clinical findings. In the next chapter, we further address these limitations and present a method to estimate heritability at gene-level resolution.

3 Gene-level heritability analysis

3.1 Introduction

In the previous chapter, we presented an overview of the findings of cancer GWAS and the methods to estimate the amount of inherited cancer risk due to high-frequency variants. We have shown that GWAS hits alone are often providing a non-exhausting characterisation of cancer risk, and we presented heritability as a useful metric to account for the inherited risk due to all variants.

SNP heritability, h^2_{SNP}, that is the heritability explained by all genotyped SNPs, is a single estimate that accounts for the risk of cancer due to inherited variants, Fig. 3.1 [Lee et al. 2011; Golan, Lander, and Rosset 2014]. Cancer heritability is generally low and its estimates often suffer by the low prevalence of cases in the population, hence most studies on GWAS heritability focus on frequent traits and qualitative phenotypes. Nonetheless, cancer heritability has been studied using multiple methods and data panels, especially for more frequent malignancies. Available estimates of GWAS heritability, see Tab. 3.1, show that estimates vary greatly between studies and methods. In many cases h^2_{SNP} is around 0.10, with prostate and testicular tumors being those with the strongest signal reaching 0.3 of heritability. Conversely, for ovarian and

	[1] h_l^2 **(95% CI)**	**[2]** h_l^2 **(SE)**	**[3]** h_l^2 **(95% CI)**	**[4]** h_l^2 **(se)**	h_l^2
Bladder	0.123 (0.086-0.160)			0.169 (0.067)	
Breast	0.096 (0 - 0.199) *	0.14 (0.012)	0.14 (0.09–0.18)	0.11 (0.020)	
Colorectal		0.09 (0.0089)	0.11 (0.07–0.14)	Colon: 0.12 (0.042) Rectum: 0.068 (0.073)	0.072 [9]
Kidney	0.147 (0.023 - 0.270)			0.142 (0.097)	
Lung	0.206 (0.142 - 0.271)	0.075 (0.011)	0.13 (0.08–0.19)	0.117 (0.058)**	
Ovarian		0.033 (0.0065)	0.07 (0.02–0.12)	-0.048 (0.117)	
Pancreas	0.098 (0.037 - 0.160)		0.05 (0–0.10)	0.104 (0.165)	0.21 [5]
Prostate	0.378 (0.244 - 0.513)	0.18 (0.021)	0.27 (0.21–0.33)	0.111 (0.037)	0.28 [6] 0.27 [7]
Testes	0.299 (0.084 - 0.513)			0.381 (0.386)	0.374 [8] (CI: 0.28-0.47)

Table 3.1: Snapshot of the available SNP heritability estimates for most frequently studied malignancies. The first four columns are the estimates from studies on multiple cancers , while the last column recapitulates data from different studies. We selected only those malignancies for whom we found more than one estimate and clear cancer types. [1] Sampson et al. 2015, [2]Jiang et al. 2019, [3]Lindström et al. 2017, [4]UKBB `https://nealelab.github.io/UKBB_ldsc/`, [5]Chen et al. 2019, [6]Gusev et al. 2016, [7]Mancuso et al. 2015, [8]Litchfield et al. 2015, [9]Jiao et al. 2014. *) ER-, **) Brunchus and Lung

pancreatic cancer SNP heritability is very low, which is also possibly due to a lack of power in rarer malignancies.

Genome-wide estimates of heritability do not provide any local or functional information on how SNPs explain cancer risk, and they can hardly be used to inform further studies on SNP function. SNP heritability, though, is an additive measure, hence it is straightforward to obtain partitioned estimates of h^2_{SNP} like the heritability explained by each chromosome [Yang et al. 2011]. More relevantly, we can obtain estimates of the percentage of total heritability explained by a panel of significant SNPs, e.g. those that are significant in the GWAS or those in cancer susceptibility genes. Nontheless, these SNPs, often exclusively considered for *post-hoc* investigations, do not fully explain cancer heritability. For instance, top GWAS hits and their flanking regions ($\pm 500kb$) have been shown to account for less than 53% of the total heritability, 28% on average, [Jiang et al. 2019] while previously known CSGs explain on average 12% of total heritability [Sampson et al. 2015]. It is then clear that known CSG and GWAS hits do not provide a full picture of cancer risk and that the polygenic signal is attributable to more loci.

Methods to explicitly partition the heritability allow to obtain estimates for groups of SNPs. As we described in Chapter 2, there are two main approaches to this issue: functional partitioning and local partitioning, see Fig. 3.1. The former splits the genome in multiple, overlapping, binary annotations; heritability estimates for specific functional attributes are obtained and can be used to prioritise further investigation [Finucane et al. 2015; Finucane et al. 2018]. However, no local target can be extracted from such methods, as each functional estimate is gathered from genome-wide SNPs. Differently, local partitioning returns non-overlapping heritability estimates, which have proven useful to understand the co-heritability of different traits at the haplotype level [Shi, Kichaev, and Pasaniuc 2016; Shi et al. 2017]. While state-of-the-art local partitioning methods can estimate heritability for up to 1700 genomic regions, they are still too coarse-grained to provide testable targets.

We then developed a method, Bayesian gene-level heritability analysis (BAGHERA), see Fig. 3.1, that applies local partitioning to heritability, and reaches gene-level resolution. In the paper below, we describe BAGHERA in depth, and we benchmark its performance. We then applied our method to all 38 cancer types of the UK Biobank [Sudlow et al. 2015] and we gathered a comprehensive picture of the heritability loci, which are those explaining a significant amount of heritability for each malignancy.

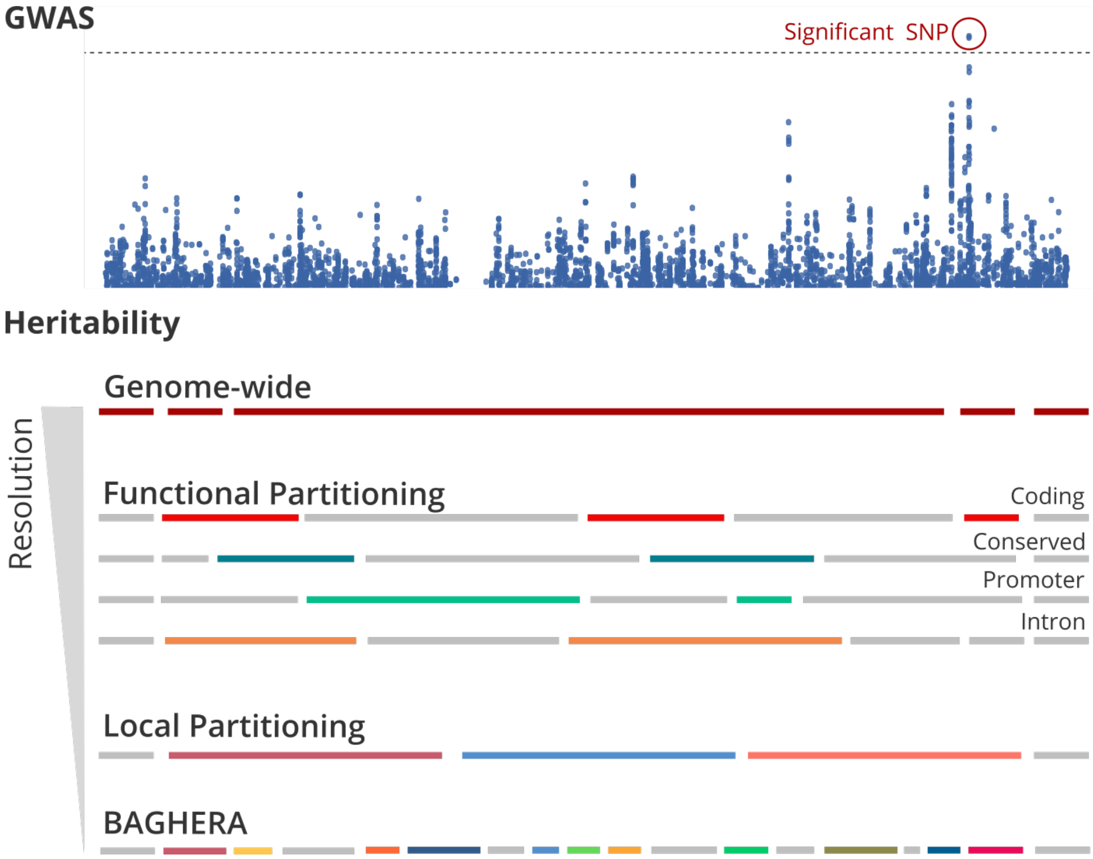


Figure 3.1: *Partitioning the heritability. Here we show an illustration of the different levels of resolution of heritability estimates. The last row, BAGHERA, is the method for estimating gene-level heritability that is presented in this chapter.*

3.2 The Landscape of the Heritable Cancer Genome

The whole manuscript has been drafted by V. Fanfani, with the supervision and contributions of G. Stracquadanio. The development of the method and data analysis has been carried out by V.Fanfani with the supervision of G. Stracquadanio and L.Citi. A.L. Harris and F. Pezzella contributed to the editing of the manuscript and the validation of the conclusions.

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

The Landscape of the Heritable Cancer Genome

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

ABSTRACT

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

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

See related commentary by Bader, p. 2586

Introduction

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

It has been hypothesized that part of cancer risk could be apportioned to high-frequency low-penetrant variants, such as single nucleotide polymorphism (SNP). Genome-wide association studies (GWAS; ref. 6) have been instrumental in identifying SNPs associated with increased risk of cancer in the broader population (1), including breast (7), prostate (8), testicular (9), and blood malignancies (10, 11). However, the vast majority of SNPs account only for a limited increase in cancer risk (1) and are usually filtered out by multiple hypotheses correction procedures applied in GWAS analysis (12), which ultimately leaves most of the cancer risk unexplained (5).

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

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

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

[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, 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.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: Giovanni Stracquadanio, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom. Phone: 4401-3165-07193; E-mail: giovanni.stracquadanio@ed.ac.uk

Cancer Res 2021;81:2588–99

doi: 10.1158/0008-5472.CAN-20-3348

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

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

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

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

Materials and Methods

Estimation of heritability at the gene level

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

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

Hereby, we denote as genome-wide heritability the amount of heritability explained by all genotyped SNPs, M, whereas we refer to the amount of heritability explained by the SNPs in a gene as gene-level heritability. In a model where each SNP has equal contribution to the genome-wide heritability, the per-SNP heritability is $\hbar^2 = h^2_{SNP}/M$. Conversely, if variants can have varying contribution to the genome-wide heritability, we can model the per-SNP heritability as a random variable, $\hbar^2_M$, whose expectation is $\hbar^2_M = \mathrm{E}[\hbar^2_j]_{j=1,\cdots,M}$, 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 nonoverlapping genes in the human genome, each of them with M_k SNPs, the genome-wide heritability can be expressed as $h^2_{SNP} = \sum_{k=1}^{K} \sum_{j \in k} \hbar^2_j = \sum_{k=1}^{K} M_k \hbar^2_{M_k}$, where $M_k \hbar^2_{M_k}$ is the amount of heritability explained by all the SNPs in the k-th gene. Thus, let the number of SNPs in each gene and the gene-level per-SNP heritability be independent random variables, it is straightforward to prove that the expectation of the gene-level per-SNP heritability is the per-SNP genome-wide estimate $h^2_{SNP}/M = E[\hbar^2_{M_k}]_K$. However, estimating h^2_{SNP} only from SNPs assigned to genes would lead to biased estimates, because the contribution of the SNPs in intergenic regions would be neglected; thus, SNPs outside genic regions are assigned to a single intergenic locus, such that the heritability is correctly estimated from all genotyped SNPs.

A hierarchical Bayesian model for heritability estimation

The estimation of heritability can be modeled 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:

$$\begin{aligned} h^2_{SNP} &\sim F_1()\ \mathit{with}\ \mathit{supp}(F_1()) \in [0,1] \\ h^2_k |\ h^2_{SNP} &\sim F_2\left(h^2_{SNP}\right) \\ Y |\ h^2_k &\sim F_3\left(h^2_k\right) \end{aligned} \tag{A}$$

where $F_1, F_2,$ and 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 nonadditive and environmental effects, these quantities can be modeled as random variables with $\sigma^2_g \sim \Gamma(\alpha,\theta)$ and $\sigma^2_e \sim \Gamma(\beta,\theta)$, respectively. Because $\Gamma(\alpha,\theta)/(\Gamma(\alpha,\theta)+\Gamma(\beta,\theta)) \sim \mathrm{Beta}(\alpha,\beta)$, a suitable distribution for F_1, in Eq. A, would be an uninformative Beta distribution, e.g., $\mathrm{Beta}(1,1)$. In practice, the use of a Beta distribution as prior for h^2_{SNP} allows us to obtain accurate heritability estimates in the unit range even for low-heritability diseases, where classical methods are usually inaccurate (28).

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

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

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

$$\chi^2_{jk} \sim \mathcal{N}\left(N l_j h^2_k/M + e, \sqrt{l_j}\right) \tag{B}$$

where χ^2_{jk} 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 within a certain genomic window (e.g., 1 Mb) in a given

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

population. Importantly, LD scores can be conveniently computed from large-scale genetic studies, such as the 1000 Genomes project.

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

BAGHERA software

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

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

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

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

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

UK BioBank summary statistics processing and curation

We used summary statistics of the UK BioBank GWAS for cancers classified using the ICD10 disease classification (source: https://nealelab.github.io/UKBB_ldsc/); importantly, data are uniformly processed with state-of-the-art methods, which prevents any methodologic bias. Here, we developed a custom pipeline to assign LD scores to SNPs, and SNPs to human genes (see **Fig. 1**). Specifically, we used precomputed LD scores for SNPs on autosomal chromosomes with minor allele frequency MAF>0.01 in the European population (EUR) of the 1000 Genomes project. We then removed the SNPs on chr6:26,000,000–34,000,000, because this region contains the major histocompatibility complex that have unusual genetic patterns and is known to affect GWAS result interpretation (25, 32). Ultimately, our analysis is conducted on $1,285,620$ SNPs over 22 chromosomes.

We then used Gencode v31 to determine the genomic coordinates of protein coding genes in the GRCh37 human genome. First, we merged overlapping genes by creating a new multigene locus, whose name denotes the overlapping genes and whose boundaries are defined as the first and last base-pair of these loci. We then assigned to a locus all SNPs within or no more than ±50 kb away from its boundaries (**Fig. 1**); this strategy allows us to account for *cis*-regulatory elements while retaining gene-level resolution. All other SNPs are assigned to the intergenic locus. Overall, 55% of SNPs were mapped to a locus, while the rest of them are assigned to the intergenic term. Finally, to mitigate false positives due to poorly genotyped regions, we considered only gene-loci harboring at least 10 variants. Ultimately, our dataset consists of $15,025$ loci; $12,042$ (80.1%) of them are harboring more than 10 SNPs, which were considered in our heritability study. The results of our analyses are deposited in CSV format on Zenodo (doi: 10.5281/zenodo.3968269).

Enrichment analyses

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

Results

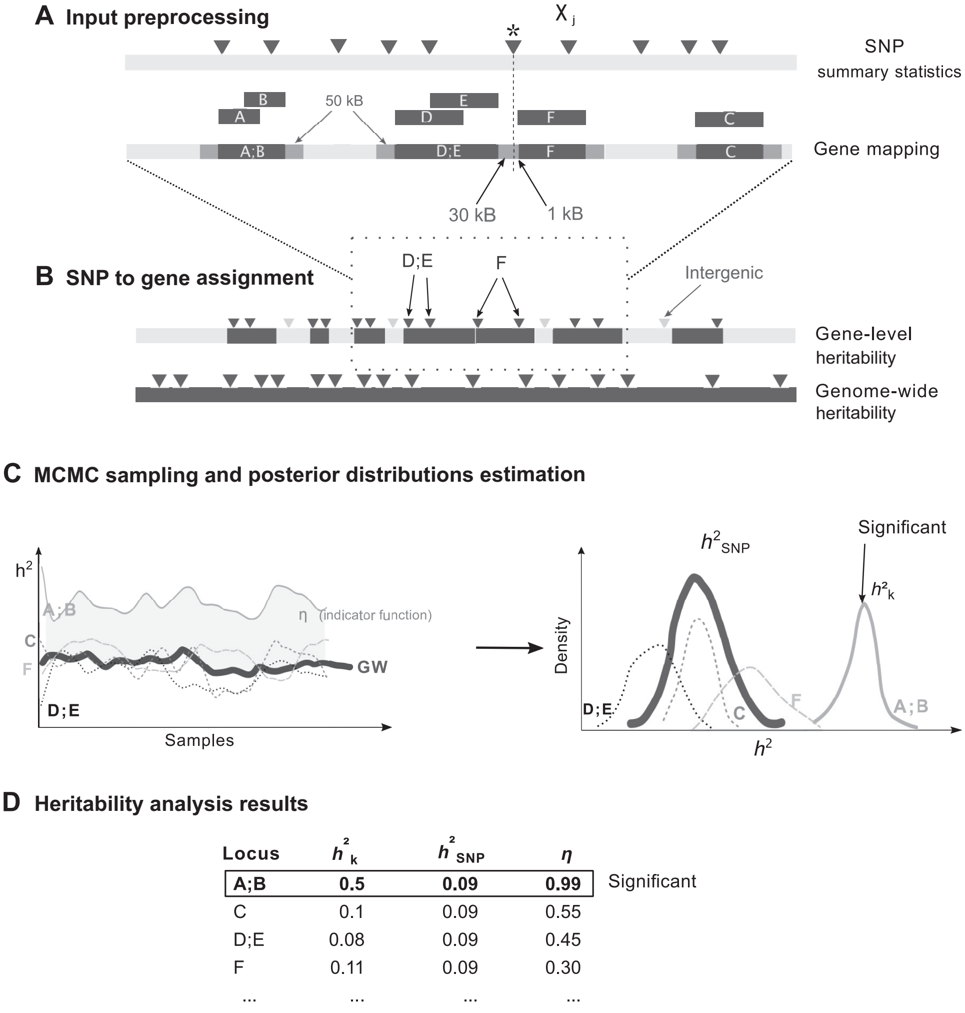
Simulations assessing robustness of genome-wide and gene-level estimates for low heritability traits

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

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

We then assessed whether BAGHERA was able to identify heritability loci, that is loci harboring SNPs with a contribution to heritability higher than expected under a constant per-SNP heritability contribution. To do that, we selected 1% of the loci on chromosome 1 (≈ 13) as heritability loci and computed receiver operator

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348



Locus	h^2_k	h^2_{SNP}	η
A;B	**0.5**	**0.09**	**0.99**
C	0.1	0.09	0.55
D;E	0.08	0.09	0.45
F	0.11	0.09	0.30
...	...	...	...

Figure 1.

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

characteristic (ROC) and precision recall (PR) curves at varying levels of genome-wide heritability (see Supplementary Methods). For all curves, we evaluated the area under the curve (AUC). Here, we found that BAGHERA correctly identified heritability loci (ROC AUC: 0.89), although precision and recall were consistently higher for higher genome-wide heritability levels (PR AUC: 0.41 for $h^2 = 0.01$, >0.58 for $h^2 > 0.01$; Supplementary Figs. S1 and S2A–S2C).

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

We addressed this limitation by simulating summary statistics using only linkage disequilibrium information (see Supplementary Methods). This approach provides a tractable framework to test varying

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

levels of heritability enrichment, reported in terms of fold change with respect to the genome-wide estimate, and to simulate SNPs across the entire genome, rather than a single chromosome.

We then assessed the performance by computing ROC and PR curves, the true positive rate (TPR), and the FDR. BAGHERA correctly identifies heritability loci, even with fold changes in heritability as low as $f_c = 5$ (ROC AUC range: $0.70-0.99$). Importantly, we found BAGHERA to be conservative with a low FDR across all scenarios (FDR range: $0\%-5\%$); this result suggests that our method is suitable for exploratory analyses, and that significant results are associated to true biological signal (see Supplementary Figs. S3–S5).

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

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

Genome-wide estimates were compared with LD score regression (LDsc) results (28). Gold-standard methods require raw data; however, previous studies have shown that LDsc has comparable performance in most scenarios (22). Because LDsc is routinely used to estimate heritability for the traits in the UK BioBank, we retrieved the results for all 38 cancers and compared them with BAGHERA estimates. We found strong consensus between the estimates of the two methods (see Supplementary Fig. S6), consistent with the fact that BAGHERA uses a similar genome-wide estimator. Nonetheless, BAGHERA is more robust for low heritability traits, because our Bayesian formulation guarantees correct heritability estimates in the unit domain, whereas LDsc incorrectly provides negative values.

Performances on local heritability analysis were compared with the heritability estimation from summary statistics (HESS) method (27), which is the only available approach to estimate local heritability from summary statistics. Here, we used BAGHERA to estimate the heritability of 1703 regions, as defined in the HESS original study (see Supplementary Methods). We then restricted our analysis to breast and prostate cancer data, because these malignancies are those with the highest h^2_{SNP} estimates; this was necessary to ensure a fair comparison between the two methods, because HESS is not designed for low heritability traits. Here, we found a statistically significant correlation between HESS and BAGHERA estimates (Pearson ρ: 0.76 for prostate and 0.78 for breast, see Supplementary Figs. S7 and S8). However, because BAGHERA provides robust estimates for as much as 15,000 regions, it enables more detailed analyses compared with HESS.

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

Genome-wide estimates of cancer heritability in the UK Biobank

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

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

Estimating heritability from nontargeted cohorts can be challenging, due to the small prevalence of the disease. To test whether we had sufficient signal for each cancer, we reasoned that if the SNP test statistic follows a χ^2 distribution with 1 degree of freedom, under the null hypothesis of no association, its expected value is $E[\chi^2] = 1$; thus, similarly to other studies, we expected to have sufficient polygenic signal for our analysis if the average χ^2 was greater than 1 (25). Here, we found the vast majority of cancers to have an average $\chi^2 \approx 1$, with only 17 having a deviation greater than 1% from the expected value of the test statistic. We also did not consider cancers assigned to other malignant neoplasm of the skin (C44), because (i) most tumors belong to unspecified anatomic regions (C44.3, C44.9); (ii) are predominantly caused by sun exposure in Europeans; and (iii) and includes poorly characterized rare skin cancers. Ultimately, we restricted our study to 16 cancers for which we had sufficient power to perform our analysis. Nonetheless, all our results are consistent with those we obtained when considering all 38 cancer types (see Supplementary Figs. S9–S12D and S13A–S13C; Supplementary Tables S1–S3).

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 lymphoma and up to 31% for testis (see **Table 1**) consistent with other available estimates for this cohort (see Supplementary Materials and Supplementary Figs. S14, 15A–S15D, and S16A–S16C; Supplementary Table S4). While comparison between cancer heritability estimates is usually difficult across studies, due to differences in histologic classification and genetic confounders, we found our heritability estimates on the liability scale to be consistent with those reported for other cohorts, in particular for breast, prostate, testes, and bladder (17, 18, 20, 34). The heritability of testicular cancer is the highest among all malignancies ($h^2_{SNP_L} = 0.3158$), consistent with the hypothesis that germline variants have stronger effects in early onset and young adult cancers. However, early onset cancers are underrepresented in the UK Biobank, because children and young adults were not enrolled in the study, and thus, an accurate estimation of the correlation between age of onset and heritability is not possible. Nonetheless, it is interesting to note that many malignancies with onset in late adulthood, such as prostate or bladder, still display a significant heritable component, ranging from $h^2_{SNP_L} = 0.25$ for brain tumors (age of onset: 59) to $h^2_{SNP_L} = 0.08$ for diffuse non-Hodgkin lymphoma (age of onset: 60). Overall, 14 of 16 cancers (87%) show heritability higher than 10%, suggesting a consistent contribution of SNPs to the heritable risk of cancer.

Heritability loci across 16 malignancies

We identified 783 heritability loci (η>0.99), harboring $1,146$ protein-coding genes, across 16 cancers (see **Fig. 2**), with 53 heritability loci per malignancy on average, ranging from 5 loci in

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

Table 1. Genome-wide heritability of the 38 cancers in the UK BioBank.

ICD10	Malignancy	Cases	Prevalence	$\hat{\chi}^2$	h^2_{SNP}	$h^2_{SNP_L}$	HL
C44	Other malignant neoplasms of skin	9,086	0.0252	1.1408	0.0341	0.2422	422
C50	**Malignant neoplasm of breast**	8,304	0.0230	1.0869	0.0170	0.1285	267
C61	**Malignant neoplasm of prostate**	4,342	0.0120	1.0765	0.0191	0.2320	271
C18	**Malignant neoplasm of colon**	2,226	0.0062	1.0399	0.0070	0.1416	33
C43	**Malignant melanoma of skin**	1,672	0.0046	1.0288	0.0051	0.1293	52
C15	**Malignant neoplasm of esophagus**	519	0.0014	1.0236	0.0035	0.2296	24
C67	**Malignant neoplasm of bladder**	1,554	0.0043	1.0222	0.0047	0.1254	39
C34	**Malignant neoplasm of bronchus and lung**	1,427	0.0040	1.0208	0.0035	0.1010	17
C20	**Malignant neoplasm of rectum**	1,118	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 leukemia**	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 lymphoma**	587	0.0016	1.0104	0.0014	0.0824	14
C82	**Follicular (nodular) non-Hodgkin 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 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 leukemia	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

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

mesothelioma, to 271 loci for prostate (see **Table 1**; **Fig. 3A**); here, we are using the term heritability loci when referring to the nonoverlapping genomic regions tested by BAGHERA, which might also include multigene loci. Gene-level heritability across the selected 16 cancers has a long-tail distribution (**Fig. 3B**), with a median 16-fold increase compared with the genome-wide estimate, ranging from 4.4-fold for the *phosphodiesterase 4D (PDE4D)* gene locus to 276-fold for the *fibroblast growth factor receptor 2 (FGFR2)* gene locus in breast cancer. Interestingly, 87% of heritability loci show per-SNP heritability 10-fold higher than the genome-wide estimate. Only 3 loci have fold changes below 5 and more than 99% of loci with fold changes below 10 are found in the breast and prostate datasets, which have $h^2_{SNP} > 0.01$. On the basis of our simulations, our set of heritability loci is expected to have a limited number of false positives.

Interestingly, heritability loci represent less than 1% of all the loci in the genome, but they are significantly more than those harboring genome-wide significant SNPs (see Supplementary Material, Supplementary Figs. S17 and S18, and Supplementary Table S5); this result is consistent with cancer being polygenic. Although we identified a polygenic signal, heritability loci account for up to 38% of all the heritable risk (breast cancer), suggesting that a significant amount of heritability could be explained by only few loci across the genome (**Fig. 3A**). Consistent with our hypotheses, when we looked at the contribution of SNPs in intergenic regions, we did not find any heritability enrichment.

We then tested whether heritability loci were shared among multiple cancers to identify any potential genomic hotspot for pan-cancer heritability. We found that only 59 (≈ 8%) of the 783 heritability loci show a significant heritability enrichment in at least 2 cancers, and 8 (≈ 1%) in 3 or more (**Fig. 3C** and **D**). This observation is consistent with results from tumor sequencing studies, which have shown that pleiotropic effects are limited to few master regulators, such as *TP53* (35). Nonetheless, after performing literature curation, we found evidence for a cancer-

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

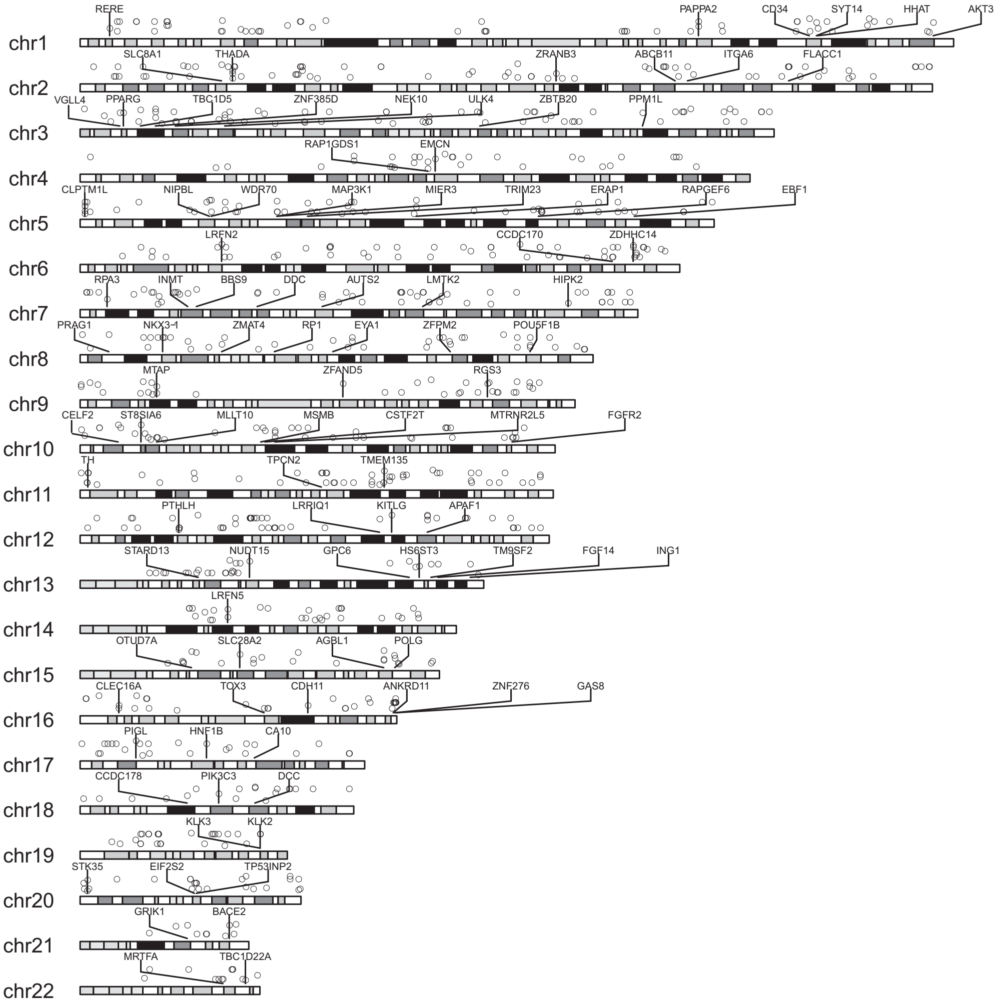


Figure 2.

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

mediating role for 7 of the 11 unique protein coding genes found in at least 3 cancers (see Supplementary Table 6), including 4 genes (*CLPTM1L, APAF1, THADA, AGBL1*) involved in apoptosis and 3 genes (*PCDH15, DLG2, POU5F1B*) involved in cell division, migration, and tumorigenesis (36, 37). It is important to note that the *cisplatin resistance-related protein 9* (*CLPTM1L*) is the heritability locus found in most cancers (4/16) and is one of the gene in the *5p15.33* locus (the other being *TERT*), which has been consistently associated with different cancer types (38).

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

CHGs are recurrently mutated in tumors

Tumor sequencing projects, including The Cancer Genome Atlas program and the PCAWG project, have identified a number of driver genes, which promote tumorigenesis when acquiring a somatic mutation.

There is also increasing evidence that genes harboring germline and somatic mutations can mediate cancer phenotypes (14, 39); thus, we tested whether CHGs are significantly enriched among known cancer driver genes. To do that, we obtained a curated list of driver genes using the COSMIC Cancer Gene Census (Supplementary Table 7). Interestingly, we found that a significant proportion of CHGs, 60 of 1, 146 (≈5%), are also known cancer driver genes

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

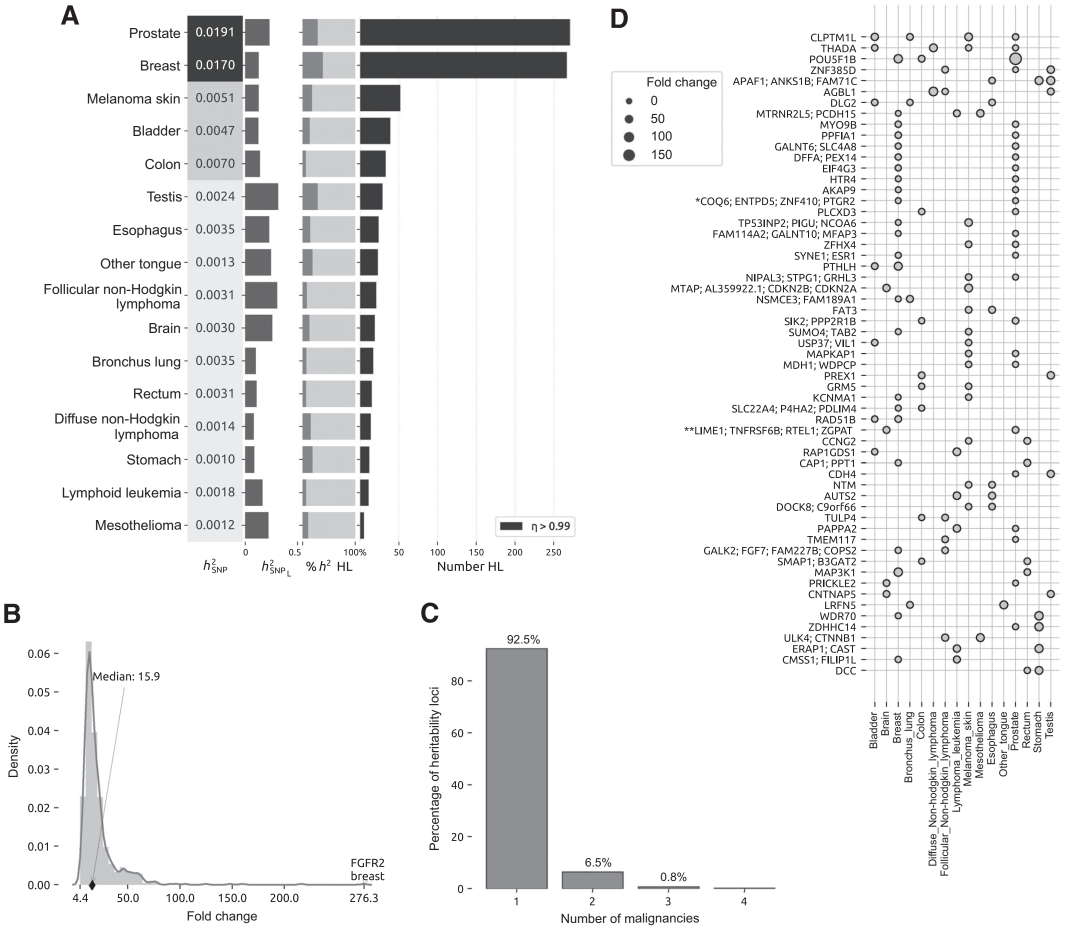


Figure 3.

Heritability loci across cancers in the UK Biobank. **A,** For each malignancy, we report the observed heritability (h^2_{SNP},left box), the heritability on the liability scale ($h^2_{SNP_L}$, dark barplot, between 0 and 0.5), the percentage of h^2_{SNP} explained by heritability loci (middle barplot; the percentage explained by heritability loci (HL) is highlighted with a darker shade), and the number of heritability loci (right barplot). **B,** Gene-level heritability distribution across heritability loci, expressed as fold change with respect to the genome-wide estimate. The *x*-axis is bound to the minimum and maximum values of fold change. We highlighted the top locus (FGFR2) and the median (15.9) fold change across all cancers. **C,** Percentage of cancer heritability loci associated with multiple cancers. Approximately 8% of heritability loci are common to multiple malignancies. **D,** Cancer heritability loci associated with multiple cancers. We report the 59 heritability loci common to at least two cancers; here, the size of the dot is proportional to the fold change of the locus in the specific cancer.

(OR = 1.75; $P : 1.3 \times 10^{-4}$). These genes include members of the p53 pathway, such as the *cyclin-dependent kinase inhibitor 2A* (*CDKN2A)*, the *tumor protein 63* (*TP63*), and *MDM4 regulator of p53* (*MDM4*), as well as genes mutated across multiple types of cancer, including *FGFR2* and the *anaplastic lymphoma kinase (Ki-1; ALK*) gene (**Fig. 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 (15), the Precision Oncology Knowledge Base (OncoKB; ref. 40), Memorial Sloan Kettering Impact and Heme gene panels (41), and the curated list of cancer genes by Vogelstein and colleagues (42). Here, we found that CHGs are significantly enriched in each cancer driver gene annotation analyzed, with an enrichment ranging from OR = 1.55 for the PCAWG annotation to OR = 2.47 for OncoKB tumor suppressors (Supplementary Table S7). 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.

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

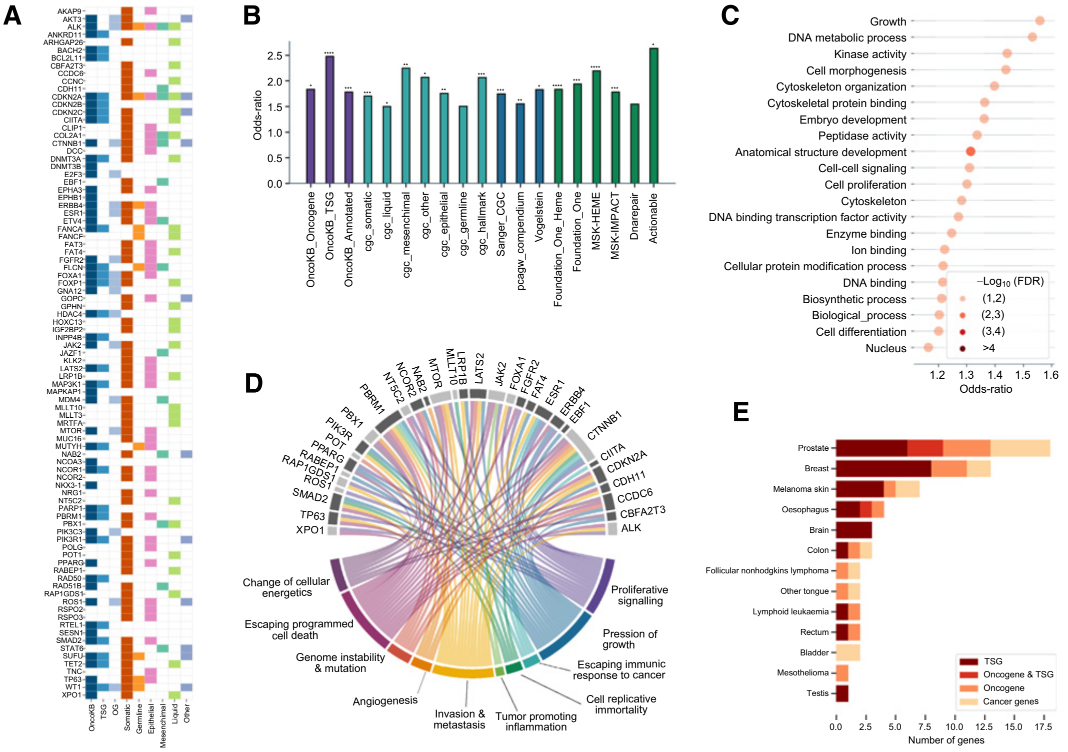


Figure 4.

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

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

CHGs underpin biological processes affecting tumorigenesis

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

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

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

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

Table 2. Gene ontology enrichment analysis of cancer heritability genes.

GO term	No. CHGs	OR	*P* value	FDR
Anatomic structure development	352	1.31	0.000044	0.006133
Kinase activity	126	1.44	0.000237	0.012169
Growth	84	1.55	0.000263	0.012169
DNA metabolic process	82	1.53	0.000481	0.016723
Cytoskeleton organization	120	1.39	0.000861	0.023924
Ion binding	431	1.22	0.001248	0.028903
Biosynthetic process	361	1.21	0.002711	0.041872
Biological_process	505	1.20	0.002224	0.041872
Cell morphogenesis	81	1.43	0.002419	0.041872
Cell proliferation	146	1.30	0.003404	0.047312
Cytoskeleton	141	1.28	0.005851	0.054216
Cellular protein modification process	275	1.21	0.004476	0.054216
Cell–cell signaling	123	1.30	0.005097	0.054216
Peptidase activity	103	1.33	0.005513	0.054216
DNA binding transcription factor activity	160	1.27	0.005068	0.054216
Enzyme binding	178	1.24	0.006568	0.057059
Cell differentiation	268	1.20	0.007776	0.063577
Embryo development	77	1.36	0.009437	0.069042
Cytoskeletal protein binding	77	1.36	0.009173	0.069042
Nucleus	347	1.16	0.014507	0.097916
DNA binding	174	1.21	0.014793	0.097916

Note: We report the gene ontology terms significantly associated with cancer heritability genes, at 10% FDR. For each term, we report the number of annotated CHGs, the odds ratio, the *P* value from the Fisher exact test, and the adjusted *P* value after applying the Benjamini–Hochberg procedure.

genome instability and mutation, and escaping cell death. Interestingly, we found 33 CHGs associated with at least one hallmark $(\mathrm{OR}: 2.062; P: 3 \times 10^{-4})$. Consistent with our previous analysis, cancer heritability loci are involved in escaping cell death, mediating proliferative signaling, invasion and metastasis (**Fig. 4D**; Supplementary Table S8). We then went further to understand whether CHGs mediate these cancer processes by acting either as tumor suppressor genes (TSG) or oncogenes (see **Fig. 4E**). To do that, we used the Precision Oncology Knowledge Base (OncoKB; ref. 40), a curated list of 519 cancer genes, including 197 TSGs, 148 oncogenes, and other cancer genes of unknown function. We found that 27 CHGs are tumor suppressors (OR: 2.47, $P: 7.9 \times 10^{-5}$), whereas 17 are reported as oncogenes (OR: 1.83; $P: 0.0198$), of which, 4 can function both as TSGs and as oncogenes (**Fig. 4A**, **D**, and **E**; Supplementary Tables S7 and S8); importantly, this result has been also confirmed when using the COSMIC Cancer Gene Census TSG annotation (OR: 2.036; $P: 2.07 \times 10^{-4}$). Tumor suppressor CHGs include well-known cancer driver genes, such as *CDKN2A and SMAD2*, which regulate cell growth, and DNA repair genes, such as *MUTYH* and *FANCA* (45).

Taken together, we found evidence that CHGs directly mediate processes underpinning tumorigenesis; interestingly, while we did not observe pleiotropic effects at genomic level, we found that CHGs 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 CHGs consistent with a two-hit model (46); while inherited mutations associated with oncogene activation are likely to be under purifying selection, mutations in TSGs can be observed at higher frequency because deleterious effects are only observed upon complete loss of function.

Discussion

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

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

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

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

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

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

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

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

V. Fanfani: Conceptualization, data curation, software, formal analysis, validation, investigation, methodology, writing–original draft, writing–review and editing. **L. Citi:** Validation, methodology, writing–review and editing. **A.L. Harris:** Validation, investigation, writing–review and editing. **F. Pezzella:** Validation, investigation, writing–review and editing. **G. Stracquadanio:** Conceptualization, software, formal analysis, supervision, validation, investigation, methodology, writing–original draft, writing–review and editing.

Acknowledgments

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

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received October 12, 2020; revised January 15, 2021; accepted March 2, 2021; published first March 17, 2021.

References

1. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nat Rev Cancer 2017;17:692–704.
2. Malkin D, Li FP, Strong LC, Fraumeni JF, Nelson CE, Kim DH, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 1990;250:1233–8.
3. Miki Y, Swensen J, Shattuck-Eidens D, Futreal P, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66–71.
4. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789–92.
5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461: 747–53.
6. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet 2017;101:5–22.
7. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 2012;44:312–8.
8. Eeles RA, Olama AAA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet 2013;45: 385–91.
9. Wang Z, McGlynn KA, Rajpert-De Meyts E, Bishop DT, Chung CC, Dalgaard MD, et al. Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor. Nat Genet 2017; 49:1141–7.
10. Law PJ, Berndt SI, Speedy HE, Camp NJ, Sava GP, Skibola CF, et al. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. Nat Commun 2017;8:14175.
11. Vijayakrishnan J, Studd J, Broderick P, Kinnersley B, Holroyd A, Law PJ, et al. Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia. Nat Commun 2018;9:1340.
12. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet 2017;18:117–27.
13. Stracquadanio G, Wang X, Wallace MD, Grawenda AM, Zhang P, Hewitt J, et al. The importance of p53 pathway genetics in inherited and somatic cancer genomes. Nat Rev Cancer 2016;16:251–65.
14. Zhang P, Kitchen-Smith I, Xiong L, Stracquadanio G, Brown K, Richter P, et al. Germline and somatic genetic variants in the p53 pathway interact to affect cancer risk, progression and drug response. bioRxiv Cold Spring Harbor Laboratory; 2019;835918.
15. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature 2020;578:82–93.
16. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era - Concepts and misconceptions. Nat Rev Genet 2008;9:255–66.
17. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, et al. Analysis of heritability and shared heritability based on genome-wide association studies for thirteen cancer types. J Natl Cancer Inst 2015;107: djv279.
18. Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, et al. The contribution of rare variation to prostate cancer heritability. Nature Genetics 2016;48:30–5.
19. Chen D, Cui T, Ek WE, Liu H, Wang H, Gyllensten U. Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability. Carcinogenesis 2015;36:992–8.
20. Litchfield K, Thomsen H, Mitchell JS, Sundquist J, Houlston RS, Hemminki K, et al. Quantifying the heritability of testicular germ cell tumour using both population-based and genomic approaches. Sci Rep 2015;5:13889.
21. Sapkota Y. Germline DNA variations in breast cancer predisposition and prognosis: a systematic review of the literature. Cytogenet Genome Res 2014; 144:77–91.
22. Fanfani V, Zatopkova M, Harris AL, Pezzella F, Stracquadanio G. Dissecting the heritable risk of breast cancer: from statistical methods to susceptibility genes. Semin Cancer Biol 2020;S1044–579X:30134–6.

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

23. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet 2013;9:e1003449.
24. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet 2018;50:621–9.
25. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 2015;47:1228–35.
26. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 2020;578:102–11.
27. Shi H, Kichaev G, Pasaniuc B. Contrasting the genetic architecture of 30 complex traits from summary association data. Am J Hum Genet 2016;99:139–53.
28. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291.
29. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–9.
30. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76–82.
31. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in python using PyMC3. PeerJ Computer Science 2016;2:e55.
32. Kennedy AE, Ozbek U, Dorak MT. What has GWAS done for HLA and disease associations? Int J Immunogenet 2017;44:195–211.
33. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018.
34. Jiang X, Finucane HK, Schumacher FR, Schmit SL, Tyrer JP, Han Y, et al. Shared heritability and functional enrichment across six solid cancers. Nat Commun 2019;10:431.
35. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell 2018;173:371–85.
36. Shao YW, Wood GA, Lu J, Tang Q-L, Liu J, Molyneux S, et al. Cross-species genomics identifies DLG2 as a tumor suppressor in osteosarcoma. Oncogene 2019;38:291–8.
37. Hayashi H, Arao T, Togashi Y, Kato H, Fujita Y, De Velasco M, et al. The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer. Oncogene 2015;34:199–208.
38. Rafnar T, Sulem P, Stacey SN, Geller F, Gudmundsson J, Sigurdsson A, et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nat Genet 2009;41:221–7.
39. Qing T, Mohsen H, Marczyk M, Ye Y, O'Meara T, Zhao H, et al. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. Nat Commun 2020;11:1–8.
40. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A precision oncology knowledge base. JCO Precis Oncol 2017;1:1–6.
41. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015;17:251–64.
42. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
43. Stuelten CH, Parent CA, Montell DJ. Cell motility in cancer invasion and metastasis: Insights from simple model organisms. Nat Rev Cancer 2018;18:296–312.
44. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell 2011;144:646–74.
45. Lange SS, Takata K, Wood RD. DNA polymerases and cancer. Nat Rev Cancer 2011;11:96.
46. Knudson AG. Mutation and cancer: Statistical study of retinoblastoma. Proc Natl Acad Sci U S A 1971;68:820–3.
47. Vosoughi A, Zhang T, Shohdy KS, Vlachostergios PJ, Wilkes DC, Bhinder B, et al. Common germline-somatic variant interactions in advanced urothelial cancer. Nat Commun 2020;11:1–3.
48. Zhang Y, Zhang Y, Sun K, Meng Z, Chen L. The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. J Mol Cell Biol 2019;11:1–3.

Published OnlineFirst March 17, 2021; DOI: 10.1158/0008-5472.CAN-20-3348

AACR American Association for Cancer Research

Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

The Landscape of the Heritable Cancer Genome

Viola Fanfani, Luca Citi, Adrian L. Harris, et al.

Cancer Res 2021;81:2588-2599. Published OnlineFirst March 17, 2021.

3.3 Conclusions

BAGHERA provides a pan-cancer compendium of genes explaining a portion of cancer risk larger than expected by chance. While the method is agnostic to the partitioning choice, gene-level estimates, with flanking regulatory regions, are better suited to generate novel testable hypotheses and are a reasonable trade-off between resolution and robustness of the estimates.

Being able to pinpoint specific heritability genes, we proceeded to understand their functional relevance for tumorigenesis and progression. Consistently with what we observed for single GWAS hits, most cancer heritability genes are involved in key processes underpinning cancer, suggesting that multiple low-penetrance variants might show phenotypic convergence at the process level and might have a synergistic effect on cancer risk. Indeed, extremely polygenic heritability is consistent with the effects of negative selection on high-penetrance germline variants [O'Connor et al. 2019].

Moreover, we have found evidence of the overlap between heritability loci and somatic driver genes, reinforcing the possibility that further, unexplored, genetic-environmental effects could explain tumorigenesis. While this topic is vastly understudied, especially for the difficulties in assessing the effects of SNPs, the interplay between germline and somatic mutation has been observed in urothelial cancer [Vosoughi et al. 2020] and it has been shown to mediate the effects of the *TP53* pathway on cancer progression and treatment [Zhang et al. 2021].

Taken together, this study is a stepping stone for further investigation of the role of low penetrance variants in cancer risk; accordingly, it provides interesting suggestions and exhibits some limitations.

Our results on cancer heritability loci are consistently hinting towards possible genetic-environmental interactions. SNPs might be mediating cancer risk both by directly regulating pathways disrupted by driver mutations [Zhang et al. 2021; Carter et al. 2017] or by affecting pathways that have a secondary impact on cancer risk, for instance by affecting the BMI [Di Giovannantonio et al. 2020]. This is consistent with the observed inverse correlation between genetic predisposition, measured as PRS, and the somatic mutation burden and age at diagnosis of the neoplasm [Zhu et al. 2016; Qing et al. 2020]. Further evidence in this direction might open novel avenues for prevention, early detection, and personalised treatments.

Nonetheless, we are aware of the current limitations of this work, some of which could be tackled by extending the method, while others would require additional experimental evidence.

First, BAGHERA does not explicitly include any functional signal, e.g. functional effects of the variants measurable by eQTLs [Geijn et al. 2021; Yao et al. 2020]. The prioritization of SNPs based on their functional annotation improves fine-mapping and recovery of heritability [Weissbrod et al. 2020; Kichaev et al. 2019; Wen, Pique-Regi, and Luca 2017] and cis-regulatory effects could be included as a prior on gene-level heritability. Conversely, we would hardly be able to include trans-regulatory effects in our non-overlapping mapping of SNPs.

One issue we have not addressed is the population representativity of this work; all our results have been focused on studies directly carried out on European populations. It is indeed well known that subjects non-European ancestry are underrepresented in GWAS and research is increasingly addressing this matter [Duncan et al. 2019; Fritsche et al. 2021]. On this issue, a multi-ancestry GWAS project, the Pan-UK Biobank (https://pan.ukbb.broadinstitute.org/), was recently started with the goal of testing for association subject non-European ancestry that are less frequent, but still present, in the UKBB. While the practical and ethical considerations on the lack of diversity in research are outside the scope of this work, it is worth noting and remarking that our panel of heritability genes has been found in the European populations and would need to be tested and validated in different ones. An interesting avenue would be to study whether gene-level heritability can find phenotypic convergence between different populations.

Eventually, this study can be considered exploratory and needs further computational and experimental validation of the results. The identified heritability loci could be used to prioritize gene-editing experiments to test the effects of variants [Whitington et al. 2016; Lawrenson et al. 2015; Dimitrakopoulos et al. 2019]. Similarly, we expect that other methods for the estimation of heritability might provide orthogonal evidence of the relevance of these loci for cancer risk.

4 From single experiments to system biology

The cellular machinery is a complex system where all biological functions are the result of interactions between multiple genomic elements [Vidal, Cusick, and Barabási 2011]. Complex phenotypes can only be understood when the cooperation of these elements is considered, from regulatory elements mediating gene expression to entire pathways involving multiple proteins.

High-throughput technologies enable the whole-genome characterisation of samples with remarkable detail [Rozenblatt-Rosen et al. 2020]; transcript abundance can be quantified at single-cell resolution [Stuart and Satija 2019] and genome-wide mutations are identified for thousands of samples [Abascal et al. 2021; Taylor-Weiner et al. 2019]. However, these experiments provide snapshots of the status of a DNA sequence, a transcript, or protein, in isolation, without any direct measure of how they interact with each other.

Similar to how NGS experiments capture different properties of DNA, transcripts, and proteins, biological networks can describe a wide spectrum of relationships between them, see Fig. 4.1. The nodes of these networks represent genomic elements such as proteins, transcripts, transcription factors. The edges between the nodes are instead specific interactions between them; in Protein-Protein Interaction (PPI) networks each link describes the physical contact between proteins [Rual et al. 2005; Lehner and Fraser 2004], while in regulatory networks they capture how DNA elements interact with genes to regulate expression [Ravasi et al. 2010; Califano et al. 2012]. The direct quantification of these interactions in an experiment is non-trivial and hard to scale, with the search space growing quadratically with respect to the number of genomic elements. However, in the last two decades, considerable efforts have been made to collect and aggregate data to recapitulate whole-genome interactome [Stark et al. 2006; Chatr-Aryamontri et al. 2017; Luck et al. 2020; Huttlin et al. 2021; Huttlin et al. 2015; Szklarczyk et al. 2019].

While the phenotype of cancer cells is characterized by known, recurrent, hallmark processes [Hanahan and Weinberg 2011], the background genetics of tumors reveals a polygenic and heterogeneous disease [Vogelstein et al. 2013]. In the previous chapters, we have shown that different low-penetrance germline variants tend to target the same biological processes [Stracquadanio et al. 2016], often occurring in genes that functionally interact with many others [Fagny et al. 2020] and in regulatory elements [Lawrenson et al. 2015]. Hence, the status of single genes may be insufficient to detect cancer-driving mechanisms. Conversely, understanding how genes and proteins interact can elucidate how polygenic signals arise. Interaction data is key to inform how heterogenous mutations target the same biological processes, see Fig. 1.1, and to interpret the results of multi-modal experiments. Unsurprisingly, networks have been widely used in cancer studies to gain insight into different aspects of tumor biology [Ozturk et al. 2018]; for the identification of cancer-driving mutations, genes, and pathways [Ruffalo, Koyutürk, and Sharan 2015; Khalighi, Singh, and Varadan 2020; Leiserson et al. 2015; Jia and Zhao 2014; Hristov and

Singh 2017; Hristov, Chazelle, and Singh 2020; Reyna, Leiserson, and Raphael 2018; Tuncbag et al. 2016; Paull et al. 2013; Vanunu et al. 2010; Mezlini and Goldenberg 2017; Cho et al. 2016; Horn et al. 2018; Silverbush et al. 2019], to stratify patients based on their genetic and transcriptional signatures [Hofree et al. 2013; Vandin, Upfal, and Raphael 2011; He et al. 2017; Lopes-Ramos et al. 2018; Hansen and Vandin 2016], for the identification of therapeutics opportunities [Margolin et al. 2006; Alvarez et al. 2018; Zitnik et al. 2019; Ruiz, Zitnik, and Leskovec 2021].

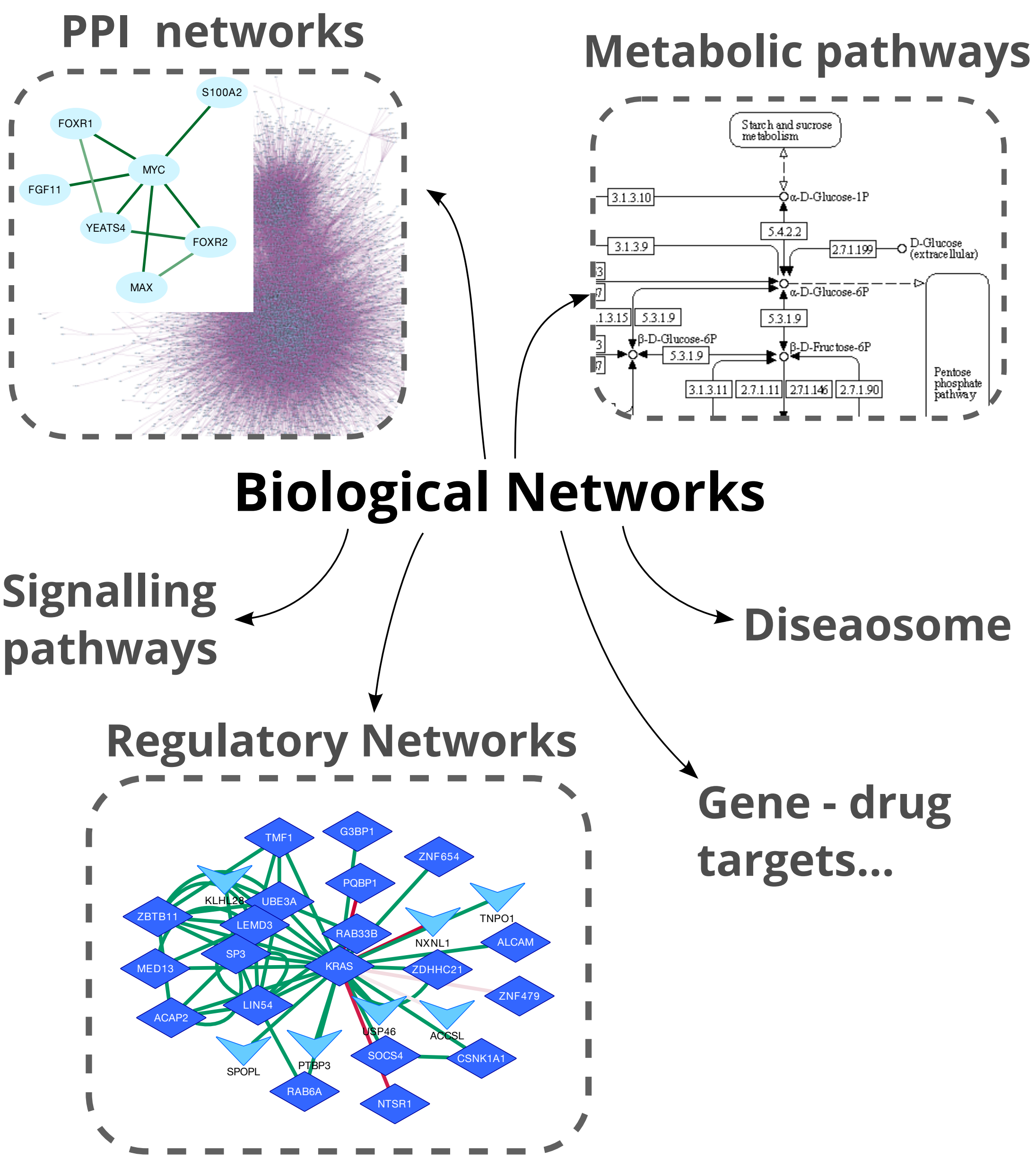


Figure 4.1: *Biological Networks. There are multiple ways of representing genomic elements as graphs, in each case nodes and edges between them represent specific interactions [Pillich et al. 2017]. PPI networks are large-scale interactomes [Luck et al. 2020], where each link represents physical contact between the proteins. Here we show a detail of the neighbourhood of the MYC Proto-Oncogene [Huttlin et al. 2021]. Metabolic pathways are detailed and directed graphs that describe the reactions involved in a metabolic process, like glycolysis [Kanehisa and Goto 2000], and sometimes a transfer function is available for each reaction. Regulatory Networks describe the interactions between genes and transcription factors, or other regulatory elements. Here we show the regulatory elements for KRAS specific for prostate cancer (PRAD), [Aytes et al. 2014]. This is however a non-exhaustive list of biological networks, as many others have been defined and used in the literature.*

4.1 Protein-Protein Interaction Networks

Protein-Protein Interaction (PPI) networks provide mechanistic insight into how proteins physically interact with each other [Vidal, Cusick, and Barabási 2011; Costanzo et al. 2010]. Proteomics methods such as yeast two-hybrid (Y2H) [Rolland et al. 2014; Rual et al. 2005] and affinity purification followed by mass spectrometry (AP-MS) [Huttlin et al. 2015] are able to detect protein-protein contacts and protein complexes, and they can be scaled up to detect PPIs between multiple protein coding genes. Since we can now scan and aggregate the interactions between thousands of proteins, PPI networks aim to provide an exhaustive reference of possible contacts on top of which novel hypotheses on cellular functions, aided by orthogonal experimental data, can be formulated [Luck et al. 2020]. We refer to these large networks, when they include almost all known proteins, as interactomes.

The main issue with the generation of complete PPI maps is scalability. Indeed, AP-MS and Y2H require careful planning and parallelization to obtain system-level interactomes. Moreover, these methods have low sensitivity ($\sim 20\%$ [Venkatesan et al. 2009]) which contributes to hinder discovery. Commonly used networks [Szklarczyk et al. 2019; Stark et al. 2006] overcome incompleteness by aggregating data from published research, and report a collection of known interactions. For instance, BIOGRID [Stark et al. 2006; Chatr-Aryamontri et al. 2017], which provides interactomes for multiple organisms, has a human PPI network compiled from more than 30 thousand published articles with more than $26,000$ genes and $500,000$ edges between them.

Resources such as BIOGRID [Stark et al. 2006] and STRING [Szklarczyk et al. 2019] have enabled researchers to study interactions between thousands of genes. However, since they aggregate data from a deluge of experiments found in the literature, they can be noisy and are prone to false discoveries [Von Mering et al. 2002]. Both BIOGRID and STRING mitigate this issue by curating their data, and they provide confidence scores for each interaction, that can be then used to calibrate the analysis model [Chatr-Aryamontri et al. 2017]. Moreover, novel efforts, such as the Huri [Luck et al. 2020] network, recently published, enabled to scan the complete ORFeome ($\sim 17,000$ genes) with the same protocol, obtaining $52,548$ interactions between 8272 proteins. The HuRI network has been shown to provide more reliable data than literature curated resources, with many of the interactions validated by orthogonal data. Further efforts to generate whole-ORFeome maps [Huttlin et al. 2021] will improve the reliability of PPI networks and will mitigate the issues of aggregated interactomes.

Finally, PPI networks also have intrinsic limitations that require careful consideration before their use and interpretation [Futschik, Chaurasia, and Herzel 2007; Peng et al. 2016]. Indeed, the experimental protocols used to draw PPI networks are not representative of the actual environment proteins face; expression and translation are tissue- and condition-dependent, and it is likely that many of the observed interactions will not be actually occurring in the cell, might be occurring only in some tissues, or might be happening at specific times and locations. Hence, while large scale PPI networks provide thousand of possible, experimentally observed interactions at the whole ORFeome level, orthogonal and functional data are key to infer the actual pattern of interactions and to link it to phenotypes [Kuenzi and Ideker 2020; Silverbush and Sharan 2019; Zheng et al. 2021; Haenig et al. 2020].

4.2 Using interactome topology to test omics data

As we described in the previous section, PPI networks aim at encompassing the physical contacts of the whole ORFeome, the protein-coding genome. Thus, given the availability of large-scale interaction data and whole-genome (exome or proteome) experiments, single-gene information can be combined onto the interactome to expand the interpretability of a single experiment.

The fundamental idea behind most omics experiments is to quantify an appropriate signal, representing the biological properties of interest, for all genomic elements, to then detect deviations and variability from what is expected [Meyerson, Gabriel, and Getz 2010]. For instance, differential expression experiments aim at finding transcriptional differences between two or more conditions, by comparing transcript abundance and detecting the genes with condition-dependent expression levels [Love, Huber, and Anders 2014; Pimentel et al. 2017; McCarthy, Chen, and Smyth 2012]. While their data and modalities are very different from each other, most NGS analyses attribute a final score for each gene and statistically test them. Here we use the term ‘gene’ broadly as a synonym of transcripts, exomes, loci, to refer to any genomic element that is tested by the experiment.

High-throughput experiments are often statistically tested with ‘standard’ pipelines, specific for the modality, whose results are readily comparable to published literature. Downstream, this data is characterised by mapping the results onto a functional annotation database; pathway analysis tools [Jassal et al. 2020; Subramanian et al. 2005] are ubiquitously used to reveal the processes disrupted by the genomic and transcriptomic aberrations found in cancer. Networks, either PPIs, metabolic, signaling pathway graphs, are used to enhance the pathway analysis by adding information on the connectivity between the geneset and the pathway [Ihnatova, Popovici, and Budinska 2018; Ma, Shojaie, and Michailidis 2019]. Consistently with the poligenicity hypothesis, pathway analysis methods map and aggregate gene-level data to reveal functional patterns that would otherwise be undetected. However, these methods rely only on predefined definitions of pathways and oftentimes use specific graphs that do not capture the entirety of the interactions.

Conversely, interactomes can be used to characterise a set of genes in terms of their connectivity, cross-talk, topological proximity without any bias on the subgraph of interest. NGS results are mapped onto the corresponding protein in the network to test their topological properties and infer functional properties [Jeggari and Alexeyenko 2017; Liao et al. 2019]; highly mutated genes that have a large degree could be master regulators and a group of differentially expressed genes that are strongly interacting with each other might reveal a novel pathway [Menche et al. 2015].

With the deluge of available large-scale datasets and publicly curated genesets, we reasoned that network topology methods could be routinely used to scan experimental results. Surprisingly, many tools often used in the literature cannot be integrated into bioinformatics pipelines; they are web applications, visualization plugins, or scarcely documented scripts, that are better suited for targeted analyses.

We hence developed PyGNA, which is a Python package and command-line interface tool that enables the statistical analysis of the topological properties of genesets and networks. PyGNA implements multiple statistics that have been shown to be representative of key properties of a network, such as the average internal degree and the diffusion scores. PyGNA allows to statistically test the properties of a single geneset, e.g. whether the set of genes is strongly interacting with each other, or to assess the connectivity between two genesets, e.g. to infer comorbidities.

The manuscript in the next section details PyGNA's implementation and describes the methods used for statistical testing. Moreover, PyGNA is able to generate simulated networks and genesets under different models. In the manuscript, we use the simulated data to benchmark and compare all statistical tests' robustness and specificity. Eventually, we retrieved differential expression datasets from the TCGA consortium for 6 cancer types and we applied PyGNA showing how network topology testing can become a routine step for the characterisation of high-throughput experiments.

4.3 PyGNA: a unified framework for geneset network analysis

The whole manuscript has been drafted by V. Fanfani, with the contributions of F. Cassano, and with the supervision and the contributions of G. Stracquadanio. V.Fanfani first developed the tool and carried out the data analysis. F.Cassano contributed to code reformatting and to the analysis of the TCGA datasets.

BMC Bioinformatics

SOFTWARE Open Access

PyGNA: a unified framework for geneset network analysis

Viola Fanfani, Fabio Cassano and Giovanni Stracquadanio*

*Correspondence:
giovanni.stracquadanio@ed.ac.uk
School of Biological Science, The University of Edinburgh, Edinburgh EH9 3BF, UK

Abstract

Background: Gene and protein interaction experiments provide unique opportunities to study the molecular wiring of a cell. Integrating high-throughput functional genomics data with this information can help identifying networks associated with complex diseases and phenotypes.

Results: Here we introduce an integrated statistical framework to test network properties of single and multiple genesets under different interaction models. We implemented this framework as an open-source software, called Python Geneset Network Analysis (PyGNA). Our software is designed for easy integration into existing analysis pipelines and to generate high quality figures and reports. We also developed PyGNA to take advantage of multi-core systems to generate calibrated null distributions on large datasets. We then present the results of extensive benchmarking of the tests implemented in PyGNA and a use case inspired by RNA sequencing data analysis, showing how PyGNA can be easily integrated to study biological networks. PyGNA is available at http://github.com/stracquadaniolab/pygna and can be easily installed using the PyPi or Anaconda package managers, and Docker.

Conclusions: We present a tool for network-aware geneset analysis. PyGNA can either be readily used and easily integrated into existing high-performance data analysis pipelines or as a Python package to implement new tests and analyses. With the increasing availability of population-scale omic data, PyGNA provides a viable approach for large scale geneset network analysis.

Keywords: Geneset Network Analysis, Biological Networks, Network analysis workflow

Background

The availability of high-throughput technologies enables the characterization of cells with unprecedented resolution, ranging from the identification of single nucleotide mutations to the quantification of protein abundance [1]. However, these experiments provide information about genes and proteins in isolation, whereas most biological functions and phenotypes are the result of interactions between them. Protein and gene interaction information are becoming rapidly available thanks to high-throughput screens [2], such as the yeast two hybrid system, and downstream annotation and sharing in public databases [3, 4]. Thus, it is becoming obvious to use interaction data to map single gene information to biological pathways.

Integrating interaction information with high throughput experiments has proven challenging. The vast majority of existing analytical methods are based on the concept of over-representation of a candidate set of genes in expert curated pathways or networks [5, 6]; however, this approach is strongly biased by the richer-get-richer effect, where intensively studied genes are more likely to be associated with a pathway [7], ultimately limiting the power of new discoveries. Many methods have now been proposed to directly integrate network information for function prediction [8–10], module detection [11], gene prioritization [12] and structure recognition [13]. However, results are usually sensitive to the underlying network interaction model used and test statistics [14], and performing analyses across different tools is not feasible, as the vast majority of this software comes either as a web application or visualization plugins. While web applications are simple to use for targeted analyses, they are also difficult to integrate in high-throughput data analyses pipelines.

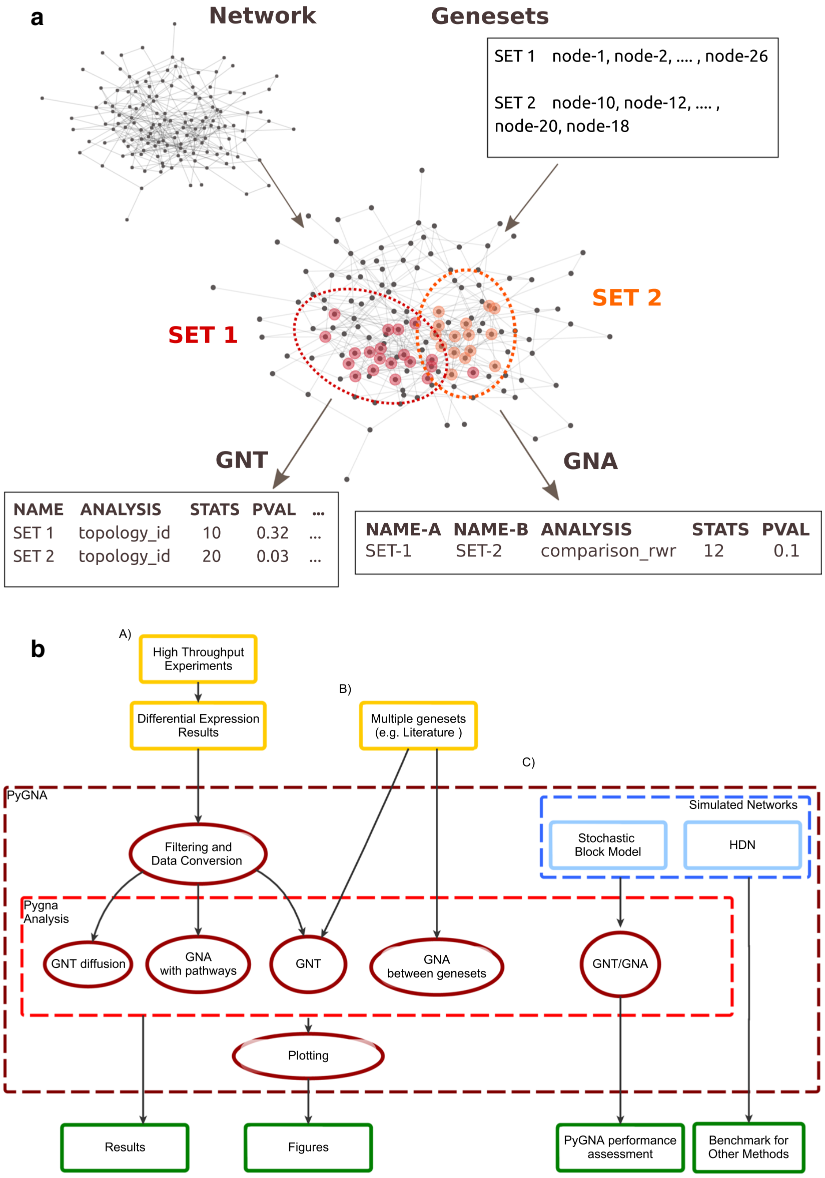
With the increasing availability of biological interaction resources and the development of standardized high-throughput analysis pipelines, a unified and easy to use framework for network characterization of genes and proteins could generate useful information for downstream experimental validation.

Here we build on recent advances in network theory to provide an integrated statistical framework to assess whether a set of candidate genes (or geneset) form a pathway, that is genes strongly interacting with each other. We then extended this framework to perform comparisons between two genesets to find similarities with other annotated networks, as a way to infer function and comorbidities. We called our statistical tests geneset network topology (GNT) and geneset network association (GNA) tests, respectively (Fig. 1a). We implemented our tests into a Python package, called Python Gene Network Analysis (PyGNA). It is important to note that the tests implemented in our software are not an exhaustive list of all the approaches presented in literature; here we favoured well established models with test statistics easy to interpret [14]. Nonetheless, PyGNA provides a flexible API to implement and benchmark new network-based statistical tests, while taking advantage of our data processing and statistical testing framework.

We tested the GNT and GNA tests implemented in PyGNA on synthetic datasets to assess the performance (true positive rate and false positive rate). We then present how to use PyGNA to analyse high-throughput RNA sequencing data generated by The Cancer Genome Atlas (TCGA, [15]) and how to interpret network analysis results.

PyGNA is released as an open-source software under the MIT license; source code is available on GitHub (http://github.com/stracquadaniolab/pygna) and can be installed either through the PiP or Anaconda package managers, and Docker. Our software is designed with modularity in mind and to take advantage of multi-core processing available in most high-performance computing facilities. PyGNA facilitates the integration with workflow systems, such as Snakemake [16], thus lowering the barrier to introduce network analysis in existing pipelines.

The manuscript is organized as follows; "Methods" section describes the statistical network framework implemented in PyGNA, whereas "Implementation" section describes PyGNA APIs and command line interface (CLI) options. In "Results" section,


a
Network
Genesets
SET 1 node-1, node-2, ... , node-26
SET 2 node-10, node-12, ... , node-20, node-18
SET 1
SET 2
GNT
GNA
b
A)
High Throughput Experiments
Differential Expression Results
B)
Multiple genesets (e.g. Literature)
C)
PyGNA
Simulated Networks
Stochastic Block Model
HDN
Filtering and Data Conversion
Pygna Analysis
GNT diffusion
GNA with pathways
GNT
GNA between genesets
GNT/GNA
Plotting
Results
Figures
PyGNA performance assessment
Benchmark for Other Methods

NAME	ANALYSIS	STATS	PVAL	...
SET 1	topology_id	10	0.32	...
SET 2	topology_id	20	0.03	...

NAME-A	NAME-B	ANALYSIS	STATS	PVAL
SET-1	SET-2	comparison_rwr	12	0.1

Fig. 1 The PyGNA analysis workflow. **a** Outline of the GNT and GNA tests. Given an input network, PyGNA maps genes to network nodes, performs GNA and GNT tests, and then outputs the results in CSV format. **b** Complete workflow. We recognize three main use-cases where PyGNA can be used, including (i) network analysis of high-throughput experiments, (ii) network analysis of curated genesets and iii) simulations of networks and genesets for algorithms benchmarking. PyGNA can perform GNT analysis on single or multiple genesets, along with GNA analysis to identify network associated with other genesets or pathways. Results are provided as CSV files and as high quality PDF figures

we present benchmarking results on simulated data and how to apply PyGNA to analyse RNAseq experiments. We conclude by discussing how PyGNA compares to other existing tools and why it represents an advancement for geneset network analysis.

Methods

We hereby introduce basic notation and properties for network analysis, describing interaction models, test statistics and hypothesis testing methods implemented in PyGNA.

Let $G = (V, E)$ be a network, or graph, with $|V|$ nodes and $|E|$ edges. Let A be a matrix $|V| \times |V|$, with $A_{ij} = 1$ if there is an edge between node i and j and 0 otherwise; we denote A as the adjacency matrix of the network G. We hereby consider only undirected graphs, thus the adjacency matrix is symmetrical $A_{ij} = A_{ji}$; however, all the tests we present can be applied to directed networks and weighted networks. Moreover, unless otherwise stated, we consider only the largest connected component (LCC) of the network; while this is not strictly necessary, distance measures are often not informative when computed over disconnected graphs. We denote as degree of a node i, deg(i), the number of edges associated with it. In this context, nodes represent genes or proteins, whereas edges the intervening interactions, e.g. physical, genetic interactions.

Let $S = s_1, \ldots, s_n$ be a geneset consisting of n genes, we want to quantify the strength of interaction between genes in the geneset (geneset network topology, GNT) and with genes in another geneset (geneset network association, GNA).

Interaction models

We denote as interaction model, a function that quantifies the strength of interaction between any two nodes in a network. Here we introduce three interaction models with different properties and complexity.

A direct interaction model assumes that two nodes interact only if there is an edge between them; this is the most efficient model to evaluate as it requires only the inspection of the adjacency matrix.

Under a shortest path interaction model, instead, we assume that the strength of interaction between two genes is a function of their distance on a network G, that is closer genes are more likely to interact. Thus, we denote with $i \to j$ a path in G from node i to node j, whose length, l_{ij}, is the number of edges from i to j. We then quantify the strength of interaction between two genes, i and j, as the length of the shortest path from i to j, denoted as s_{ij}; w.l.o.g, shortest paths can be also computed over directed and weighted networks.

Finally, we introduce a probabilistic model of gene interactions, namely the Random Walk with Restart (RWR) model. Let W be a stochastic matrix inferred from the adjacency matrix A, the probability of reaching node i from node j after k steps is $(W^k)_{ij}$ [17]. However, for k big enough, the probability of interaction between nodes converges to a quantity proportional to the degree of the nodes, thus neglecting local structure information. We here instead consider a random walk with restart model (RWR), where it is possible to return to the starting node with fixed probability β (set to 0.85 unless otherwise stated [18]). We can then estimate analytically the probability of interaction at steady state as follows:

$$H = \beta(I - (1 - \beta)\bar{A})^{-1} \quad (1)$$

where $\bar{A}$ is the normalized adjacency matrix obtained as $\bar{A} = AD^{-1}$, with D being the diagonal matrix of node degrees. In this case, the matrix H can be interpreted as the heat exchanged between each node of the network [11]. It is also worth noting that the above formulation is agnostic to direction and weights of the edges.

These three interaction models capture different topological properties. Direct models provide information about the neighborhood of a gene and its observed links. However, they might not be sufficiently powered to detect mid- and long-range interactions, thus statistics defined under these models are usually sensitive to missing links. Conversely, modelling gene interactions using shortest path provides a simple analytical framework to include local and global awareness of the connectivity. However, this approach is also sensitive to missing links and small-world effects, which is common in biological networks and could lead to false positives [19]. Propagation models provide an analytical model to overcome these limitations, and have been shown to be robust for biological network analysis [20]. While its interpretation is not necessarily straightforward, the RWR model is more robust than the shortest path model, because it effectively adjusts interaction effects for network structure; it rewards nodes connected with many shortest paths, and penalizes those that are connected only by path going through high degree nodes.

Based on the above interaction models, we have implemented and tested different statistics, which are described in detail below.

Geneset network topology statistics

Let $S = s_1, \ldots, s_n$ be a geneset of n genes, each mapped to a node in $G = (V, E)$. We are interested in testing whether the strength of interaction between nodes of the geneset is higher than expected by chance for a geneset of the same size.

Under a direct interaction model, the importance of a geneset S can be quantified as the number of edges connecting each node in S to any other node in the network; we refer to this quantity as the total degree of the node. Thus, we define the total degree statistic for a geneset S as:

$$T_{TD} = \frac{1}{n} \sum_{i \in S} deg(i) \tag{2}$$

While T_{TD} could be helpful to have an idea of how relevant and well characterized the nodes in the geneset are, we do not expect this statistic to be informative on the strength of interaction withing a geneset.

Conversely, with the direct interaction model, the strength of interaction for a geneset S can be quantified as the number of edges connecting each node in S to any other node in the geneset; we refer to this quantity as the internal degree of the node. Thus, we define the internal degree statistic for a geneset S as:

$$T_{ID} = \frac{1}{n} \sum_{i \in S} \frac{deg(i, S)}{deg(i)} \tag{3}$$

where $deg(i, S)$ is the internal degree of gene i in geneset S. In practice, the internal degree statistic captures the amount of direct interactions between genes in a geneset, and thus a geneset showing a network effect should have T_{ID} values close to 1. However, the main limitation of this model lies in the fact that it only captures direct interactions, whereas biological networks are usually characterized by medium and long range interactions.

Another way to assess the strength of a network effect is the size of the largest connected components of the graph induced by the geneset S, hereby denoted as T_M. A main concern regarding direct interaction methods is that they could fail in presence of missing links, which is a well-known problem in biological networks analysis, where experimental screens are often not sensitive enough to detect all existing gene/protein interactions.

A shortest path interaction model allows to overcome this limitation by explicitly taking into account the distance between nodes. Here we define the test statistic T_{SP} for the geneset S as follows:

$$T_{SP}(S) = \frac{1}{n}\sum_{i=1}^{n} \min_{j \in S} s_{ij} \tag{4}$$

which is the average of the minimum distance between each gene and the rest of those in S [21].

Conversely, under a RWR model, we can consider $h_{ij} \in H$ as the heat transferred from node i to node j, which can be used as a measure of interaction strength between the nodes in the geneset S, as follows:

$$T_H(S) = \sum_{i,j \in S, i \neq j} h_{ij} \tag{5}$$

Geneset network association statistics

Let S_1 and S_2 be two geneset with n and m genes respectively, we want to estimate the association between S_1 and S_2 as a function of the strength of interaction between their nodes.

Under a shortest path model, the association statistics U_{SP} is defined as follows:

$$\begin{aligned} U_{SP}(S_1, S_2) = & \frac{1}{n+m} \sum_{i \in S_1} \min_{j \in S_2} s_{ij} + \sum_{j \in S_2} \min_{i \in S_1} s_{ij} \\ & - \frac{1}{2}(T_{SP}(S_1) + T_{SP}(S_2)) \end{aligned} \tag{6}$$

whereas, under a RWR model, we measure association as a function of the heat, U_H, transferred between the two genesets as follows:

$$U_H(S_1, S_2) = \sum_{i \in S_1, j \in S_2} h_{ij} + h_{ji} \tag{7}$$

where we consider also the heat withhold by a gene, when there are overlapping genes between S_1 and S_2.

Hypothesis testing

The topological and association statistics are ultimately used for hypothesis testing. To do that, we need a calibrated null distribution to estimate whether the observed statistics are more extreme than what expected by chance. Closed form definition of null distributions is possible only for very simple network models, which are often unrealistic. Therefore, we reverted to a bootstrap procedure to estimate null distributions of the test statistics, conditioned on the geneset size; while this approach can be computationally taxing, in practice, we observed that ≈ 500 bootstrap samples are sufficient to obtain a stable distribution (see Additional file 1).

Thus, w.l.o.g, let Q be the null distribution of the test statistic q estimated for a geneset of size n, and $\bar{q}$ the observed value. It is possible to derive an empirical p-value as follows:

$$P(\bar{q} \geq Q) = \frac{(\sum_{i=1}^{|Q|} I(Q_i \geq \bar{q})) + 1}{|Q| + 1} \tag{8}$$

where I is the indicator function returning 1 if and only if the evaluated condition is true, and unit pseudo-count is added for continuity correction. It is straightforward to adapt this formula to the case of testing whether a test statistic is smaller than expected by chance.

The default sampler generates null distributions by sampling nodes uniformly at random. However, certain metrics might be particularly sensitive to local network structure, especially when they solely rely on degree-related statistics to characterize a geneset. To overcome this problem, we also implemented an additional sampler that generates null distributions matching the degree distribution of the tested dataset.

For the GNA tests, it is important to note that we are now dealing with two genesets. Hence, a null distribution can be computed either by sampling two random genesets or by sampling only one of the two; we recognize that the latter is more conservative, and is recommended when checking for association with known pathways (see Additional file 1).

Benchmarking geneset network tests

Rigorous benchmarking of network analyses tools is challenging, because there is no ground truth for geneset network analysis [14].

Stochastic block models (SBM) have been shown to be a reasonable model for analyzing biological networks [22]; importantly, since SBM define a generative process over networks, they can be used to create networks with controllable features, including modules (also often referred as clusters). Let $M : k \times k$ be a stochastic block model with k blocks, where M_{ij} represent the probability of a node in block i to be connected to (or interact with) a node in block j. A new network with n nodes can be generated by assigning each node to a block and adding edges probabilistically using the block model matrix. It is straightforward to note that if $M_{ii} >> M_{ij}$ for any j, the genes in block i are likely to show a network effect. Hence, by modulating the values on the diagonal of the block model matrix, we can assess the performance of GNT tests by analyzing the genesets made of the genes in a block. Conversely, we expect to find a significant association between two blocks i and j if $M_{ij} >> M_{kl}$, with $i, j \neq k, l$. By parametrizing the off-diagonal terms of the block model matrix, it is possible to assess the performance of GNA tests (see Additional file 1 for a graphical representation of the SBMs).

While the SBM are useful to simulate networks with controllable structures, they are difficult to adapt to modelling networks with highly connected nodes (hubs), which are common in biological networks. Thus, here we introduce a stochastic generation procedure to build networks with hubs, which can then be used for assessing the performances of GNT tests. We hereby describe each model in detail.

SBM for GNT benchmarking

We use the SBM framework to simulate a network with k blocks, with a baseline probability of interaction within and between blocks, p_0. We then select $k^+ < k$ blocks from the SBM matrix and set their within probability of connection $M^+_{ii} = \alpha p_0$, where $\alpha > 1$ is a scaling factor controlling the strength of interaction of the genes within block i compared to the rest of the genes in any other block. Intuitively, each of the k^+ blocks represents a geneset with a significant network effect, thus a robust GNT test should be able to detect them.

Ultimately, by varying the size of highly connected blocks, the baseline probability of interaction p_0 and the strength of interaction α, it is possible to assess the power, true positive rate (TPR) and false positive rate (FPR) of GNT tests under different conditions.

SBM for GNA benchmarking

Similar to the approach outlined for GNT benchmarking, we used the SBM framework to generate network with multiple gene clusters to assess the performance of GNA tests.

We use the SBM framework to simulate a network with k blocks, with a baseline probability of interaction within and between blocks, p_0. We then selected k^+ blocks at random and set their within block connection probability to $M^+_{ii} = \alpha p_0$ and their between blocks connection to $M^+_{ij} = \gamma p_0$ for $i \neq j$ and $\alpha, \gamma > 1$. We then reparametrize γ as a function α, in order to control the relationship between the within and between block connection probability. Let $\beta = \gamma/(\alpha - 1)$, we can set the between block connection probability as $M^+_{ij} = p_0 + \beta p_0(\alpha - 1)$. With this parametrization, we can directly simulate 3 different scenarios:

1 if $\beta = 0 \Rightarrow M^+_{ij} = p_0$, the connection probability between blocks is equal to the baseline, thus genes in a block are highly connected.
2 if $0 < \beta < 1 \Rightarrow p_0 < M^+_{ij} < M^+_{ii}$, then the connection probability between the blocks is higher than the baseline, and thus we obtain assortative genesets.
3 $\beta > 1 \Rightarrow M^+_{ij} > M^+_{ii}$, then we have non assortative genesets, thus we expect them to be detected by a GNA test.

After building a network, we then generate genesets by selecting two distinct blocks, i, j, with m nodes each, and add $\pi \times m$ nodes from block i and $(1 - \pi) \times m$ nodes from block j; for simplicity, we picked genes from blocks containing the same number of genes. The GNA testing is then performed between the SBM blocks and the novel mixture blocks. By varying the size of highly connected blocks and their interaction probability, along with the geneset composition, it is possible to assess the true positive rate (TPR) and false positive rate (FPR) of GNA tests.

High degree nodes model for GNT benchmarking

The high degree nodes (HDN) model generates networks with a controllable number of hubs, n_{hd}, whose probability of connection with another node, p_{hd}, is higher than the baseline probability p_0 assigned to any other node in the network. The model is fully specified by four parameters, namely the number of nodes in the network, n, the number of HDN nodes, n_{hd}, the baseline connection probability, p_0, and the HDN connection probability, $p_{hd} > p_0$.

In order to benchmark GNT tests in presence of HDN nodes, we created geneset as a mixture of HDNs and non HDN nodes; we denoted these genesets as extended genesets. Specifically, each geneset is made of $\pi_{hd} \times n_{hd}$ nodes, with $\pi_{hd} \in (0, 1]$, and $\rho\pi_{hd}n_{hd}$ random high degree nodes, where ρ is the ratio between high degree nodes and other nodes in the network (see Additional file 1 for a graphical representation).

With the HDN model, we can replicate a common scenario where the tested geneset is made of a few master regulators and many, possibly, unrelated genes. Here, the idea is that a robust GNT test should have a low false positive rate, even when observed statistics might be skewed by few highly connected nodes.

Implementation

PyGNA is implemented as a Python package and can be used as a standalone command-line application or as a library to develop custom analyses. In particular, our framework is implemented following the object oriented programming paradigm (OOP), and provides classes to perform data pre-processing, statistical testing, reporting and visualization. Here we provide an overview of the package structure and available interfaces, although the complete API documentation is available at: https://github.com/stracquadaniolab/pygna. Our basic workflows are summarized in Fig. 1b.

Input/output functions

Our software can read genesets in Gene Matrix Transposed (GMT) and text (TXT) format, while networks can be imported using standard Tab Separated Values (TSV) files, with each row defining an interaction. For diffusion analysis, instead, we require a Comma Separated Value (CSV) file specifying weights for each gene. It is important to note that parsers for new data can be easily implemented by extending the ReadData abstract class.

To facilitate the integration in bioinformatics pipelines, e.g. downstream analysis of DESeq2 results [23], we implemented a Utility class to enable input filtering, gene name conversion and GMT file creation.

PyGNA stores results as CSV files, for downstream manipulation and sharing, although new formats can be supported by extending the Output class. It is important to note that performing tests on large networks using either shortest path or random walk models is computationally taxing. However, since the node pairwise metrics are dependent only on the network structure, they can be computed upfront as part of a pre-processing step. Here, we save matrices in Hierarchical Data Format (HDF5) format, using the pytables framework [24], for efficient matrix storage. On this point, we designed PyGNA to performs efficiently both on low-memory machines, using memory mapped input output, and high-performance computing environments, by loading matrices directly into memory.

Analysis functions

The GNT and GNA analysis are implemented by the StatisticalTest, and the StatisticalComparison classes, respectively. It is important to note that PyGNA can be easily extended to use different test statistics by defining new Python functions; on this point, in our online documentation, we provide a complete example on how to build GNT tests based on closeness centrality of the nodes.

A bottleneck of our network analysis framework is the bootstrap procedure used to obtain a null distribution for hypothesis testing. However, the resampling procedure is a seamlessly parallelizable process, since each randomly sampled set of nodes is independent from the others; thus, we implemented a parallel sampler using the multiprocessing Python library, allowing the user to set the number of cores to use. If only one core is requested, the multiprocessing architecture is not set-up, sparing the overhead incurred by setting up a scheduler for running only one thread (see Additional file 1). It is important to note that, currently, Python 3.8 is required in order to process large matrices on multi-core CPUs.

Visualization functions

PyGNA has been developed to generate high quality figures for each analysis and to export networks and genesets in standard formats compatible with graph visualization software, such as Cytoscape [25]. The visualization functions are implemented as part of the PygnaFigure class, which comes with sensible default parameters to maximize figures readability.

There are four main types of figures currently implemented in PyGNA, namely bar plots, point plots, heatmaps and volcano plots, to visualize to GNT and GNA results.

Barplots are used to plot the GNT results for a single statistic. For each geneset a red bar represents the observed statistic, whereas a blue one represents the average of the empirical null distribution. To denote significance of each test we annotate the plot with stars, according to the $-log_{10}(p\text{-}value)$. An example is presented in Fig. 4c, as part of our results.

Conversely, a dot plot can be used to summarize multiple tests for the same geneset. In order to show all the results in the same figure, the observed values are transformed in absolute normalized z-scores, such that all significant tests have z-score > 0 and are marked with a red dot. An example is discussed in Fig. 4a.

GNA results can instead be visualised on heatmaps, with the color gradients used to report the strength of association between two genesets. When an all-vs-all test is conducted, as in Fig. 5, a lower triangular matrix is shown, with stars denoting significance. If, instead, a M-vs-N test was conducted, a complete heatmap would be included in the plot.

Alternatively, volcano plots can be used to visualize one-vs-many GNA results, for testing a geneset against a large number of datasets (e.g. gene ontologies). The plot shows the normalized z-score on the x-axis and the $-log_{10}$ of the p-value adjusted to control the False Discovery Rate (FDR) on the y-axis. Significant results are shown with red crosses, whereas not significant associations are represented by blue dots. We also annotate the plot with the top 5 scoring terms. An example of this plot is presented in the Additional file 1.

We provide more detailed information and tutorials in our online documentation.

Network properties

On top of the statistical testing framework, we provide functions for the basic characterization of the network and geneset. General information, such as number of nodes and edges, average degree, connected components of the graph can all be retrieved from command line and saved in textual formats or shown in a GraphML file.

Network simulation functions

PyGNA provides a comprehensive simulation framework to generate networks with different structures and properties for benchmarking purposes, as described in "Benchmarking geneset network tests" section. Moreover, since we allow the user to implement further statistical tests, we provide a full pipeline to generate a benchmark dataset to compare the results with those available in this paper.

Model descriptions and implementation details are also available in our online documentation.

Command line interface and workflow system integration

PyGNA implements a standard Unix-like command line interface with robust default options set for all functionalities. Using a CLI interface facilitates integration with workflow analysis systems, such as SNAKEMAKE [16]. We have developed SNAKEMAKE pipelines to perform network analysis, available at https://github.com/stracquadaniolab/workflow-pygna, which can be readily integrated into existing workflows.

Results

We designed PyGNA as a tool to streamline network analysis of biological data. Here we perform an extensive analysis of the performances of the GNT and GNA tests implemented in PyGNA and then, we present a common use case regarding the analysis of cancer RNA sequencing (RNA-seq) experiments, providing basic guidelines to interpret PyGNA results.

Network simulations and algorithm benchmarking

GNT benchmarking with SBM and HDN

We used the SBM and HDN network models to assess the performance of the GNT tests implemented in PyGNA.

To do that, we first generated networks using the SBM model using the parameters reported in Table 1 (GNT-SBM). Given the large number of parameters, we restricted our analyses to networks generated using $k = 7$ blocks. For each network, we set $\lfloor k/2 \rfloor$ blocks with connection probability αp_0 to simulate genesets with a network effect, which we denoted as positive genesets, whereas the remaining $k - \lfloor k/2 \rfloor$ were denoted as

Table 1 GNT and GNA benchmark parameters

Parameter	Description	GNT-SBM	GNT-HDN	GNA-SBM
n	Number of nodes	1000	1000	1000
k	Number of blocks	7	–	9
p_0	Baseline connection probability	0.01, 0.02, 0.05	0.006, 0, 02	0.01, 0.02
m	Size of the geneset	20, 50, 100	9, . . . , 200	50, 80
α	Within block connection probability scaling	2, 3, 5, 10	–	2, 5
β	Between block connection probability scaling	0	–	0, 10
p_{hd}	HDN connection probability	–	0.5, 0.2, 0.1, 0.08, 0.05, 0.01	–

For each model, we report the name of the parameter, a short description and its setting. A dash is reported when the parameter is not used

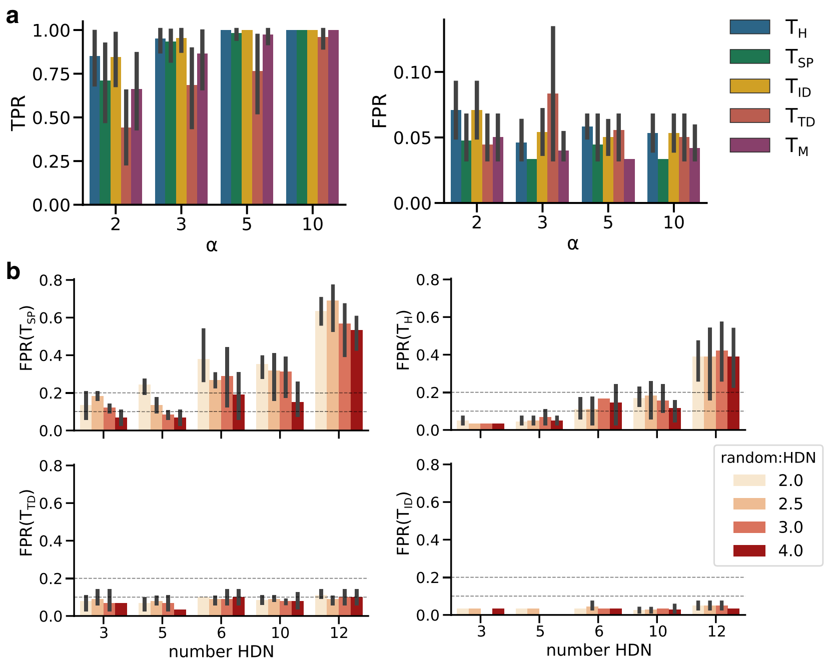


Fig. 2 GNT benchmarking. **a** Performance of all GNT tests on the SBM networks. We show True Positive Rate (TPR) and False Positive Rate (FPR) (y-axis) of each GNT test (colors) for different values of α (x-axis). As expected, as the value of α increases, all tests improve their detection performance, with T_H and T_{ID} having consistently TPR > 0.75. Conversely, for FPR we do not see a strong effect as α increases, with most tests having FPR ∼ 5%. **b** Extended geneset high degree nodes (HDNs) networks used to quantify FPR. Genesets have been selected with increasing number of HDNs (x-axis) and random nodes to HDNs ratios (colors); for each analysis, we report the False Positive Rate (FPR). As the ratio between random and HDNs increases (ρ), we notice that T_{SP} has better performances. Interestingly, T_{ID} is the only one with FPR < 5% in all conditions

negative genesets. For each possible parameter setting, we generated 10 networks and corresponding genesets for a total of 5400 positive and 5400 negative genesets.

We found that T_{ID}, T_H and T_{SP} are the statistics with the best overall performances (Fig. 2a), with TPR > 70% for all instances, whereas T_M and T_{TD} were able to detect a network effects only for highly connected genesets. In general, we found that all tests are robust to false positives ($FPR < 10\%$ for all tests), with T_{SP} being the most conservative.

We then used the HDN model to estimate the FPR of the GNT tests with respect to networks with hubs. Here we generated networks using the parameters reported in Table 1 (GNT-HDN) and generated 10 networks for each parameter setting. For each network, we then created extended genesets with by varying $\pi_{hd} = 0.1, 0.2, 0.5$ and $\rho = 2, 2.5, 3, 4$. For each combination of π_{hd} and ρ we generated 3 random genesets, for a total of 30 datasets for each combination of network and geneset parameters. It is important to note that for increasing p_0, p_{hd} and π values, the extended genesets begin to form connected clusters; these cannot be considered false positives, albeit being generated at random. Thus, for each geneset, we first computed the size of the largest connected component (LCC), and discarded those genesets with more than 75% of the genes belonging to the LCC.

Here we found that our tests have a low FPR ($< 10\%$) regardless of geneset composition and network structure. Interestingly, while T_{SP} was the most robust on SBM networks, it is the most sensitive to HDN in the networks, with FPR as high as 20% even for genesets with only 3 HDNs (Fig. 2b). In this case T_{ID} is the most robust test (FPR$<$ 10%), while T_H has FPR$>$ 0.2 when the number of HDN increases.

Taken together, the T_{ID} statistic is the one achieving the best performances and it is faster to compute respect to the other best performer, T_H, which requires the computation of a random walk matrix. Nonetheless, for exploratory analyses, we recommend using the T_H test, which is confirmed to be well powered to detect network effects and has a low FPR, and might less sensitive to missing links. We would also point out that, since PyGNA provides implementations of the GNT analysis under different models of interaction, ensemble analyses could be useful in practice to increase the power of detecting network effects.

GNA benchmarking with SBM

We tested also the performance of GNA tests by generating networks and genesets as outlined in "SBM for GNA benchmarking" section and using the parameters reported in Table 1. For each network, we set two groups of blocks, $k^+ = 4$ and $k^- = 4$, both of size m, along with another one including the remaining $N - k \times m$ nodes. We then set $M_{ij} = \gamma p_0$, for $i = 1, \ldots, 7$ and $j = i + 1$. For each pair of blocks, we generated genesets with a varying mixture of nodes $\pi = \{0.04, 0.06, 0.1, 0.12\}$; with these genesets, we can test associations between highly connected and partially overlapping genesets. For each network and geneset parameter, we generated 10 runs, for a total of 2640 datasets. For both U_H and U_{SP}, we then assessed the TPR, as the ratio of significant tests between genesets with $\beta = 10$, and the FPR, as the ratio of significant tests between genesets with $\beta = 0$.

We found that U_H has higher TPR than U_{SP}, regardless of network structure and geneset composition (see Fig. 3). However, it is more prone to false discoveries when the number of overlapping nodes increases. In particular, when two genesets do not have high inter-connectivity, but share more than 5 out of 50 nodes the test is always significant. Importantly, all tests between non overlapping genesets are not significant.

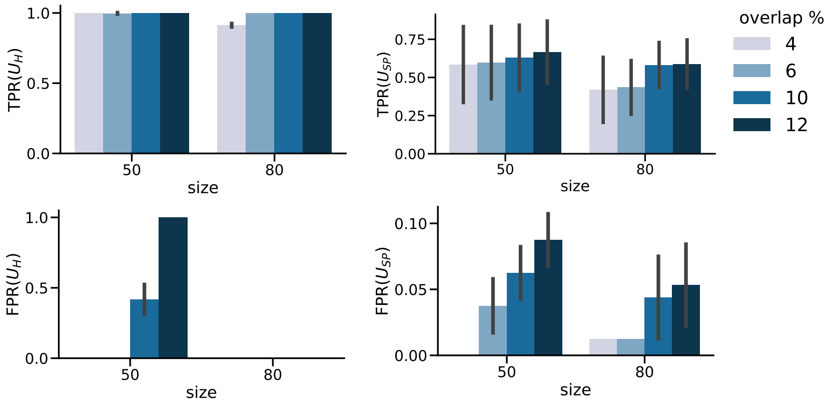


Fig. 3 GNA benchmarking. **a** Performance of all GNA tests on SBM benchmark data. On the left column, we report the True Positive Rate (TPR) and False Positive Rate (FPR) for U_H, while on the right column we report the same metrics for U_{SP}. On the x-axis, we show different geneset sizes, while we denote the overlap between the tested genesets with colors. For example, for size 50 and 4% of overlap the two geneset share 2 nodes. We notice that U_H has TPR $>$ 0.95, while U_{SP} is consistently below 0.75. Moreover, the FPR analysis confirms better performance for U_H, albeit it is skewed by many overlapping nodes. On this point, when two genesets share 6 or more nodes out of 50, U_H always considers them as positives

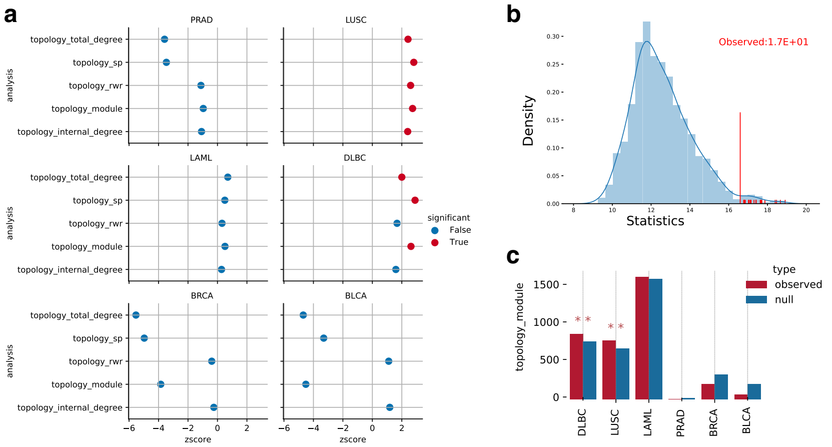


Fig. 4 GNT analysis of TCGA RNA sequencing experiments. **a** Summary of the GNT results on the TCGA datasets. For each geneset analysed, a summary of all test results is reported. In order to make results comparable, observed test statistics are transformed in normalised z-scores. All results are in a scatter plot, where significant tests are marked with a red dot. We can notice that only the TCGA Lung Squamous Cell Carcinoma geneset is significant for all topology tests. **b** Null empirical distribution (blue) and observed value (red bar) for a significant rwr test on the TCGA Lung Squamous Cell Carcinoma geneset. **c** Barplot of the GNT module analysis on all TCGA datasets. For each geneset, we report both the observed statistic and the empirical null distribution average. Stars are used to identify significance of the test. Here, DLBC (p-value:6.99×10^{-3}) and LUSC (p-value: 8.99×10^{-3}) are significant, while the other terms are not

Taken together, our results suggest that for U_H is a well powered test for exploratory analyses, whereas U_{SP} might be more appropriate for verifying known associations.

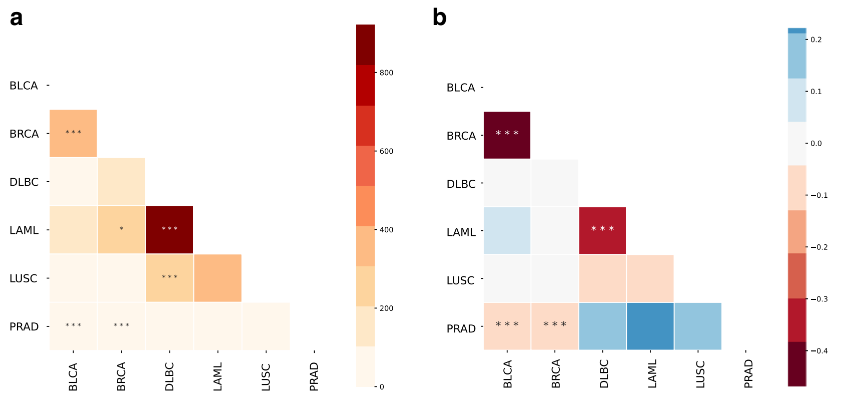


Fig. 5 GNA analysis of TCGA RNA sequencing experiments. **a** Heatmap of the observed values of the GNA test under a RWR interaction model, U_H, where darker colors denotes larger observed U_H values and stars denote statistical significance. **b** Heatmap of the observed values of the GNA test under a shortest path interaction model, U_{SP}. A divergent palette marks distant datasets with blue hues, and close ones with red hues, ($U_{SP} < 0$)

Use case: network analysis of RNA sequencing experiments

RNA-seq experiments aim at finding genes that are up or down regulated between two or more conditions. As a use case, we analyzed RNA sequencing data generated by The Cancer Genome Atlas (TCGA) project [15] for 6 different types of cancer (see Additional file 1). Specifically, we selected 4 epithelial tumors, including 2 from urogenital tissues (BLCA and PRAD), 1 from breast (BRCA) and 1 from lung (LUSC), and 2 from liquid cancers (LAML and DLBC).

Here we are interested in finding whether differentially expressed genes in each cancer show a network effect, and whether they are similar to any other cancer analysed. It is possible to address these questions using the GNT and GNA tests implemented in PyGNA.

To do that, we retrieved TCGA data and performed differential expression analysis (DEA) using the TCGABiolinks package [26]. Here we found that there are no control samples in TCGA for LUSC, LAML, and DLBC; in this case, we instead used gene expression data from the Genotype-Tissue EXpression (GTEX) project [27], as control, and the TCGA tumor data processed by the Recount2 project [28], in order to avoid biases introduced by different RNA quantification pipelines (see Additional file 1). Taken together, we retrieved 6 datasets providing mRNA abundance for ≈ 15000 genes for each tumor and performed differential expression analysis. For each dataset, we consider significant all genes with $FDR < 0.01$ and $|logFC| > 3$ (see Additional file 1).

We then used PyGNA to perform GNT analysis and GNA analysis between all cancer datasets, using the BioGRID interaction network [29], a publicly available repository of protein interactions defining a human protein interaction network of 17331 nodes and 283991 edges. For each test, PyGNA returns the results as a CSV file, which includes descriptive statistics and the parameters of the null distribution used for hypothesis testing. Our workflows are summarized in Fig. 1 and Snakemake pipelines are available at: https://github.com/stracquadaniolab/workflow-pygna.

We then used the PyGNA plotting tool (PAINT-SUMMARY-GNT) to visualize a summary of the GNT results for all datasets (Fig. 4a), where we report the test statistic as a z-score, to make them comparable across different tests. Interestingly, only differentially expressed genes in lung and lymphoid cancers show a significant network effect, albeit this is detected by all tests for lung cancer and only by T_H and T_{ID} lymphoid neoplasm. Interestingly, we did not observe any network effect for the other cancers; this could be explained by the fact these cancers might be controlled not by one highly connected network, but by multiple distinct ones.

We then used PyGNA diagnostic plot generated by the GNT analysis to visualize the effect size and the null distribution of the test statistic for one of our significant datasets; in Fig. 4b, the plot shows the observed value of the T_H statistic for lung cancer (vertical red line) being located in the upper-tail of the null distribution (blue area), suggesting that a network effect has been detected.

We again used the PyGNA plotting tool (PAINT-DATASETS-STATS) to present a summary of the GNT T_M results for all datasets (Fig. 4c). For each geneset, we report both the observed statistic and the empirical null distribution average. Stars are used to identify significance of the test. Here, DLBC (p-value:0.00699) and LUSC (p-value: 0.00899) are the only cancers with a statistically size of the induced module.

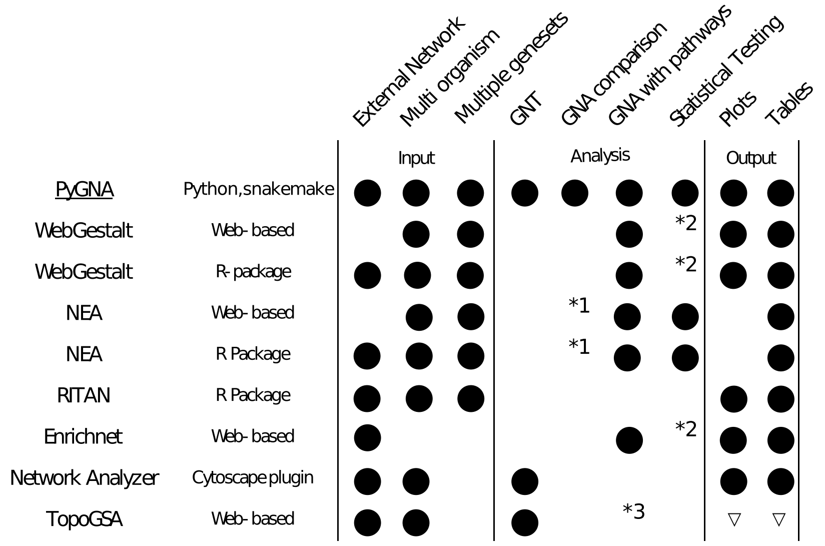
We then performed a GNA analysis between all differentially expressed genesets using the command (PAINT-COMPARISON-MATRIX) in PyGNA. While most of them does not seem to show a consistent network effect, we can use the GNA to test whether each set of differentially expressed genes are more connected with each other than expected by chance. Using either U_H and U_{SP} tests, we found a significant association between breast, bladder, and prostate carcinomas, and between leukemia and lymphoid neoplasms (Fig. 5); this is clearly shown through darker gradients for strongly associated genesets, and by the star notation to report statistical significance. This result is consistent with other gene expression analyses, which have shown that anatomically related cancers or with similar histopathology share similar changes in gene expression [30]. Interestingly, we found a significant association between lung and lymphoid neoplasms; this might be explained by the fact that lungs contain a vast lymphatic network, which might also be dysregulated in lung tumors.

Taken together, we have shown how PyGNA enables network analysis of RNA sequencing datasets and provide useful biological insights. The availability of informative diagnostic and descriptive plots provides a simple entry point for downstream expert analyses.

Discussion

The availability of biological interaction data has propelled the development of a plethora of network analysis methods, with the promise of linking single genes and protein information into networks to understand biological processes.

We surveyed publicly available, documented and actively maintained network analysis tools and found that, currently, PYGNA is the only available framework for comprehensive statistical network analysis under different interaction models (see Fig. 6). Currently, most software is available as web applications rather than stand-alone tools, usually performing only quantitative analyses with no statistical testing. This brings



		Input			Analysis				Output	
		External Network	Multi organism	Multiple genesets	GNT	GNA comparison	GNA with pathways	Statistical Testing	Plots	Tables
PyGNA	Python,snakemake	●	●	●	●	●	●	●	●	●
WebGestalt	Web- based		●	●			●	*2	●	●
WebGestalt	R- package	●	●	●			●	*2	●	●
NEA	Web- based		●	●		*1	●	●		●
NEA	R Package	●	●	●		*1	●	●		●
RITAN	R Package	●	●	●					●	●
Enrichnet	Web- based	●					●	*2	●	●
Network Analyzer	Cytoscape plugin	●	●		●				●	●
TopoGSA	Web- based	●	●		●		*3		▽	▽

1) Not explicitly defined, but could be applied by using proprietary pathway sets
2) Statistical testing is done with ORA
3) No GNA, it allows to find similarities between the topological properties of pathways and the studied geneset
▽) only interactive, no downloadable

Fig. 6 State-of-the-art tools for geneset network analysis. Comparison between publicly available, documented and actively maintained network analysis tools. For each tool, we reviewed the type of networks and genesets that can be given as input (e.g. multi-organism, external/custom defined), and whether a tool can generate tables and figures. The majority of tools provides only one type of network analysis, either GNT or GNA, with few of them providing association tests between multiple user defined genesets. We also noted that, for many tools, there are no statistical testing procedures. Conversely, PyGNA enables comprehensive statistical network analysis under different interaction models, testing both single geneset topology and multiple genesets association. Moreover, PyGNA takes input user-defined networks, regardless of their type and organism, and provides results in comma separated value (CSV) files and PDF figures

major limitations both for data interpretation and downstream integration into existing data analysis pipelines (e.g. RNA-seq and variant calling workflows); PyGNA directly addresses these problems, by implementing statistical analysis tools into a modular software package.

We further reviewed available tools by classifying their functionalities either as GNT or GNA, whether they perform statistical analysis and whether they provide a command line interface (CLI). We found two tools performing GNT on user defined genesets: TopoGSA [31] and NetworkAnalyzer [32]. TopoGSA is a web application implementing network topology geneset analysis. It evaluates topological properties of the subnetwork induced by an input geneset, such as average shortest path length, node degree and clustering coefficient. An empirical p-value is obtained through permutations, but the limited number of samples generated do not ensure a stable distribution for hypothesis testing. TopoGSA checks also for similarities with known pathways just by comparing network properties, but no statistical testing is performed, which ultimately limits its utility for interpreting the data. The application presents results in interactive tables and plots, and facilitate access to pre-computed networks of several organisms, along with the option to import user-defined networks. NetworkAnalyzer is a Cytoscape plugin, which estimates topology features of the subnetwork induced by a geneset, including centrality measures, average shortest path, node degree distribution. Differently from TopoGSA, it only provides descriptive statistics but no statistical analysis can be performed.

PyGNA instead provides robust topological statistical testing under different interaction models, which enables in depth analysis of the data, and represents a better solution for topology analysis.

Interestingly, we found GNA analysis to be a more popular application, in particular to study association with known pathways. The vast majority of tools perform association analysis using either over-representation analysis (ORA), which is usually a variant of Fisher's exact test, or geneset enrichment analysis (GSEA, [33]). However, none of them explicitly allows association analysis between multiple user-defined genesets. There are three available tools commonly used for GNA analysis with known pathways: Webgestalt [34], network enrichment analysis (NEA, [35]), and Enrichnet [10].

Webgestalt is a comprehensive suite for geneset analysis, which implements conventional ORA and GSEA analysis, and performs association testing as network topology association (NTA) test using Gene2Net (http://www.gene2net.org/). First, a subnetwork is built from the input geneset by adding relevant neighbours using a random walk model, as implemented in NetWalker [36]. Then, the application performs ORA between the genes in the inferred subnetwork and a pathway databases. We found Webgestalt to be the most comprehensive tool for GNA, as it includes multiorganism and multiplatform support, interactive plots and downloadable results. NEA performs GNA by computing an enrichment score between an input geneset and a pathway, as a function of the number of edges shared between the two. The statistical significance of the score is assessed by randomly permuting the edges in the network; recently, a binomial test has been implemented to reduce the running time. NEA has been implemented both as an R package, NEArender [37], and as a web application, EviNet [38], which provides access to multiple network and pathway repositories (e.g. GO, KEGG, Biocarta). Enrichnet performs GNA analysis between a user-defined geneset and a predefined list of biological pathways. The application uses RWR to compute interaction probabilities between the input geneset and each pathway. The interaction probabilities are transformed into a score, X_d; intuitively, X_d is a measure of how close the geneset is to the pathway compared to all the others. As the X_d score does not allow a direct statistical testing, it is combined with Fisher-test FDR corrected p-values from an ORA, to find the threshold for significance. Enrichnet is a useful tool for direct comparison of ORA and a GNA, albeit it was last updated in 2012 and new pathways cannot be imported.

Since PyGNA provides also API for statistical network analysis, we also reviewed Ritan [39], an R package that provides functions for genesets and networks analysis in R. Ritan provides ORA testing between a geneset and pathways, provides functions to export networks for Cytoscape and iGraph, but it does not include any GNT or GNA off-the-shelf functionality. Finally, we would like to point out that the vast majority of GNA tests, are designed and optimized to perform association tests with specific datasets (e.g. KEGG pathways or Gene Ontology), rather than addressing the more general problem of network association; this poses substantial technical challenges for any rigorous benchmarking experiment based on synthetic networks.

Taken together, our software is the only available solution to easily investigate network properties under different interaction models and perform statistical testing. We recognize that web applications are easier to interact with, fast to use for small scale and targeted analyses, since they do not require any setup and integrate many network and pathway genesets. However, we have designed PyGNA with flexibility and scalability in mind; we provide both command line interface and open APIs to extend GNT and GNA analysis using different topology measures. Moreover, the support for multi-core processing and easy integration with SNAKEMAKE allows to run PyGNA on multiple datasets and experiments at a glance

Conclusions

The availability of gene and protein interaction data provide unique opportunities to understand the cellular wiring underpinning most common complex phenotypes. However, integrating network and gene-level information has been challenging. Geneset network analysis provides a statistical framework to test the presence of interactions between genes associated with a phenotype, thus providing a useful tool for downstream analysis of high-throughput data. However, there are only few tools for statistical geneset network analysis, and usually are limited to specific interaction models, lack statistical testing methods or are only accessible through web applications.

Here we present a modular Python package, called Python Geneset Network Analysis (PyGNA), to perform statistical geneset network analysis under different interaction models. As networks analysis results are sensitive to the underlying gene and protein interaction model, it is important to perform these analyses using different models to gain confidence on the observed network effects. Different from existing applications, we designed PyGNA to be easily integrated into workflow systems and rapidly provide a comprehensive network characterization of input genesets. Our software takes advantage of multi-core architectures and can work both on desktop and high-performance computing environments, thus lowering the computational requirements to perform network analysis. Our software is available on GitHub (http://github.com/stracquadaniolab/pygna) and can be easily installed from PyPi, Anaconda and as a Docker container.

We have shown how PyGNA can be used as part of biological data analysis pipelines, in particular as downstream analysis tool for differential expression experiments, exploratory geneset analyses, and as a network simulation framework. It is also worth mentioning that, while the package development has been motivated by the need for an integrated tool for biological data analysis, ranging from RNAseq experiments to evolutionary genomics [40] PyGNA is agnostic to input data types and could easily be adopted to analyse non-biological networks, including social and communication networks, where the information can be summarized into sets of nodes (e.g. users of a Facebook group).

PyGNA is not only a stand-alone application, but also a Python library that can be easily integrated into other software; thus, we envision our framework as an open-source platform to develop network statistical tests.

Availability and requirements

Project name: PyGNA
Project home page: https://github.com/stracquadaniolab/pygna
Operating system(s): Platform independent
Programming language: Python
Other requirements: pandas, numpy, scipy, matplotlib, pyyaml, tables, seaborn, palettable, networkx, statsmodels, argh, mygene (Python 3.8 is required to use large matrix analysis on multiple processors)
License: MIT license
Any restrictions to use by non-academics: Not applicable

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03801-1.

Additional file 1. contains all supplementary materials and figures referenced in the main manuscript. Section 1 1.1 describes more in depth th paralle sampling performance, Section 1 1.2 describes the stability of empirical null distributions, Section 1 1.3 describes the geneset network association bootstrapprocedures, Section 1 1.4 describes materials and preprocessing stepts for the TCGA data analysis. Section 2 is instead dedicated to the supplementary figures that are referenced in the main text.

Abbreviations

PyGNA: Python Geneset Network Analysis; GNT: Geneset network topology; GNA: Geneset network association; TCGA : The cancer genome atlas; CLI: Command line interface; LCC: Largest connected component; RWR: Random walk with restart; SBM: Stochastic block model; TPR: True positive rate; FPR: False positive rate; HDN: High degree nodes; OOP: Object oriented paradigm; BLCA: Bladder urothelial carcinoma; BRCA: Breast invasive carcinoma; PRAD: Prostate adenocarcinoma; LUSC: Lung squamous cell carcinoma; LAML: Acute myeloid leukemia; DLBC: Lymphoid neoplasm diffuse large B-cell lymphoma; DEG: Differentially expressed gene; GTEX: Genotype-tissue expression project; ORA: Over-representation analysis.

Acknowledgements

We would like to thank Prof. Francesco Pezzella at University of Oxford for useful discussions on the tumor transcriptome analysis.

Authors' contributions

VF and GS conceived the study. VF and FC wrote PyGNA and performed experiments under GS supervision. VF, FC and GS wrote the manuscript. All authors have read and approved the manuscript.

Funding

This work has been supported by the Wellcome Trust Seed Award in Science (207769/A/17/Z) to G.S.

Availability of data and materials

Data have been deposited on Zenodo and are freely accessible at: https://doi.org/10.5281/zenodo.3922015.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not Applicable

Competing interests

The authors declare that they have no competing interests.

Received: 11 January 2020 Accepted: 6 October 2020
Published online:

References

1. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20:257.
2. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, et al.: A reference map of the human protein interactome. bioRxiv, 605451 (2019)

3. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. Biogrid: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(suppl–1):535–9.
4. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012;28:451–7.
5. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2018;47(D1):419–26.
6. Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012;28(13):1805–6.
7. Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74:47–97.
8. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(suppl–2):214–20.
9. Chen J, Bardes EE, Aronow BJ, Jegga AG. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37(suppl–2):305–11.
10. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. Enrichnet: network-based gene set enrichment analysis. Bioinformatics. 2012;28(18):451–7.
11. Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106–14.
12. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). Bioinformatics. 2013;29(21):2757–64.
13. Cho A, Shim JE, Kim E, Supek F, Lehner B, Lee I. Muffinn: cancer gene discovery via network analysis of somatic mutation data. Genome Biol. 2016;17(1):129.
14. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, Lin J, Hescott B, Hu X, Mercer J, et al. Assessment of network module identification across complex diseases. Nat Methods. 2019;16(9):843–52.
15. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemp Oncol. 2015;19(1A):68.
16. Köster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2.
17. Newman M. Networks. Oxford: Oxford University Press; 2018.
18. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab 1999.
19. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
20. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18(9):551–62.
21. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):841.
22. Ghasemian A, Hosseinmardi H, Clauset A. Evaluating overfit and underfit in models of network community structure. IEEE Transactions on Knowledge and Data Engineering. 2019;.
23. Love M, Anders S, Huber W. Differential analysis of count data-the deseq2 package. Genome Biol. 2014;15(550):10–1186.
24. Team PD. PyTables: Hierarchical Datasets in Python (2002). http://www.pytables.org/
25. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2010;27(3):431–2.
26. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic Acids Res. 2015;44(8):71.
27. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The genotype-tissue expression (gtex) project. Nat Genet. 2013;45(6):580–5.
28. Collado-Torres L, Nellore A, Jaffe AE. recount workflow: accessing over 70,000 human rna-seq samples with bioconductor. F1000Research 6 (2017)
29. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, et al. The biogrid interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):369–79.
30. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291–304.
31. Glaab E, Baudot A, Krasnogor N, Valencia A. Topogsa: network topological gene set analysis. Bioinformatics. 2010;26(9):1271–2.
32. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc. 2012;7(4):670–85.
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Genetics. 2005;102(43):15545–50.
34. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47(W1):199–205.
35. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J, Pawitan Y. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinform. 2012;13(1):226.
36. Zhang B, Shi Z, Duncan DT, Prodduturi N, Marnett LJ, Liebler DC. Relating protein adduction to gene expression changes: a systems approach. Mol BioSyst. 2011;7(7):2118–27.
37. Jeggari A, Alexeyenko A. NEArender: an R package for functional interpretation of 'omics' data via network enrichment analysis. BMC Bioinform. 2017;18:118.

38. Jeggari A, Alekseenko Z, Petrov I, Dias JM, Ericson J, Alexeyenko A. EviNet: a web platform for network enrichment analysis with flexible definition of gene sets. Nucleic Acids Res. 2018;46(W1):163–70.
39. Zimmermann MT, Kabat B, Grill DE, Kennedy RB, Poland GA. Ritan: rapid integration of term annotation and network resources. PeerJ. 2019;7:6994.
40. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol. 2008;25(4):762–77.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.4 Conclusions

PyGNA is a tool that allows to topologically characterize genesets routinely obtained from the analysis of high-throughput experiments. We implemented robust and frequently-used statistics that shed light on the topological properties and localisation of a set of genes. Importantly, the detailed implementation of simulated networks allows generating extensive benchmarks that we used to assess the power and robustness of the implemented properties. Novel results can be readily tested by integrating the provided pipeline steps in existing workflows. Alternatively, PyGNA is a modular Python package that can be included in personalised scripts and other tools or extended with user-defined statistics.

Nevertheless, we have to address some of the limitations of PyGNA and topological testing in general. Taken together, PyGNA is a computational tool applicable downstream to current NGS analysis methods. We provide a full framework to benchmark the statistical tests and to characterise the topology of a geneset. However, we do recognise that the results require some careful interpretation which is less straightforward than a pathway analysis. Geneset network topology tests (GNT) such as module, internal degree, and diffusion, aim at identifying whether a set of genes could be a putative ‘pathway’, that is a set of genes associated with a phenotype and strongly interacting with each other. Unfortunately, extensive analysis of available pathways and disease-associated geneset has shown that they often are fragmented and sparsely mapped onto the PPI network, lacking the properties expected by topological testing [Agrawal, Zitnik, and Leskovec 2018].

Moreover, our rigorous benchmarking of the statistics shows that shortest path and module measures, previously used to show network effects and comorbidity patterns [Menche et al. 2015], are prone to false positives. We also have evidence that testing the same geneset onto two different PPI networks, that are expected to be capturing similar and common properties of the interactome, lead to different results. Considering the low sensitivity of PPI networks and the challenge of aggregating data from the literature, we reason that methods that only rely on the observed interactions might fail to fully capture the properties of a geneset. Methods that try to predict interactions or that filter spurious links [Silverbush and Sharan 2019] will possibly improve the reliability, and interpretability, of topological testing.

Eventually, the biggest limitation of the approaches presented in this chapter is the mapping of gene scores onto the network. Embedding multi-modal data onto the network structure is not trivial, and all methods discussed here are condensing the results of entire high-throughput experiments onto a score the represents each gene status. While this is a necessary step for many graph inference methods, collapsing multiple measures might result in information loss. Multi-omics studies are becoming increasingly important and frequent, as they directly capture the complexity of a cell. Henceforth, in the next chapter, we present and discuss the methods for inference on multi-modal graph data.

5 Data integration on networks

In the previous chapter, we presented the state-of-the-art for the analysis of graph topology in biological networks, with particular focus on the computational tools available for interactome-level data analysis. We have shown that topology alone is helpful to describe the connectivity of a subnetwork, but detecting the full complexity of the structural organisation is much harder. Moreover, we briefly discussed the limits of using single scores onto the network structure and the difficulty of providing easily interpretable results.

Twenty years after the first genome was sequenced, a wealth of high-throughput technologies can characterise samples at the genomics, transcriptomics, and proteomics level [Reuter, Spacek, and Snyder 2015; Aslam et al. 2017; Lowe et al. 2017]. In practice, we are often dealing with multi-dimensional datasets that capture various properties of the same samples. Multi-omics analysis methods [Bersanelli et al. 2016; Reel et al. 2021] are increasingly being used for cancer research; for instance, they have been applied to find sources of heterogeneity in Chronic Lymphocytic Leukaemia [Argelaguet et al. 2018], to distinguish between tumor subtypes in liver [Chaudhary et al. 2018] and oligodendroglial tumors [Kamoun et al. 2016].

This amount and detail of data and the increased availability of PPI networks suggest that methods for the integration of the interactome with multi-modal data will produce novel insights into tumor biology. For instance, using previous knowledge of somatic driver mutations to guide diffusion on the network results in better predictions of putative drivers [Hristov, Chazelle, and Singh 2020] and methods applied to multi-omics data to detect cancer-driving modules outperform results on single-omics methods [Silverbush et al. 2019].

However, as it was discussed in the previous chapter, combining graph-structured data and multiple node attributes (or features) is not trivial. Depending on the task, either class prediction or node clustering, there are appropriate strategies to do inference on attributed graphs. First, the attributes are manipulated to obtain scores that can be directly used with graph inference methods; the features are transformed into a similarity matrix that weights the edges of the network [Kuijjer et al. 2019], or multiple features are summarised in a single node score [Leiserson et al. 2015; Silverbush et al. 2019]. Alternatively, the graph is encoded onto the features space, such that they can be directly analysed with methods for the analysis of tabular data. This is what is done with node embedding methods [Goyal and Ferrara 2018] which apply a graph-dependent transformation to the features, that are then analysed with standard machine learning methods. Both these approaches can be thought of as preprocessing steps, since the actual inference method does not directly deal with either the features, in the first case, or the network, in the latter.

Eventually, node features can be used within the inference method; they can guide a random walk [Hristov, Chazelle, and Singh 2020], can be transformed into a distance metric to guide clustering [Bothorel et al. 2015], can be used within an expectation-maximization algorithm to find subnetworks [Newman and Clauset 2016]. In these cases, the features are incorporated into the model's parameters resulting in more computationally expensive algorithms or a reduced number of possible attributes.

Graph Neural Networks (GNNs) are a class of deep learning methods that apply convolutions, or other nonlinear transformations, to graph-structured data with multiple node features [Wu et al. 2019; Zhou et al. 2020]. GNNs explicitly use the graph structure within their layers and optimization. Features are transformed by using connectivity information, possibly at all hidden layers of the Neural Network (NN); they apply multiple, sequential, embedding operations both capturing short and long-range interactions within the network. Also, both structure and node features can be used in the optimisation process by specifying appropriate losses. Moreover, GNNs are built on top of the well-established field of deep learning and borrow most of its optimization procedures and, in practice, their community-maintained computational frameworks. The key advance of GNN has been introducing differentiable transformations that correspond to meaningful operations in the non-Euclidean graph space. From both a mathematical and a practical point of view, NNs are easier to optimise than graph inference methods and provide a flexible framework for testing and tuning a model.

5.1 Graph Neural Networks

Neural Networks are a powerful resource that has become widely applied to many different fields; for computer vision, natural language processing, image classification, deep learning is by now ubiquitous, and recently it has also been increasingly applied to biological data. Naturally, the first and more successful implementations of NNs in biology have been for image processing [McKinney et al. 2020; Haibe-Kains et al. 2020] and prediction of sequence function and binding [Zhou and Troyanskaya 2015; Avsec et al. 2021; Yuan et al. 2019].

These methods are applied to one- or bi-dimensional data, that allow to efficiently define and apply matricial operations. Graphs, however, have much more complex structures and cannot be directly projected onto a Euclidean space. The first deep learning methods for graph data [Grover and Leskovec 2016; Perozzi, Al-Rfou, and Skiena 2014] borrowed the strategies used for text documents to embed the node features into a Euclidean space, before applying conventional NN layers. However, recently, NNs have been extended to explicitly handle graphs, resulting in a deluge of novel GNN methods [Wu et al. 2019; Zhou et al. 2020]. While the taxonomy of the field is still being updated, we can anticipate that we are interested in the methods of Convolutional GNNs that generalise the convolution operation onto the graph. In the remainder of this section, we will introduce the basic intuition behind convolutional GNNs.

5.1.1 Convolutional Neural Networks

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, with $\mathcal{V}$ being the set of n nodes and $\mathcal{E}$ being the set of m edges between nodes in $\mathcal{G}$. Edges can be represented by an adjacency matrix $A \in \mathbb{R}^{n \times n}$, such that $A_{uv} = 1$ iff node u is connected to node v in $\mathcal{G}$, and 0 otherwise. The graph $\mathcal{G}$ can be attributed, that is each node v correspond to a vector $\mathbf{x} \in \mathbb{R}^f$ of properties. Contextually, each node could belong to a group, such that each node corresponds to an assignment vector $\mathbf{y} \in \{0, 1\}^c$ that specifies to which of the c classes the node belongs, s.t. $\mathbf{y}_j = 1$ iff the node belongs to the j-th class.

Classical inference tasks would predict nodes' classes, knowing only a few of them (semi-supervised learning) or all of them (supervised learning), or cluster the nodes in unknown groups depending on their features and connections (unsupervised learning). The idea behind the GNNs is that considering the features of a node and those of its neighbours enhances the performance prediction. Hence, GNNs apply a filter to the graph that describes how to use the attributes in the neighborhood to update those of each node. While this is easily managed into the Euclidean space through convolutions, graph nodes can each have a variable number of neighbours, and distance measures are not readily available in the non-Euclidean space. GNNs can apply such convolution-like, or message-passing tasks [Scarselli et al. 2009], defining aggregation rules that are independent of the size of the neighborhood.

First, we introduce the message-passing function in the spatial domain. With $\mathbf{x}_i^{(k-1)} \in \mathbb{R}^f$ denoting the features of node i in layer $(k-1)$, the simplest message passing graph neural networks can be described as [Hamilton, Ying, and Leskovec 2017]

$$\mathbf{x}_i^{(k)} = W^{(k)}\left(\mathbf{x}_i^{(k-1)}, \text{AGG}_{j\in\mathcal{N}(i)}\left(\mathbf{x}_j^{(k-1)}\right)\right), \tag{5.1}$$

where AGG denotes a differentiable, permutation invariant function, e.g., sum, mean or max, and W are the weights of the Multi Layer Perceptron (MLP). The gist of such layer is to represent the properties of the node at the k-th layer of the NN as the combination of the properties of the node at the $(k-1)$-th layer and the aggregated features of the neighbouring nodes. The AGG operator is equivalent to a spatial filter in the Euclidean space that defines, for instance, the value of a one-dimensional temporal signal as the average of its value and that of the preceding and following samples.

Generalizing Eq. 5.1, we can rewrite the message-passing layer as

$$\mathbf{x}_i^{(k)} = \gamma^{(k)}\left(\mathbf{x}_i^{(k-1)}, \text{AGG}_{j\in\mathcal{N}(i)}\,\phi^{(k)}\left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}\right)\right), \tag{5.2}$$

where both γ and ϕ are differentiable functions with ϕ transforming the neighborhood features. Different architectures have been proposed by changing $\phi, \gamma,$AGG; the Graph Attention Network [Veličković et al. 2018; Zhang et al. 2018], which applies the *attention* operation, well known in NN [Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017], follows the same scheme of Eq.

5.1, but it adds a learnable operation by using an MLP in ϕ that weights the concatenated features of both node i and its neighborhood. Conversely, the isomorphism operator, GIN [Xu et al. 2019], approximates the Weisfeiler-Lehman (WL) graph isomorphism test by using the sum aggregator, an MLP for γ, and a scalar weight for $\mathbf{x_i}$.

However, convolutional filters can be also defined in the spectral domain where the convolution operation is applied as matrix multiplication. Here, the key issue is to transform the attributed graph in the spectral domain. First we need to introduce the degree diagonal matrix D s.t. $D_{i,i} = \sum_j A_{i,j}$ and the graph Laplacian for an undirected matrix $L = D - A$. Let $L = U \Lambda U^T$ be the decomposition to the eigenvalues of L, where U are its eigenvectors and Λ the eigenvalues. Given the feature vector $\mathbf{x}$, its Fourier transform in the graph space is $\mathcal{F}\{\mathbf{x}\} = U^T\mathbf{x}$. Given then a filter g_θ, with Θ Fourier transform, the convolution between the node and the filter becomes $\mathcal{F}\{g_\theta \circledast \mathbf{x}\} = \Theta U^T \mathbf{x}$.

Given the definitions above, a NN layer where a convolutional filter is applied in the graph spectral domain is defined below:

$$H^{(l)}_{(:,q)} = \sigma\left(\sum_{i=1}^{f_l - 1} U\, \Theta^{l-1}_{p,q}\, U^T\, H^{l-1}_{:,p}\right) \quad \text{with} \quad j = \{1, \dots, f_l\} \tag{5.3}$$

f_l are the feature dimensions for the l-th layer, that is the output of the NN layers, and $\Theta^{(l-1)}_{p,q}$ is a diagonal matrix that defines the learnable filter between the p-th input feature and the q-th output feature, as a transformation of the graph eigenvalues. Eventually, σ is a typical non-linearity applied to the layer. This formulation is general and allows to explicitly follow the transforms to and from the spectral domain, as well as the general definition of the filter. However, it is easy to spot some problems. First, the spectral domain is graph-dependent, hence both the eigenvectors and the filters cannot be applied to a different graph structure. Moreover, the decomposition to the eigenvalues of a graph is computationally expensive ($\mathcal{O}(n^3)$). All methods that are used in practice approximate Θ to be independent of the graph, such that perturbations to the graph would not invalidate the full NN training [Bruna et al. 2013; Henaff, Bruna, and LeCun 2015].

From the general definition above, the simplest, but also most frequently used convolutional layer, is the Graph Convolutional Network (GCN) [Kipf and Welling 2016] that approximates the filter using a Chebyshev polynomial of the first order and can be rewritten as:

$$H^{(l+1)} = \sigma\left(\tilde{A}\, H^{(l)}\, W^{(l)}\right) \quad \text{with} \quad H^0 = X \tag{5.4}$$

where $\tilde{A} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is the degree normalised adjacency matrix with $\hat{A} = A - I$, $W^{(l)}$ are the learnable parameters of the layer, and $H^{(l)}$ are the hidden features, or the input ones, X, for the first layer. We can see that from an operative point of view, this formulation is easily computed as matrix multiplications with the learnable parameters W independent of the graph. It is interesting to see that the GCN defined in the spectral domain is equivalent to Eq. 5.1 with the $\mathbf{x_i}$ values normalised by degree and the mean function as AGG. Similar to GCN, the Adaptive GCN (AGCN, [Li et al. 2018]), uses generalised Mahalanobis distance to be able to apply the filter to a non-Euclidean space and to simultaneously learn a modified adjacency matrix, that could be detecting unseen connections or remove spurious ones.

The convolutional layers we have described above, although non-exhaustive, are the fundamental models that have been revised, manipulated, and stacked in most subsequent works on GNN. Similar to what happens with the other NN models, optimal performance and best models require field-specific knowledge and hyperparameter tuning [You, Ying, and Leskovec 2020; Ghasemian et al. 2020]. GNNs are now being increasingly used for PPIs [Li and Zitnik 2021]; they have been applied to cancer subtype prediction tasks [Rhee, Seo, and

Kim 2018], cancer driver prediction [Schulte-Sasse et al. 2021], drug adverse reactions [Zitnik, Agrawal, and Leskovec 2018], and drug repurposing [Gysi et al. 2021; Ruiz, Zitnik, and Leskovec 2021]. Hence, while we cannot directly draw detailed architectures for gene function prediction, we believe that GNNs are sufficiently powerful and flexible to be applied for the integration of multi-modal cancer data.

5.2 An interpretable model for function and structure prediction

GNNs have been shown to perform well for supervised and semi-supervised node learning tasks [Kipf and Welling 2016; Veličković et al. 2018], hence they could be applied to the discovery of novel cancer drivers [Hristov, Chazelle, and Singh 2020; Reyna et al. 2020]. Nonetheless, deep learning methods are classically considered a *black box*; NNs do not provide interpretable results, often using multiple hidden layers, with thousands of parameters. As a result, trained models that achieve quasi-perfect learning performances need additional steps to retrieve and interpret the network decisions [Baldassarre and Azizpour 2019]. This issue is even more compelling for biomedical applications where it is particularly important to justify the algorithm predictions with a clear underlying model. Specifically for cancer, we have discussed the need for methods that are able not only to detect novel driver genes but also to provide the putative subnetworks implicated in the phenotype.

The Stochastic Block Model (SBM) [Holland, Blackmond, and Leinhardt 1983] is a generative model that stochastically defines the network structure through multiple communities and the probability of connection within and between them [Lee and Wilkinson 2019]. For a non-attributed graph, the model is fully specified by the SBM matrix $B \in [0,1]^{k \times k}$, and a membership matrix

$Z \in \{0, 1\}^{n \times k}$. For k communities (or blocks), each value B_{ij} is the probability of connection between two nodes of the i and j community, while each entry of the membership matrix, Z_{vk} denotes whether the node a node v belongs to the k-th block. This model can also be used to infer the communities in an observed network by finding B and Z such that the likelihood of observing the network from the model is maximum [Karrer and Newman 2011]. Compared to the simplest formulation above, the SBM has been extended with more complex, and realistic, versions of the model; for instance, models with weighted assignments of the nodes to the block [Airoldi et al. 2009] or overlapping blocks [Peixoto 2015], and models with node attributes [Stanley et al. 2019; Newman and Clauset 2016]. The SBM has also been used multiple times for genomic networks [Larremore, Clauset, and Buckee 2013; Ghasemian et al. 2020; Airoldi et al. 2009; Kavran and Clauset 2020; Padi and Quackenbush 2018], confirming that it is a suitable model to infer communities within a biomedical network setting.

In the next section we present the SBM-GNN architecture, that combines SBMs with GNNs to simultaneously infer communities within the network and to perform supervised learning on multi-modal data. The idea behind the model is that the node function depends both on its features and on the community it belongs to. The model allows us to directly detect pathways of genes that are strongly connected with each other, which is an unsupervised learning task, to predict novel drivers, supervised learning, and, finally, to link the prediction to an explainable higher-order network structure.

We benchmark the performances of our model by simulating networks and features with known community parameters. Desirably, the SBM can be used to generate networks with controlled properties; we use it to simulate networks with planted communities where we modulate the detectability of the block structure. First, we test the performance of the network with unsupervised learning for community detection. Moreover, we simulate block-dependent features and labels for all the nodes, and we test how SBM-GNN performs for both community detection and labels prediction.

Then, we apply SBM-GNN for the prediction of novel cancer driver genes by using a recently published genomic dataset. Our model is able to correctly detect communities and it outperforms state-of-the-art methods for driver gene prediction. Nonetheless, we believe that the major advantage of SBM-GNN is the readily interpretable description of the whole interactome, which is also directly linked to the gene prediction task. This allows explaining how each gene participates in the cancer phenotype by testing each block for the enrichment of known functional pathways.

5.3 Discovering cancer driver genes and pathways using stochastic block model graph neural networks

The whole manuscript has been drafted by V. Fanfani, with the supervision and contributions of of G. Stracquadanio. The method was developed by V. Fanfani under the supervision of G. Stracquadanio and P. Liò. P. Liò and R. Vinas-Torme, contributed to the editing of the manuscript.

Errata corrige

In Eq. 7 (page 4): '$\mathbf{\hat{A}} = \mathbf{A} + \mathbf{I_n}$ is the adjacency matrix with added self-edges used by the 'renormalization trick' to avoid numerical instabilities'.

Amended caption for Fig. 4 (page 17): '**Block level organization of cancer driver genes.** From left to right, we group the genes according to their original label (known), the block they belong to, as inferred by SBM-GNN, for the SBM of size $k_s = 20, 10, 5$ (red, blue, green hues), and their predicted labels (Predicted). The size of each box depends on the number on genes falling into each category, for instance the 'cancer driver genes' group on bottom left represents the 169, out of $\sim$ 10,000 genes in the network, that are originally labelled cancer drivers. Between different boxes, we are plotting branches representing the number of nodes shared between the two groups. The SBM block $b20\_19$, dark red, shares many genes with $b10\_3$, light blue. Interestingly, SBM-GNN is not assigning all the driver genes to the same block (the cancer drivers, dark grey box on the left, are assigned to multiple blocks of in the SBM of size 20, red hue boxes), thus the inferred structure is not only dependent on their labels as expected. Moreover, it is worth noting that between different SBM layers there are consistent rearrangings, rather than clear hierarchies, possibly reflecting the complex structural and functional organisation of the network'.

Discovering cancer driver genes and pathways using stochastic block model graph neural networks

Viola Fanfani[1], Ramon Vinas Torne[2], Pietro Lio'[2], and Giovanni Stracquadanio *[1]

[1]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FD, United Kingdom

Abstract

The identification of genes and pathways responsible for the transformation of normal cells into malignant ones represents a pivotal step to understand the aetiology of cancer, to characterise progression and relapse, and to ultimately design targeted therapies. The advent of high-throughput omic technologies has enabled the discovery of a significant number of cancer driver genes, but recent genomic studies have shown these to be only necessary but not sufficient to trigger tumorigenesis. Since most biological processes are the results of the interaction of multiple genes, it is then conceivable that tumorigenesis is likely the result of the action of networks of cancer driver and non-driver genes.

Here we take advantage of recent advances in graph neural networks, combined with well established statistical models of network structure, to build a new model, called Stochastic Block Model Graph Neural Network (SBM-GNN), which predicts cancer driver genes and cancer mediating pathways directly from high-throughput omic experiments. Experimental analysis of synthetic datasets showed that our model can correctly predict genes associated with cancer and recover relevant pathways, while outperforming other state-of-the-art methods.

Finally, we used SBM-GNN to perform a pan-cancer analysis, where we found genes and pathways directly involved in the hallmarks of cancer controlling genome stability, apoptosis, immune response, and metabolism.

1 Introduction

The classical paradigm of cancer formation suggests that tumors arise from the stochastic accumulation of somatic mutations in key genes, called cancer driver genes, which give aberrant cells the ability to escape cell death and immune response and to grow uncontrollably throughout the body [1, 2].

The advent of high-throughput sequencing technologies has enabled the study of a broad spectrum of cancers and the identification of hundreds of driver genes across different malignancies [3, 4]. While the causal role of many genes in cancer has been confirmed by in-vitro and in-vivo models, recent studies have shown that driver mutations occur also in normal tissues [5, 6]. These observations suggest that driver mutations are necessary but not sufficient for tumorigenesis, and that the order in which they are acquired and the joint alteration of non cancer driver genes is important to transform a normal cell into a malignant one.

*Corresponding author. Email: giovanni.stracquadanio@ed.ac.uk

While the evolution of cancer cells can now be studied at sufficient resolution to generate testable hypotheses, discovering pathways of driver and non-driver genes associated with cancer has been challenging [7]. Current high-throughput omic assays provide only information with single-gene resolution, whereas gene and protein interaction experiments provide only pairwise information. It has now become apparent that methods able to perform multi-omic analyses at the pathway level are pivotal to understand the aetiology of cancer and design effective therapies.

In the last ten years, there have been substantial efforts to develop network analysis methods that would capture cancer poligenicity [7]. Nonetheless, current approaches typically focus on predicting either new cancer driver genes [8] or cancer driving pathways [9] by integrating multi-omic information; however, these methods usually aggregate multiple experimental information into a gene-level score, which effectively masks the effects and relationships between multiple biological processes underpinning cancer phenotypes.

Here we addressed current limitations in network-aware cancer analysis by developing a new method to simultaneously discover cancer driver genes and pathways by integrating gene-level high-throughput experiments with protein interaction information. To do that, we built a new deep learning model, combining graph neural networks (GNNs, [10]) and stochastic block models (SBMs, [11]), called Stochastic Block Model Graph Neural Network (SBM-GNN). GNNs provide a framework to obtain network-aware embeddings of gene level features; these models have been successfully applied to a number of tasks, including node labelling and link prediction [12, 10, 13], and have been shown to provide meaningful representations of omic data [14]. Embeddings are then combined with SBMs, a robust generative framework to model network connectivity, to infer new pathways. Importantly, our model can be fit end-to-end using standard gradient descent and scales efficiently with the size of the datasets.

To assess the performances of our method for discovering cancer driver genes and pathways, we built a simulation framework to generate synthetic networks with different structures and feature-level multi-modalities; experimental results show that our method is able to detect cancer driver genes and pathways with high accuracy. We then applied SBM-GNN to pan-cancer genome data [3]; here we found that our method outperforms other state-of-the-art approaches in identifying cancer driver genes, while being able to discover pathways associated with the hallmarks of cancer.

2 Methods

2.1 Model architecture

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, with $\mathcal{V}$ being the set of n vertices (or nodes) representing genes or proteins, and $\mathcal{E}$ being the set of m edges (or links) between nodes in $\mathcal{G}$. Edges can be represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, such that $\mathbf{A}_{uv} = 1$ iff node u is connected to node v in $\mathcal{G}$, and 0 otherwise. We define $\mathbf{X} \in \mathbb{R}^{n \times f}$ as the matrix of f gene features (e.g. mRNA abundance, number of somatic mutations), and $\mathbf{Y} \in \{0, 1\}^{n \times c}$ a matrix of c gene labels (e.g. being a cancer driver gene), such that $\mathbf{Y}_{uq} = 1$ iff node u has label q and 0 otherwise.

Here we hypothesise that node labels depend on the observed gene features and their involvement in pathways mediating the phenotypes of interest; this information is obviously unknown but can be learned from the data.

We denote with $\hat{\mathbf{Y}}, \hat{\mathcal{G}} = M(\mathcal{G}, \mathbf{X}, \mathbf{Y})$ a model that takes in input a graph $\mathcal{G}$, a feature matrix $\mathbf{X}$, and a label matrix $\mathbf{Y}$ and predicts new labels $\hat{\mathbf{Y}}$ and a new graph $\hat{\mathcal{G}}$. Our model consists of three layers: a network-aware embedding layer, a community detection layer, and a label prediction layer.

The network-aware embedding layer is used to learn a latent dense representation of the biological processes mediated by each gene and its immediate neighbours. The embedding for $\mathbf{X}$ can be computed using a non-linear transformation, ψ, defined as:

$$\hat{\mathbf{X}} = \psi(\mathbf{A}, \mathbf{X}) \tag{1}$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{n \times \hat{f}}$ is an $\hat{f}$-dimensional embedding conditioned on the observed node features $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$ of the graph $\mathcal{G}$.

We then wanted to detect how genes are organised in pathways. To do that, we implemented a community detection layer, which allows us to identify groups of nodes that are strongly connected to each other and that have homogeneous features. To perform community detection, we used a layer that models network structure using Stochastic Block Models (SBM) [15]. SBM is a generative model where edges between nodes depend on a community matrix $\mathbf{B} \in [0,1]^{k \times k}$ and a membership matrix $\mathbf{Z} \in [0,1]^{n \times k}$, where k is the number of unknown blocks. Each entry of the membership matrix, $\mathbf{Z}_{ik}$ denotes the probability that a node i belongs to the k-th block, which effectively corresponds to assigning genes to pathways. In a canonical SBM, $\mathbf{Z}$ is binary; however, genes often belong to multiple pathways, thus we relaxed this constraint by allowing for mixed membership, s.t. $\sum_k \mathbf{Z}_{ik} = 1$. Each entry of the community matrix $\mathbf{B}_{ij}$, instead, denotes the probability of observing an edge between two nodes belonging to blocks i and j, respectively (see Supplementary Materials, Supplementary Figure 1). In practice, we infer $\mathbf{B}$ from the matrix of observed edges $\mathbf{C} = \mathbf{Z}\mathbf{A}\mathbf{Z}^T$, where $\mathbf{C}_{ij}$ is the number of edges between blocks i and j [16].

However, since cancer is not only mediated by short-range but also long-range interactions, a naive approach combining a network embedding layer with a SBM layer will likely lead to poor performances. To overcome this problem, we designed our model to simultaneously learn multiple SBMs with a decreasing number of blocks, as a way to force the model to capture both short and long-range interactions. Here we defined a multi SBM layer, as a layer consisting of S SBMs with different numbers of blocks, k_s. W.l.o.g. we assume $s = 0$ to index the SBM with the smallest number of blocks; intuitively, high index SBMs represent fine-grained communities, whereas low index SBMs represent coarse-grained communities.

To learn the membership matrix $\mathbf{Z}^{(s)}$ for the s-th SBM, we apply a non linear network-aware transformation ζ to the embedding of the node features, $\hat{\mathbf{X}}$, as follows:

$$\mathbf{Z}^{(s)} = \texttt{softmax}(\zeta(\mathbf{A}, \hat{\mathbf{X}})) \tag{2}$$

where the `softmax` transformation ensures that $\sum_{i=1...k_s} \mathbf{Z}_i^{(s)} = 1$.

Finally, to perform node label prediction, we concatenated S membership matrices $\mathbf{Z}^{(s)}$ column-wise, and use the resulting matrix as input for the output layer, ϕ, as follows:

$$\hat{\mathbf{Y}} = \sigma(\phi([\mathbf{Z}^{(1)}|\mathbf{Z}^{(2)}|\ldots|\mathbf{Z}^{(S)}])) \tag{3}$$

where σ is a non-linear transformation suitable for binary or multi category classification.

2.1.1 Learning parameters of an SBM-GNN model

To fit our model, we defined a differentiable loss function, $\mathcal{L}$, as follows:

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{sbm} + \mathcal{L}_{mix} \tag{4}$$

where $\mathcal{L}_{class}$ is a supervised loss function for label prediction, whereas $\mathcal{L}_{sbm}$ and $\mathcal{L}_{mix}$ are unsupervised loss functions for learning communities. Specifically, given a membership matrix $\mathbf{Z}$ and a block matrix $\mathbf{B}$, $\mathcal{L}_{sbm}$ is the likelihood of observing m edges in $\mathcal{G}$ defined as:

$$\mathcal{L}_{sbm} = -(\mathbf{1}^\mathbf{T}(\mathbf{C}\ln(\mathbf{C}))\mathbf{1} - (\ln \mathbf{1}^\mathbf{T}\mathbf{Z})\mathbf{C}\mathbf{1} - \mathbf{1}^\mathbf{T}\mathbf{C}(\ln \mathbf{Z}^T\mathbf{1})) \tag{5}$$

and which in turn is averaged across each SBM layer [16]. Moreover, the contribution of $\mathcal{L}_{sbm}$ is weighted as a function of the number of epochs, such that the learning process is forced to minimise the community loss first and to learn how to classify the nodes later.

However, fitting multiple SBMs tends to assign nodes to only few blocks, effectively skipping learning community structures. To overcome this problem, we introduced a membership loss function, $\mathcal{L}_{mix}$, defined as:

$$\mathcal{L}_{mix} = -\left[\mathcal{H}(\mathbf{Z}\mathbf{Z}^T/n)\right]^{-1} \tag{6}$$

where $\mathcal{H}$ is the entropy function and n is the number of nodes; in practice, $\mathcal{L}_{mix}$ penalizes model configurations assigning all nodes to a single community.

2.1.2 Implementation

Our architecture can be easily tailored to different type of data and analyses by using appropriate transformations for ψ, ζ, ϕ. Our goal is to predict cancer driver genes and cancer associated pathways; thus, we defined our base model as follows:

$$\begin{aligned}
&\hat{\mathbf{X}} = \psi(\mathbf{A},\mathbf{X}) = \mathtt{ReLU}(\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{W}^{(0)})\\
&\mathbf{Z}^{(s)} = \mathtt{softmax}(\zeta(\mathbf{A},\mathbf{X})) = \mathtt{softmax}(\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{X}}\mathbf{W}^{(s)}) \quad \mathtt{for\ s\ in}\ \{1,...,S\}\\
&\hat{\mathbf{Y}} = \sigma(\phi([w_1\mathbf{Z}^{(1)}|\dots|w_S\mathbf{Z}^{(S)}])) = \mathtt{sigmoid}(\mathtt{dense}([w_1\mathbf{Z}^{(1)}|\dots|w_S\mathbf{Z}^{(S)}]))
\end{aligned} \tag{7}$$

where $\hat{\mathbf{A}} = \mathbf{A}-\mathbf{I_n}$ is the and $\hat{\mathbf{D}}$ is the node degree matrix of $\hat{\mathbf{A}}$. Layers ψ and ζ are implemented using Graph Convolutional Network (GCNs, [10]) layers, and ϕ is a fully connected layer. Importantly, we rescale membership information by learning weights, w_s, in order to identify the most relevant blocks with respect to the node labelling task.

We also explored other architectures, where we changed the ψ and ζ layers, while keeping fixed the fully connected layer to perform label prediction; we denoted each model configuration using a positional notation (see Table 1).

A possible limitation of our base model is that gene labels might be strongly correlated to gene features, and this information might not be captured only by the SBM layers. Here, we addressed this limitation by concatenating the embedding of gene-level features to the membership vectors, such that $Y = \sigma(\mathtt{dense}([w_1\mathbf{Z}^{(1)}|\dots|w_S\mathbf{Z}^{(S)}|\hat{\mathbf{X}}]))$; we refer to this layer as a residual layer (denoted by the suffix RES in the name) [17].

We also considered introducing pre-processing steps to speed up learning model parameters by augmenting our input node features. Biological network analyses have shown that graph diffusion is a powerful technique to extract information from protein interaction data ([18]). Thus, we decided to also test the use of graph diffusion as a pre-processing step using a sparsified version of a random walk diffusion matrix, called Graph Diffusion Convolution layer (GDC, [19], denoted by the prefix GDC in the model name).

Taken together, we tested 4 different model architectures that we then trained with either pre-processed and raw node features, and using residuals (see Table 1).

2.2 Synthetic data simulation

2.2.1 Synthetic gene networks

We simulated networks as a mixture of a perfect communities graph ($\mathcal{G}_{SBM}$) and a random Erdos-Renyi network ($\mathcal{G}_{ER}$) [20]. By varying a noise parameter η, we parametrised the contribution of noise and planted communities as $\mathcal{G} \sim (1-\eta)\,\mathcal{G}_{SBM} + \eta\ \mathcal{G}_{ER}$, whereas network sparsity is controlled by a density factor d (see Supplementary Materials, Supplementary Figure 2). As already shown, a stochastic block model with k communities and n nodes is defined by a stochastic matrix $\mathbf{B}$ and a node assignment matrix $\mathbf{Z}$; this model allows us to simulate different network structures by simply varying $\mathbf{B}$, $\mathbf{Z}$ and η.

With this model in place, we first simulated networks with multiple non overlapping assortative communities of approximately the same size, in order to assess whether our model is able to recover known communities. In this case, $\mathbf{B}$ is a diagonal matrix with $\mathbf{B}_{ii} = p_{\text{SBM}}$, while the noise is added as an Erdos-Renyi network with a constant probability of connection p_{ER} (see Supplementary Materials); with these parameters, we then generated k blocks harbouring approximately n/k nodes each.

We then used the Signal to Noise Ratio (SNR) as an indicator of community detectability, such that we can measure whether the SBM model is distinguishable from background noise. The SNR is defined as follows:

$$\text{SNR} = \frac{(a-b)^2}{2(a+b)} \tag{8}$$

where a and b are the average degree within and outside the community ($a = np_{\text{SBM}}$ and $b = np_{\text{ER}}$) and n is the total number of nodes; by modulating p_{ER} and p_{SBM}, we can control the SNR for a given network. Theoretically, SNR > 1 is the threshold for detectability, but it has been already shown that controlling for SNR > 1.5 is more reasonable [21].

However, since SBM-GNN is designed to detect multiple communities of different size, we simulated networks with hierarchical structure as follows; given a depth value, h, we generated h hierarchical layers of 2^h blocks each. Then, $\mathbf{B}$ is obtained as the average of the h layers, such that the smallest communities are the most assortative ones. Although there is no need to plant a hierarchical structure, this is a reasonable and realistic procedure to generate a network with multiple structured communities.

2.2.2 Synthetic gene features

We have also generated community-aware features conditioned on community structure, such that nodes belonging to closer communities are more likely to share similar features. To do that, we generate correlated random variables using the feature coloring method, which is the inverse of features whitening, a method routinely used to remove correlation between random variables.

Specifically, we first generated features as independent random variables drawn from a Normal distribution, $\mathcal{N}(\mu_i, \sigma)$, where the average μ_i depends on the i-th community the node belongs to (see Supplementary Materials); then, the coloring procedure is applied such that features are conditioned on the SBM structure of the network, which leads to features that are probabilistically more similar for nodes within the same community (see Supplementary Materials).

2.2.3 Community detection metrics

We then introduced two different metrics to assess community detection performances, namely the Jaccard coefficients, J_c, and the assignment penalty.

In our case, we measured J_c between each simulated block, R_i, and each block learnt by SBM-GNN, $\hat{R}_j$, obtained by assigning the nodes to the block with highest membership probability. Since the node assignment to the blocks is order invariant, for each learnt block we used $J_c(R_i, \hat{R}_j)$, where $\hat{R}_j = \operatorname{argmax}_j \; J_c(R_i, \hat{R}_j)$.

However, Jaccard coefficients do not measure the uncertainty of the node membership. Thus, we defined the assignment penalty metric, P_{ij}, between a known block assignment $\mathbf{Z}_{:i}$, that is the known assignment for each node to the i-th block, and the SBM-GNN assignment $\hat{\mathbf{Z}}_{:\mathbf{j}}$, as $P_{ij} = \left\| \mathbf{Z}_{:i} - \hat{\mathbf{Z}}_{:j} \right\|_F^2$, where $\| \|_F^2$ is the Froebenius norm. Similar to the Jaccard coefficient, we used the penalty metric for the j-th block as $P_j = \min_i(P_{ij})$. We then aggregated penalty scores into a single term, P_{tot}, by summing all penalties and normalising them w.r.t the number of nodes and the number of blocks, such that $0 \leq P_{tot} \leq 1$.

2.3 Cancer genome data and protein interaction datasets

We downloaded genomic data from the Pan Cancer Analysis of Whole Genomes (PCAWG) project [22], used in the companion pathway and network analyses [23, 24]. We then obtained p-values associated with the confidence that a genomic locus is a driver; specifically, for each gene, we considered p-values for coding (CDS) regions and 4 different types of non-coding regions, namely 3' UTR, 5' UTR, promoters and enhancers, and then used Fisher's method to transform p-values into χ^2 statistics.

We considered five different cancer panels as gene labels, both for training our model and evaluating its performances, including the pathway implicated drivers (PID) gene list and the COSMIC Cancer Gene Census (Cosmic) [25] (see Table 3).

Finally, we used two protein-protein interaction datasets: the STRING [26] and BioGRID [27] database. We processed both datasets to keep only high confidence interactions [24] and connected components; here we found BioGRID to be smaller and less dense than STRING (see Table 2).

3 Results

3.1 Performance on simulated data

We rigorously tested the performance of SBM-GNN on simulated networks generated by SBMs with controllable parameters, which is key to prove that our model is able to detect communities.

To do that, we simulated SBM networks with two blocks by drawing $p_{\text{SBM}} \sim$ Uniform$(0.5, 0.7)$ and adjusting the diagonal values of $\mathbf{B}$ to control the Signal to Noise Ratio (SNR), while using the identity matrix as gene features. Consistent with previously reported estimates, SBM-GNN was able to correctly detect the planted communities for SNR > 1.5 (see Figure 2A).

We then assessed the performance of our method on networks with more than 2 communities and in the presence of gene features. In this case, we simulated networks of 1000 nodes with 4 blocks as a mixture of SBM structures, with noise weighted by $\eta = \{0.1, 0.3, 0.6, 0.9\}$. For each network, we then simulated genes with 5, 10, 20 features and generated 5 replicates for each possible parameters setting; in this case, we used both uncorrelated and colored features. Here we found that SBM-GNN was able to accurately detect communities in presence of more than 2 blocks and with multiple annotations (see Figure 2B). Moreover, by increasing the number of correlated features, which corresponds to strengthening the signal, performances clearly improved (see Figure 2B).

After confirming that SBM-GNN can detect communities, we tested its performances on node labelling and compared it to other GNN models, including GCN[10], GAT[13], SAGE[12], and a LINEAR model, that is a neural network made of two fully connected layers that ignores graph structure. While other, more complex, methods might have better performances on specific datasets, these three methods are usually at the foundation of most state-of-the-art approaches.We used our simulation framework to build networks and datasets with 5, 10,20 colored genes features and 10% of positive labels (see Figure 2C), ultimately generating 5 datasets for each possible parameters setting. For $\eta < 0.5$, signal is weighted more than noise, and all graph-aware methods yield better performances than a random classifier, whereas the LINEAR model always had Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) $AUC \sim 0.5$. Interestingly, as the number of correlated features increases, SBM-GNN clearly outperformed the other models, suggesting that our approach is able to better exploits high-dimensional data.

Taken together, our simulations showed that SBM-GNN is a robust and effective architecture to identify network communities and to predict node labels using high-dimensional gene-level information.

3.2 Performance on pan-cancer genome data

We then used our model to perform cancer driver gene prediction and cancer pathways discovery using pan-cancer genomic data [3] and two protein-protein interaction datasets, namely STRING and BioGRID, while using the PID gene panel as node labels for training. Here we used both our base architecture and a set of 9 other SBM-GNN extensions, since fine tuning the layers and parameters is often key to yield the best performances in deep learning (see Supplementary Materials).

Interestingly, we found that all architectures performed remarkably well (AUC > 0.7) regardless of the protein interaction dataset used, albeit STRING seems to provide consistently better results (see Figure 3A). Moreover, we found that adding the residual information consistently improved the performances of our model, and similarly, but to a lesser extent, the use of graph diffusion.

We then compared the performance of SBM-GNN with that of GCN [10], GAT [13], SAGE [12], and a basic model with two fully connected layers, LINEAR (see Supplementary Materials). Most SBM models achieved comparable performances with the other deep learning models (see Table 4), but the residual version of SBM-GNN consistently achieved the best performances, with GCN being the best alternative.

3.2.1 Comparison with state-of-the-art cancer driver prediction methods

We then compared SBM-GNN performance with state-of-the-art methods for cancer driver gene prediction (see Supplementary Materials). There is a vast literature on methods designed to identify new cancer driver genes, with many of them using network information [8, 28, 29]. Recently, the *using Knowledge In Networks* (uKIN) [8] has been shown to be the gold standard in the field, and hence we used it as a benchmark to analyse SBM-GNN performance. Conversely, methods for community detection are less established. There are a plethora of methods that detect cancer-associated submodules, that are connected subset of genes driving cancer [9, 30], but they do not explain the structure of all the nodes that are not implicated in cancer. However, for completeness, we compared SBM-GNN performance with Hierarchical Hotnet (HHotNet) [9] to explore the performance of this class of methods (see Supplementary Materials). It is worth noting that HHotNet and uKIN are unsupervised methods, albeit the latter uses information on known cancer driver genes to guide the diffusion process. For this reason, we carried out comparisons both with models trained and tested on the PID labels, and by training on the COSMIC panel and testing on PID. SBM-GNN performance was consistent with those of uKIN and outperformed it when using PID labels (see Figure 3C). Unsurprisingly HHotNet had a clearly worse performance. While results are not directly comparable, it is interesting to observe that submodule inference is not directly applicable to identify new cancer drivers.

3.2.2 Discovering cancer pathways by inspecting stochastic block models

Our method is not limited to the prediction of cancer driver genes, but more importantly provides an interpretable picture linking cancer driver genes to the pathways they are involved in. Our hypothesis is that cancer driver genes are targeting multiple, possibly distant pathways, and that complex relationships might be captured by fitting multiple SBMs. Thus, using our model, we analysed the blocks harbouring cancer driver genes, the relationship between SBMs, and how the learnt blocks can be used to identify cancer pathways.

We selected the model with the lowest loss and analysed the genes assigned to each block. As we hypothesised, cancer drivers genes are assigned to multiple blocks, both in the coarser and finer SBM layers (see Figure 4). This observation is consistent with the fact that cancer driver genes, such as *TP53* or *MYC*, are involved in multiple biological processes. Interestingly, in most cases, cancer driver genes represent less than 20% of the genes in a block, which is consistent with a model of tumorigenesis where driver genes mediate cancer phenotypes by interacting with non driver ones.

Finally, we functionally characterised the genes in the blocks identified by SBM-GNN, in order to provide a system-level picture of gene organisation encoded by the SBM. To do that, for each block of genes, we performed a Fisher's exact test using the Reactome genesets (see Figure 5). Since we are organising the network in communities, we expect to find blocks recapitulating cancer-associated pathways, alongside others not mediating cancer phenotypes.

At the higher level, our model identified blocks of genes associated with hallmarks of cancer [31], in particular programmed cell death, metabolism which is associated with deregulation of cellular energetics, and genome instability through alteration of the DNA repair and replication machineries. Specifically, by looking at the 20 block SBM, we found one ($\texttt{b}20\_16$) having 86% of its genes been predicted as cancer driver genes, which recapitulates 72% (OR:45.1, p: 1.43×10^{-8}) of genes involved in the *Calcineurin activates NFAT* process, that is a T-cell related process involved in cancer progression and metastasis [32, 33], 70% (OR: 44.4, p: 2.32×10^{-22}) of the genes involved in Adjerens Junctions Interactions, and 60% (OR: 23.6, p: 1.11×10^{-6}) of the genes associated with *Repression of WNT target genes* pathway, which are both well known processes involved in cell motility and proliferation. Moreover, we also found another block ($b20\_6$) encompassing the PIK3 cascade pathway (OR: 5.74, p: 2.07×10^{-6}) and multiple *FGFR1* and *FGFR2* signalling processes, which are known to be critical for tumorigenesis [34]. Conversely, for example, block B20_11, which does not harbor any cancer driver gene, is associated with non-cancer related processes, such as *Olfactory Signaling*.

Taken together, we have shown that our model is able to recover genes and pathways associated with well characterised cancer mediating processes; in particular, by learning SBMs with an exponentially growing number of clusters, it is possible to go from broad molecular hallmarks to specific biological processes.

4 Conclusions

Tumorigenesis is triggered by a complex molecular reprogramming mediated by genetic, genomic, and molecular alterations acquired by driver and non-driver genes. While it is now possible to quantitatively assess the impact of these changes at the single gene level, reconstructing a system-level picture of cancer cells remains a challenging task.

Here we introduced a new model, called SBM-GNN, combining recent geometric deep learning architectures with stochastic block model to simultaneously infer cancer driver genes and associated pathways; our model provides a scalable approach to integrate multi-omic data with protein-interaction information, that can be used to generate testable hypothesis at the gene and pathway level. We validated our method using an extensive set of simulations showing that SBM-GNN can correctly identify cancer driver genes and cancer related pathways.

We then applied our method to the analysis of pan-cancer genomic data, where we showed that SBM-GNN can predict driver genes with high accuracy and identify blocks of genes associated with well-know hallmarks of cancer. On this point, the ability of our model to learn an easily interpretable pathway organization of cancer genes provides new opportunities to dissect the system-level reprogramming underwent by cancer cells.

Contributions

G.S. and V.F. conceived the study. V.F., G.S. and P.L. designed the model. G.S., V.F. P.L. and R.V.T. developed the simulation framework. V.F. wrote the software and performed all analyses, supervised by G.S. G.S., V.F. and P.L. analysed the data. G.S. and V.F. wrote the manuscript with contributions from all the authors.

Acknowledgements

This work has been supported by the Wellcome Trust Seed Award in Science (207769/A/17/Z) to G.S. The GPU hardware used for this research was donated by the NVIDIA Corporation to G.S.

References

[1] B. Vogelstein et al. “Cancer genome landscapes”. In: *Science* 340.6127 (2013), pp. 1546–1558. ISSN: 10959203.

[2] W. C. Hahn et al. “An expanded universe of cancer targets”. In: *Cell* 184.5 (2021), pp. 1142–1155.

[3] P. J. Campbell et al. “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2020), pp. 82–93. ISSN: 14764687.

[4] M. H. Bailey et al. “Comprehensive Characterization of Cancer Driver Genes and Mutations”. In: *Cell* 173.2 (2018), 371–385.e18. ISSN: 0092-8674.

[5] H. Lee-Six et al. “The landscape of somatic mutation in normal colorectal epithelial cells”. In: *Nature* 574.7779 (2019), pp. 532–537.

[6] L. Moore et al. “The mutational landscape of normal human endometrial epithelium”. In: *Nature* 580.7805 (2020), pp. 640–646.

[7] K. Ozturk et al. “The emerging potential for network analysis to inform precision cancer medicine”. In: *Journal of molecular biology* 430.18 (2018), pp. 2875–2899.

[8] B. H. Hristov, B. Chazelle, and M. Singh. “uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes”. In: *Cell Systems* 10.6 (2020), 470–479.e3. ISSN: 24054720.

[9] M. A. Reyna, M. D. Leiserson, and B. J. Raphael. “Hierarchical HotNet: Identifying hierarchies of altered subnetworks”. In: *Bioinformatics*. Vol. 34. 17. Oxford University Press, 2018, pp. i972–i980.

[10] T. N. Kipf and M. Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: (2016). arXiv: `1609.02907`.

[11] B. Karrer and M. E. Newman. “Stochastic blockmodels and community structure in networks”. In: *Physical review E* 83.1 (2011), p. 016107.

[12] W. L. Hamilton, R. Ying, and J. Leskovec. *Inductive Representation Learning on Large Graphs*. Tech. rep. 2017.

[13] P. Veličković et al. “Graph attention networks”. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2018. arXiv: `1710.10903`.

[14] R. Schulte-Sasse et al. “Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms”. In: *Nature Machine Intelligence* 3.6 (2021), pp. 513–526.

[15] C. Lee and D. J. Wilkinson. “A review of stochastic block models and extensions for graph clustering”. In: *Applied Network Science* 4.1 (2019), pp. 1–50. ISSN: 23648228. arXiv: `1903.00114`.

[16] Z. Chen et al. “Neural Stochastic Block Model & Scalable Community-Based Graph Learning”. In: (2020). arXiv: `2005.07855`.

[17] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. IEEE Computer Society, 2016, pp. 770–778. ISBN: 9781467388504. arXiv: `1512.03385`.

[18] L. Cowen et al. “Network propagation: A universal amplifier of genetic associations”. In: *Nature Reviews Genetics* 18.9 (2017), pp. 551–562. ISSN: 14710064.

[19] J. Klicpera, S. Weißenberger, and S. Günnemann. “Diffusion improves graph learning”. In: *arXiv* (2019). ISSN: 23318422. arXiv: 1911.05485.

[20] B. Karrer and M. E. Newman. “Stochastic blockmodels and community structure in networks”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 83.1 (2011). ISSN: 15393755. arXiv: 1008.3926.

[21] Z. Chen, J. Bruna, and L. Li. “Supervised community detection with line graph neural networks”. In: *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, 2019. arXiv: 1705.08415.

[22] I. The, T. P.-C. A. of Whole, G. Consortium, et al. “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2020), p. 82.

[23] E. Rheinbay et al. “Analyses of non-coding somatic drivers in 2,658 cancer whole genomes”. In: *Nature* 578.7793 (2020), pp. 102–111. ISSN: 14764687.

[24] M. A. Reyna et al. “Pathway and network analysis of more than 2500 whole cancer genomes”. In: *Nature Communications* 11.1 (2020), p. 16. ISSN: 20411723.

[25] Z. Sondka et al. “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.” In: *Nature reviews. Cancer* 18.11 (2018), pp. 696–705. ISSN: 1474-1768.

[26] D. Szklarczyk et al. “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic Acids Research* 47.D1 (2019), pp. D607–D613. ISSN: 0305-1048.

[27] C. Stark et al. “BioGRID: a general repository for interaction datasets”. In: *Nucleic Acids Research* 34.90001 (2006), pp. D535–D539. ISSN: 0305-1048.

[28] A. Cho et al. “MUFFINN: Cancer gene discovery via network analysis of somatic mutation data”. en. In: *Genome Biology* 17.1 (2016). ISSN: 1474760X.

[29] M. D. Leiserson et al. “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes”. In: *Nature Genetics* 47.2 (2015), pp. 106–114. ISSN: 15461718. arXiv: 15334406.

[30] M. A. Reyna et al. “NetMix: A Network-Structured Mixture Model for Reduced-Bias Estimation of Altered Subnetworks”. In: *Journal of Computational Biology* (2021). ISSN: 1066-5277.

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

[32] C. Tran Quang et al. “The calcineurin/NFAT pathway is activated in diagnostic breast cancer cases and is essential to survival and metastasis of mammary cancer cells”. In: *Cell Death and Disease* 6.2 (2015). ISSN: 20414889.

[33] M. Mancini and A. Toker. *NFAT proteins: Emerging roles in cancer progression*. 2009.

[34] S. Forbes et al. “COSMIC: high-resolution cancer genetics using the catalogue of somatic mutations in cancer”. In: *Current protocols in human genetics* 91.1 (2016), pp. 10–11.

Tables

Model	ψ	ζ
GCN-GCN	GCN	GCN
LIN-GCN	fully connected	GCN
GCN2-GCN	GCN depth 2	GCN
GCN-LIN	GCN	fully connected

Table 1: **SBM-GNN implementations**. For each implementation used in our study, we report the type of layer used for the network-aware embedding function, ψ, and the membership assignment function, ζ.

	# nodes	**# edges**	**Density**	**Max degree**	**Median degree**	**Mean clustering**
STRING	10224	205422	0.003931	1882	11.0	0.46084
BIOGRID	5440	16346	0.001105	248	3.0	0.187394

Table 2: **Protein-protein interaction network properties.** For each network, we report the number of nodes, the number of edges, the density, the maximum and median node degree, and the median clustering coefficient.

	# genes	**BIOGRID**	**STRING**
PID	169	134	161
COSMIC	719	481	610

Table 3: **Cancer driver genes panels.** For each cancer driver gene panel, we report the number of genes, and the number of genes mapped to genes in BIOGRID and STRING.

Network	Model	Precision	Recall	BACC	AUC
BIOGRID	GAT	0.0798	0.6194	0.7174	0.8086
	GCN	0.0798	0.6294	0.7212	0.8282
	LINEAR	0.0250	**1.0000**	0.5000	0.7814
	SAGE	0.0832	0.7024	0.7516	0.8236
	GDC-LIN-GCN	0.0662	0.5316	0.6684	0.7372
	GDC-LIN-GCN-RES	**0.0894**	0.7120	**0.7626**	**0.8568**
	GCN-GCN	0.0764	0.5952	0.7046	0.7984
	GCN-GCN-RES	0.0772	0.6440	0.7226	0.8180
STRING	GAT	0.0418	0.8366	0.6620	0.6844
	GCN	0.0604	0.7554	0.7824	0.8660
	LINEAR	0.0210	**0.9346**	0.5420	0.8150
	SAGE	0.0630	0.8044	0.8052	0.8626
	GDC-LIN-GCN	0.0524	0.6654	0.7358	0.8350
	GDC-LIN-GCN-RES	**0.0694**	0.8368	**0.8270**	**0.8948**
	GCN-GCN	0.0508	0.6368	0.7218	0.8046
	GCN-GCN-RES	0.0598	0.7554	0.7810	0.8478

Table 4: **Performance analysis of different models on cancer driver genes prediction**. Models were trained using either the BIOGRID or the STRING protein-interaction network, and the PID panel of cancer driver genes. For each model, we report the precision, recall, balanced accuracy (BACC) and Area Under the ROC Curve (AUC) averaged over 5 independent runs. We report in bold face the best performance for each metric, and the best overall model.

Figures

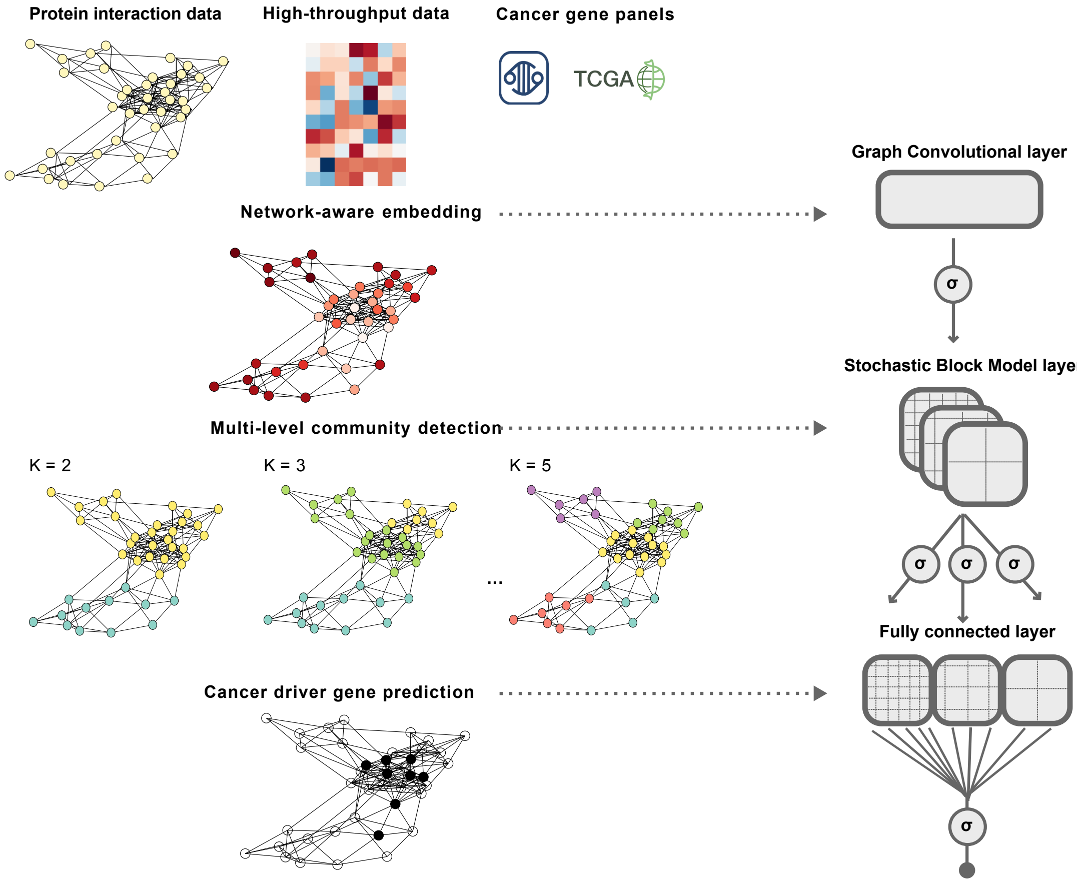


Figure 1: **An overview of the Stochastic Block Model Graph Neural Network (SBM-GNN) model.** Here we present a sketch of the processing steps (left) implemented by SBM-GNN into a deep neural network (right). Our model takes in input protein interaction data, high-throughput omics data and a cancer driver gene panel, which classifies each node as being a driver or not. Then, a Graph Convolutional layer is used to generate a gene-level network-aware feature embeddings, which in turn are used to assign genes to communities learned by multiple Stochastic Block Model (SBM) layers with varying number of blocks. Finally, the network structure learned by SBM layers is used as input for a fully connected classifier to predict cancer driver genes.

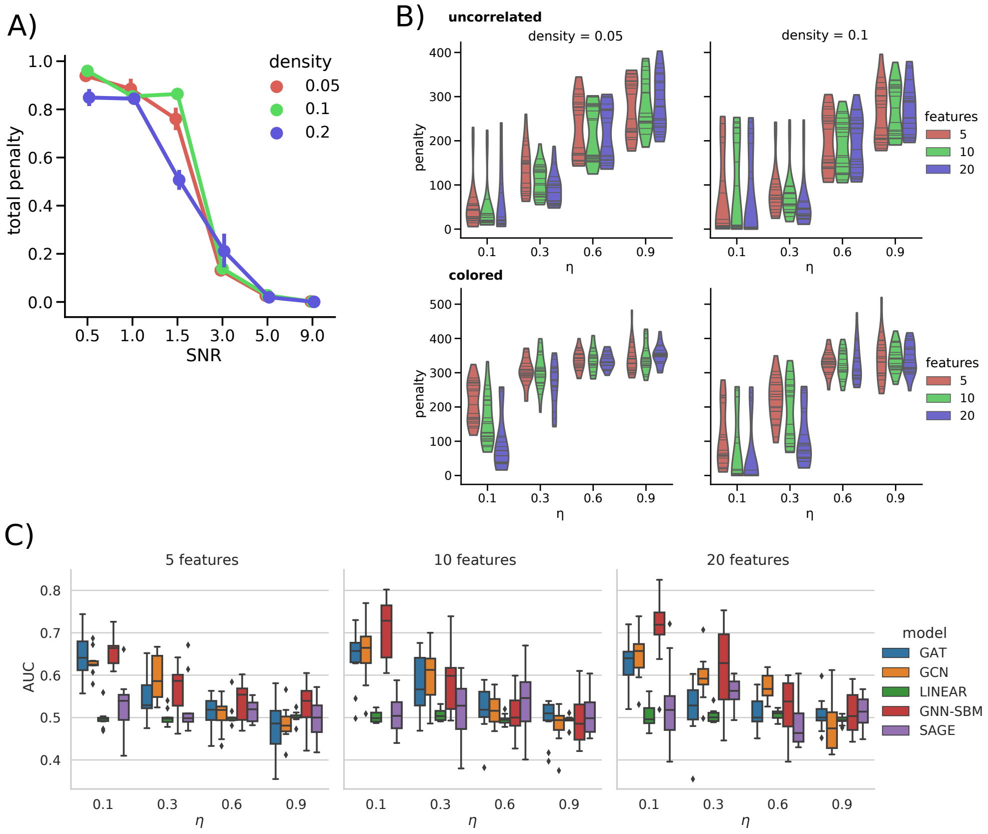


Figure 2: **SBM-GNN performance on synthetic data**. A) On the y-axis, we report the total normalised penalty obtained for 2 block SBM networks, whereas the x-axis reports SNR values and colors denote different density parameters. For for SNR$= 0.5$, non-detectable blocks, the total penalty is approximately 1, which corresponds to the maximum error. Total penalty drops for SNR> 1.5, confirming that our model is able to recover the network structure. B) Performance on uncorrelated and colored features for simulated networks with 4 blocks. On the x-axis, we show the penalty value on block assignments at varying levels of noise, η, where communities should become detectable for $\eta < 0.5$. We report results for $5, 10, 20$ gene features (colored) and for network density $d = \{0.05, 0.1\}$ (columns). C) Classification performance (AUC) of SBM-GNN , GAT, GCN,SAGE, and LINEAR architectures (color) on colored features and synthetic networks with 4 blocks. Results are shown for different η (x-axis) and for $5, 10, 20$ features (columns). For low η values, all graph neural networks have performances significantly better than a random classifier (AUC> 0.5). Interestingly, as the number of correlated features increases, we observed a significant improvement on SBM-GNN performance.

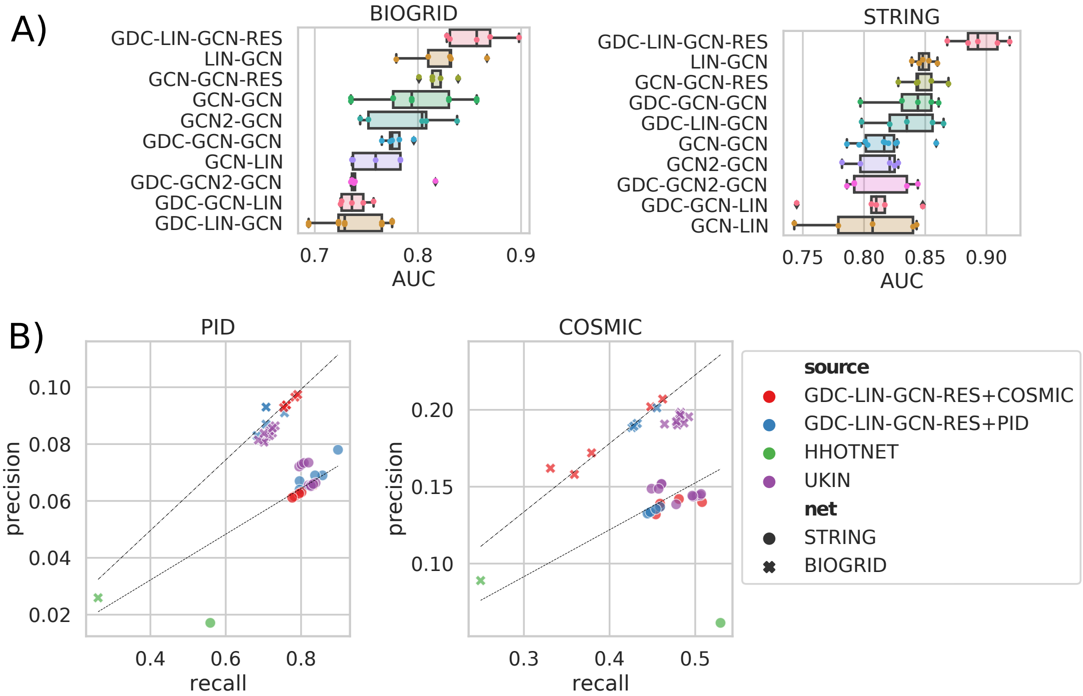


Figure 3: **SBM-GNN performance on cancer driver genes prediction**. A) Cancer driver gene prediction measured as the area under the receiver operating characteristic curve (ROC AUC), for different SBM-GNN architectures. We have sorted them by average performance over 5 runs. For both the BIOGRID and STRING networks, we found GDC-LIN-GCN-RES architecture achieves the best performances; interestingly, adding residuals information increase the AUC for both SGCN-GCN and LIN-GCN models. It is also worth noting that all neural networks without a network-aware block assignment layer, namely GCN-LIN, GDC-GCN-LIN, have consistently worse performances. B) We compare the recall (x-axis) and precision (y-axis) of hierarchical hotnet (HHotNet), UKIN, and SBM-GNN (colors) on BIOGRID and STRING networks (markers). Dashed lines are the theoretical relationship between precision and recall at the specific percentile for each network. For UKIN and SBM-GNN, we reported as cancer driver genes those above the 80th percentile of their scores. Here, we found SBM-GNN with residuals to be the best performing method, a trend we also observed when we trained our model on the COSMIC gene panel.

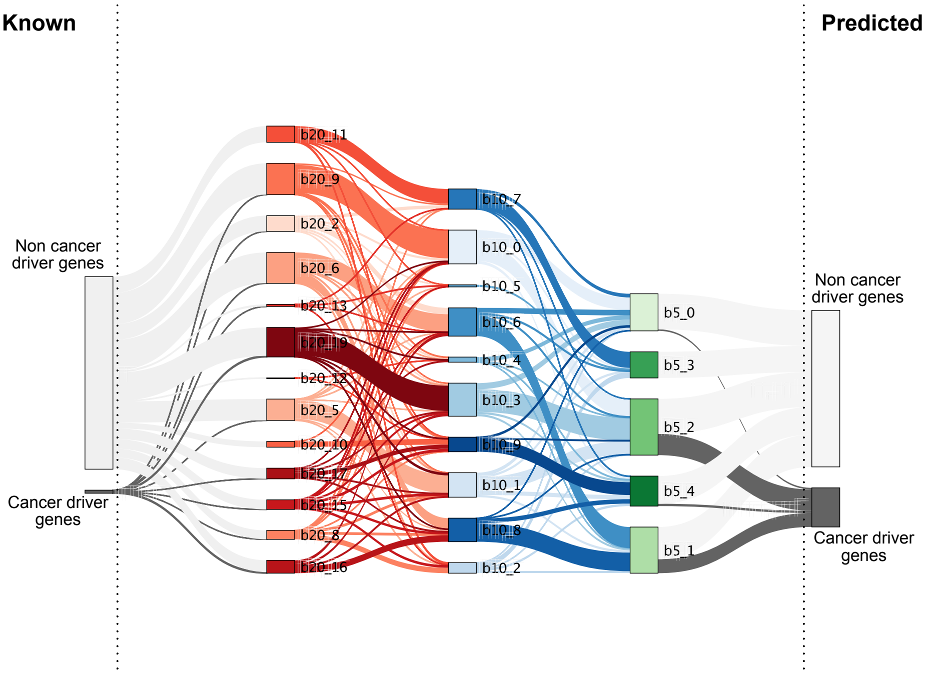


Figure 4: **Block level organization of cancer driver genes.** From left to right, we show the cancer labels, the blocks of size $k_s = \{20, 10, 5\}$ and the predicted labels. Between different nodes, we are plotting branches representing the number of nodes shared between the two blocks. The last blocks on the right are those that SBM-GNN predicts as significant.

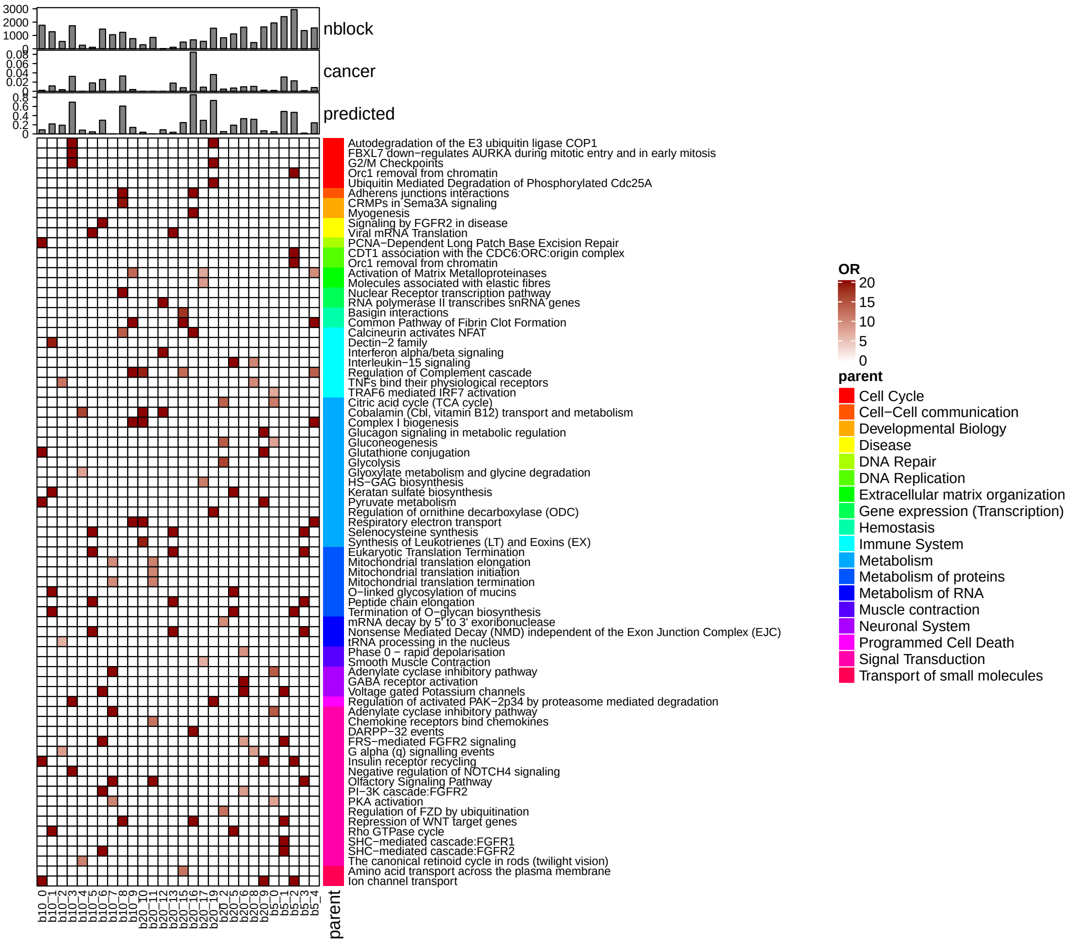


Figure 5: **Block level Reactome geneset enrichment analysis.** For each block (columns), we plot the top five statistically significant genesets ranked by odds-ratio (OR) from the Fisher's exact test. For each Reactome geneset (rows), we also report their parental group (color annotation), which allows us to identify macro functional classes. For each block, we show, from top to bottom, the number of genes in the block, the number of known cancer driver genes in the block and the fraction of those predicted by our model.

5.4 Conclusions

We have presented a new model, SBM-GNN, combining geometric deep learning and stochastic block models to simultaneously infer cancer driver genes and associated pathways. By applying it to curated pan-cancer genomic data, we have been able to obtain accurate predictions of driver genes and we have shown how the combination of network structure and node function can be used to explain our prediction.

This model has the potential to solve many of the issues raised and described in this thesis. First, using the GNN readily allows the integration of multi-modal data and PPI networks. By using the GCN and Graph Diffusion Convolution (GDC) layers we are effectively applying a multi-feature diffusion process, which has been shown to be well suited for biological applications [Cowen et al. 2017]. Moreover, the SBM can automatically detect submodules implicated in tumorigenesis and link them to known functional pathways and processes.

Nonetheless, we recognise that this work has some limitations. Surely, the model could, and should, be extended to a more comprehensive and realistic analysis of multi-omics cancer data. While we were able to obtain state-of-the-art performances with the genomic dataset, it is worth noting that the features we used are highly-curated statistics obtained only from mutational data. Actual multi-omics annotated data and comparable methods [Schulte-Sasse et al. 2021] will provide further evidence of SBM-GNN performance and explainability. Here we need to specify that the only directly comparable work [Schulte-Sasse et al. 2021] was published as we were finalising the analyses of this chapter. We do anticipate that a detailed comparison will be included in future applications of the method.

Additionally, we have shown that the general architecture of SBM-GNN can be adapted and extended by changing the layers within it. PPI network features have been shown to be predictive of node connections at distance 3 (three edges between each other) [Kovács et al. 2019]. From a biological point of view, this observation is explained with similar proteins that, rather than being in contact with each other, are affecting two proteins expressed in different tissues, through the same mechanism. Moreover, in another study, it has been hypothesised that assortative and disassortative filters [Kavran and Clauset 2020] might be improving the community detection performance. So far, we have only applied GCN or MLP, but we believe that applying different filters might be functional to detect different properties of the graph.

Eventually, we realise that we have not systematically investigated how different graphs, and features, influence the inferred communities. For instance, a comparison between PPI networks could reveal how much and when different graphs influence the learning task. While consensus procedures [Reyna, Leiserson, and Raphael 2018] can be applied to retrieve the blocks that are frequently inferred, a deeper analysis of how missing links influence prediction would be extremely valuable in practice. Moreover, future studies could employ weighted networks, e.g. edges weighted as the probability or confidence of being an actual PPI; the SBM-GNN structural loss is already accounting for the strength of interaction between and within blocks, and the improved performance of the GDC layer shows that weighted networks can be directly employed within the architecture. Finally, it would be interesting to investigate the sensitivity to feature variability. Indeed, precision medicine strives to obtain actionable patient-level predictions [Ozturk et al. 2018]; while in practice we can directly process multi-omics patient-level data with SBM-GNN, it would be interesting to see how stable the communities are and how accurate the node label prediction would be.

6 Conclusions

This thesis investigates the mechanisms underpinning tumorigenesis with particular focus on how novel computational methods can help decode and integrate the wealth of data that is being generated by high-throughput experiments. Beyond the classical paradigm of tumor formation, it is now evident that integrative studies are better suited to tackle the heterogeneity and poligenicity of cancer. PPI networks provide genome-wide maps for the investigation of the interconnection between different driving mechanisms, and we have here proposed two approaches for the detection of system-level cell reprogramming in cancer. However, for a complete picture of how tumors arise and evolve, experimental evidence, either at the genomics, transcriptomics, epigenomics level, of the effects of somatic aberrations, need to be integrated with the underlying, inherited, genetics of the organism.

In chapters 2 and 3 we have tackled the issue of cancer risk in the broader population. We developed BAGHERA, a statistical learning method for the estimation of gene-level heritability, that aggregates SNP summary statistics to evaluate the cancer risk explained by each gene and their cis-regulatory elements. Compared to state-of-the-art methods, BAGHERA is able to draw genome-wide maps of cancer heritability at higher resolution. We applied

BAGHERA to 38 histologically characterised cancer types in the UKBB and retrieved 1,146 Cancer Heritability Genes, which are those with heritability significantly higher than expected by chance. By investigating the functional role of these genes, we observed that they are recurrently involved in known cancer-related biological processes. More interestingly, we have also found that these cancer heritability genes were also previously reported somatic drivers and in particular tumor suppressors.

In chapters 4 and 5 we proceeded to study how the integration of high-throughput experimental results with Protein-Protein Interaction networks can help decode the mechanisms underpinning cancer beyond the single gene hits. PyGNA uses network topology to characterise the connectivity between phenotype-relevant genes identified by experimental data, such as differentially expressed genes. By mapping genesets onto the PPI network, PyGNA can test their topological properties, such as whether they are strongly interacting with each other, likely related to cooperation on a biological level, and whether they are hubs, affecting many others genes.

Nonetheless, graph topology alone is insufficient to provide functional explanations of the observed networks and do not fully characterise datasets with multiple modules, as is often the case of complex diseases reprogramming different pathways. Thus, we developed a deep learning method for the integration of multi-modal data and PPI networks. We hypothesise that a node's function is dependent both on its observed features and the communities it belongs to. SBM-GNN is a graph convolutional neural network that uses supervised learning to predict novel cancer drivers combining multiple attributes for each node, e.g. multi-omics datasets, and does inference of communities within the whole networks, through stochastic block models. GNNs allow to readily integrate functional and structural data, and with our architecture, we are able to provide an explainable model for tumorigenesis that links cancer drivers to the pathways within the graph.

Taken together, the integrative study of germline and somatic variation, and of the interactions between and within each other, has the potential to lead to advancement in system biology that can then improve personalised medicine.

Methods for the detection of germline and somatic cancer driver genes have largely based their predictions on the frequency, or strength of association, of a variant, in practice prioritizing those that have a clearer functional impact [Martínez-Jiménez et al. 2020]. However, difficulties in revealing driver events for all patients [Campbell et al. 2010] and cancer poligenicity [Stracquadanio et al. 2016] justify the development of integrative models to better tackle tumor heterogeneity. The aggregation of SNP effects [Huang et al. 2011] allows us to account for the effects of variants that would not reach statistical significance in the GWAS but might be targeting key programs in the cell [Fagny et al. 2020]. Moreover, the study of cancer at the interactome level has the goal of detecting groups of functionally related genes, that, when mutated, trigger similar downstream effects or that synergistically affect normal cell functions [Ozturk et al. 2018].

In chapter 3, we highlighted that BAGHERA does not explicitly include any orthogonal evidence of SNP function, for instance, eQTL data [Aguet et al. 2020]. We expect that the integration of functional information within the cancer heritability model could improve its predictive capabilities prioritizing the mechanisms that explain the increased risk [Gallagher and Chen-Plotkin 2018; Wainberg et al. 2019]. Conversely, in chapter 5 we did not explicitly explore the relationship between germline and somatic variation in the interactome context. This will surely be an ensuing work, as we have evidence of the co-occurrence of inherited and environmental aberrations [Vosoughi et al. 2020; Zhang et al.

2021; Carter et al. 2017], but we would like to explore how they organise in the network and to what extent they target the same pathways [Carbone et al. 2020]. Taken together, the integration of multi-modal data [Silverbush et al. 2019] and the detection of the effects of variations on the network modules [Deritei et al. 2019] have the potential to find emerging patterns of functional pleiotropy that go beyond the local two-hit model of tumorigenesis.

From a personalised medicine perspective, a system biology picture of the mechanisms underpinning cancer would also enable the detection of novel markers of risk and therapeutic opportunities. Indeed, while we have here focused on tumorigenesis, cancer drivers are those that confer a selective advantage to the cell, hence they can also be used to understand tumor progression and treatment. Low-penetrance variants can mediate pathways that, alongside environmental cancer risk factors, increase cancer risk, and affect tumor progression and drug response [Surakhy et al. 2020; Jeffers et al. 2021]. While BAGHERA does not inform on the specific functional effects of each heritability locus, it produces testable hypotheses and would allow the functional validation of single-locus dysregulation. In chapter 4 we reported that network studies have been able to stratify cancer types, detecting subtype-specific modules [Hofree et al. 2013; Chaudhary et al. 2018]. Furthermore, novel computational methods are trying to directly address the problem of personalised network reconstruction, for the direct identification of patient's driver genes and pathways [Ozturk et al. 2018]. Moreover, integrative network studies of cancer aberrations have led to novel drugs being clinically tested and repurposed [Hahn et al. 2021]. While cancer drivers are not always targetable, the detection of secondary risk factors and connected pathways might instead provide opportunities for patient surveillance and drug repurposing [Ruiz, Zitnik, and Leskovec 2021].

Eventually, it is worth noting that computational methods should evolve alongside experimental technologies [Muir et al. 2016]. This thesis has analysed and discussed, for the most part, bulk-sequencing data. Conversely, single-cell omics experiments are now becoming more frequent and allow to capture within- and between- tumor heterogeneity, and we should expect other cutting-edge methodologies to be developed in the future. Thus, we expect the resolution and quality of the data to increase, for instance with higher-throughput PPI network experiments and with alignment and variant calling methods that fully exploit the update reference genome. As there is no free lunch when modeling biological data and careful considerations need to be made at each step of the analysis, data integration methods will have to adapt and to even bigger datasets, but they might have further ground truth models to train the model on.

A Appendix A

A.1 Supplementary materials for “Dissecting the heritable risk of breast cancer: from statistical methods to susceptibility genes”

Dissecting the heritable risk of breast cancer: from statistical methods to susceptibility genes

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

June 1, 2020

Contents

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

[†] Department of Haematooncology, University Hospital Ostrava, Czech Republic

[‡] 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.

1 Supplementary Tables

All supplementary tables are in two files in XLSX format. We separated the tables relative to all variants and genes from those of the functional characterisation. All fields in these files are annotated.

Supplementary tables:

- Supplementary table 1: Curated list of breast cancer SNPs.
- Supplementary table 2: Curated list of genome-wide significant breast cancer SNPs. reported by more than one study.
- Supplementary table 3: Curated list of genes harboring breast cancer SNPs.
- Table 2: Curated list of breast cancer SNPs above the 95th percentile of the OR distribution.

Supplementary tables pathways:

- Supplementary table pathways 1: Curated list of genes harboring breast cancer SNP mapped to the Gene Ontology slim terms.
- Supplementary table pathways 2: Curated list of genes harboring breast cancer SNP to the KEGG pathways.
- Supplementary table pathways 3: Curated list of genes harboring breast. cancer SNP mapped to the cancer driver genes annotations and to the DNA repair genes list.
- Supplementary table pathways 4: Gene Ontology slim enrichment analysis.
- Supplementary table pathways 5: KEGG pathway enrichment analysis.
- Supplementary table pathways 6: Hallmark of cancer enrichment analysis.

2 Supplementary Methods

2.1 GWAS catalog data

We performed our analyses using the GWAS Catalog (downloaded on 12/07/2018), a curated dataset of 143963 variants from 4054 studies. Unless otherwise noted, all variants are mapped to Genome Assembly GRCh38.p12 and dbSNP Build 151.

We then considered SNPs in European populations, either in the initial or replication cohort, for studies whose primary trait of interest was breast carcinoma and breast cancer, thus removing studies related to childhood cancer, treatment, survival, mortality. With these parameters, we identified 1289 SNPs in total.

We then restricted our analyses to genome-wide significant SNPs ($p \leq 5 \times 10^{-8}$), ultimately identifying 719 variants across 26 studies, with only 421 unique loci. For each variant, we considered effect size as odds ratios (OR), either using the reported estimate in the study or by setting OR $= exp(\beta)$, where β is the reported regression coefficient.

We assigned SNPs to genes using the GENCODE annotation for Genome Assembly GRCh38.p12 (downloaded on 27/07/2019); specifically, each SNP is assigned either to all overlapping genes or to the closest gene within a 50 Kb window, if the variant is located in an intergenic region. Overall 421 SNPs are annotated to 311 genes, whereas 56 of them are intergenic. We then removed genes containing SNPs whose association has been reported only once. We also remove genes with name starting with AC**, since they do not have a Hugo Symbol but are using the accession number of NCBI. Using our criteria, we identified 104 genes harboring SNPs associated with breast cancer risk.

We provide a visual representation of the curation and filtering workflow in Supplementary Figure S2.

2.2 Biological characterisation

We performed biological characterisation of the 104 genes using standard enrichment analysis. In particular we tested whether, and to what extent, genes harboring breast cancer SNPs are enriched in any gene ontology term or KEGG pathway. Our analysis shows that many genes are associated with fundamental cellular processes regulating cellular development and proliferation, nuclear cellular components and binding functions (Fig. S8A). Moreover, aging, response to stress, and homeostatic process are biological processes correlated with increased DNA instability and metabolic changes, that are known to be hallmarks for cancer development and proliferation. Conversely, we found only the Gonadotropin-releasing hormone (GnRH) signalling pathway reaches significance for FDR < 0.05 (Fig. S8B); the signaling pathway activated with the secretion GnRH with the cascading activation of the epidermal growth factor (EGF) receptor and activation of mitogen-activated protein kinases (MAPKs).

Finally, we tested whether any gene harboring breast cancer SNPs was also a known cancer driver. To do that we used two curated datasets, namely the Cancer Gene Census and the OncoKB database; of the 104 identified in our study, 16 are cancer driver genes reported in the Cancer Gene Census and 12 in the OncoKB database, with 6 being tumour suppressor genes and 4 being oncogenes (Fig. S8C).

3 Supplementary Figures

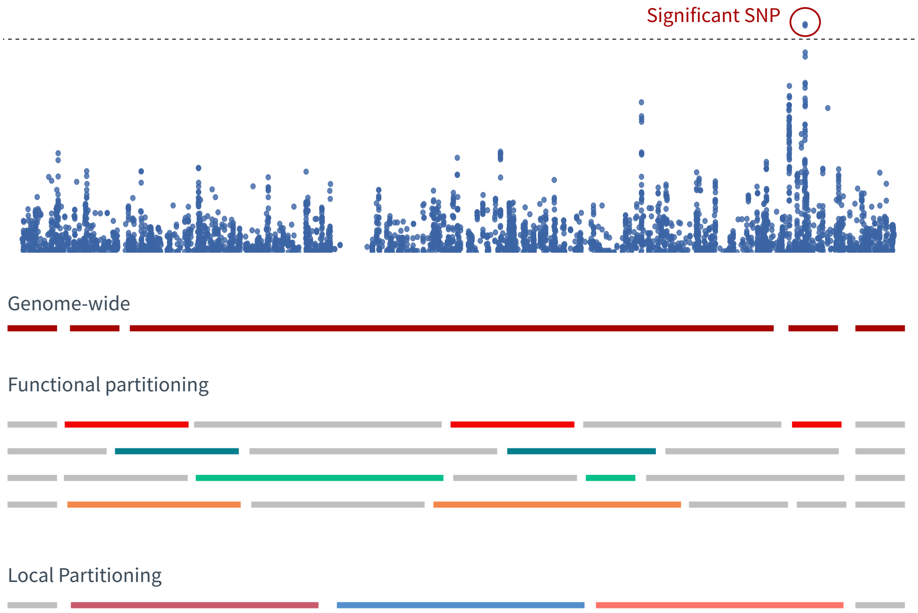


Figure S1: **Schematic representation of genome partitioning for the estimation of h^2.** This figure is providing a pictorial idea of the behaviour of different partitioning strategies. We have zoomed on a portion of a chromosome, however this extends to the whole genome. On top, we show a Manhattan plot which is a common way of showing the results of a GWAS study. In this case a single SNP is significant, with a $-log(p)$ above the threshold. Then we can see how the different methods group the SNPs together. For the estimation of the genome-wide h^2 all SNPs are considered in a single term. For the functional partitioning, all the SNPs in the genome falling into a specific category are apportioned to the same term. Eventually, for the local partitioning, the genome is divided into multiple, non overlapping, regions and each SNP then, is assigned to a local term.

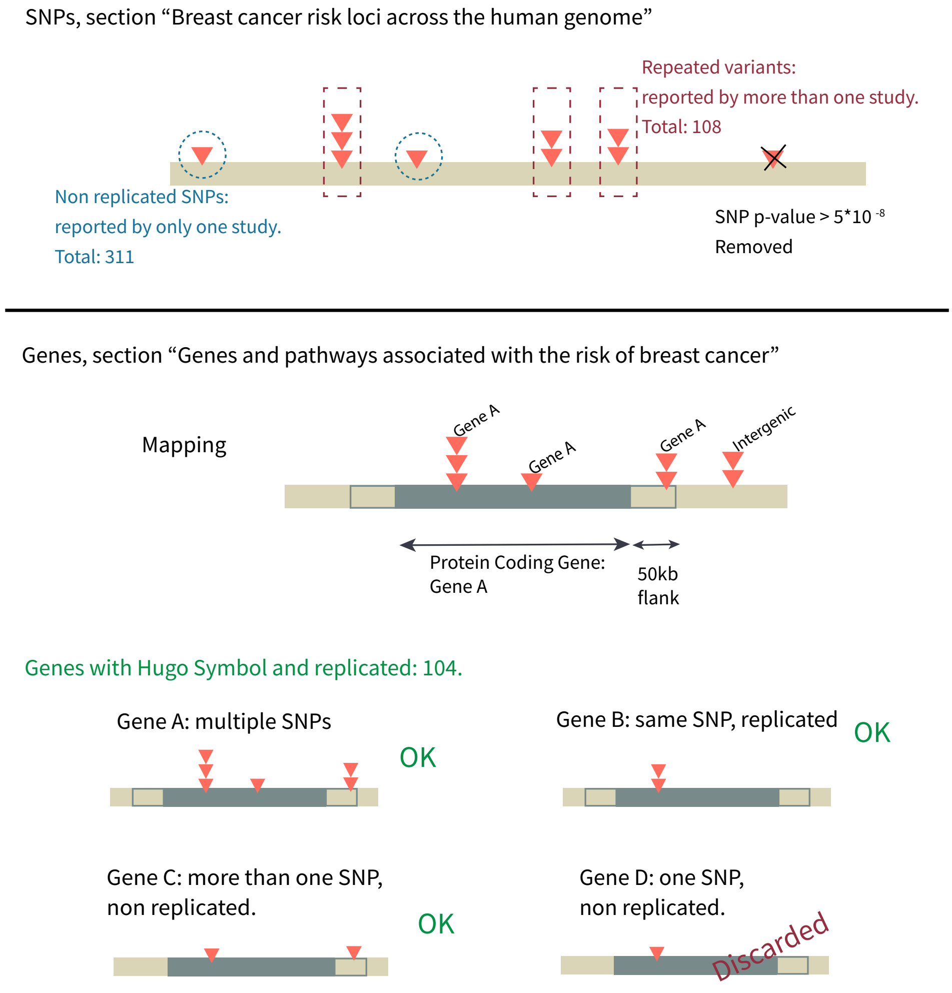

SNPs, section “Breast cancer risk loci across the human genome”
Repeated variants:
reported by more than one study.
Total: 108
Non replicated SNPs:
reported by only one study.
Total: 311
SNP p-value > 5*10 -8
Removed
Genes, section “Genes and pathways associated with the risk of breast cancer”
Mapping
Gene A
Gene A
Gene A
Intergenic
Protein Coding Gene:
Gene A
50kb
flank
Genes with Hugo Symbol and replicated: 104.
Gene A: multiple SNPs
OK
Gene B: same SNP, replicated
OK
Gene C: more than one SNP,
non replicated.
OK
Gene D: one SNP,
non replicated.
Discarded

Figure S2: **Schematic representation of the SNP filtering and curation workflow.**

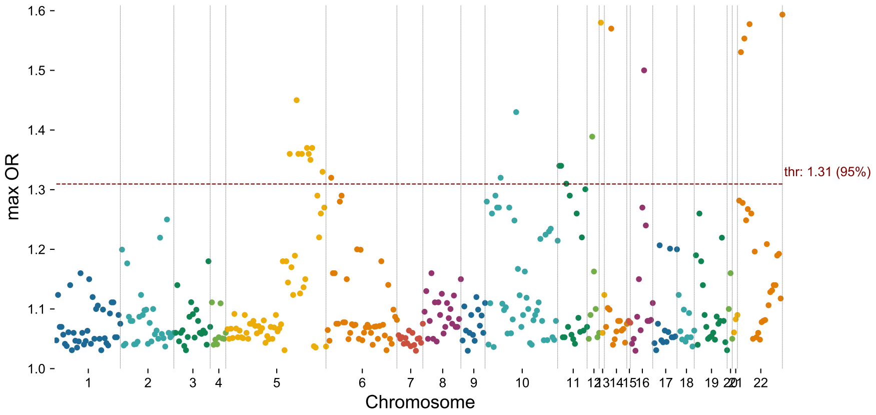


Figure S3: **Genome-wide distribution of breast cancer SNPs.** For each SNP in our dataset, we plot on the y-axis the maximum reported odds ratio (OR). It is important to note that only 5% of all SNPs have $OR > 1.31$

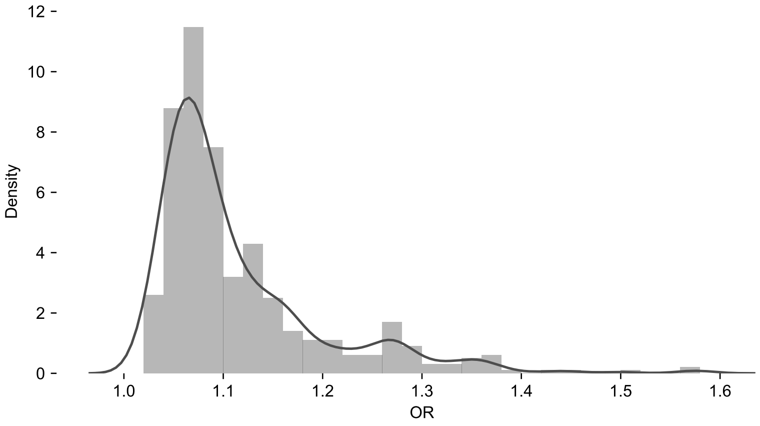


Figure S4: **Distribution of the odds ratio (OR) for breast cancer SNPs.**

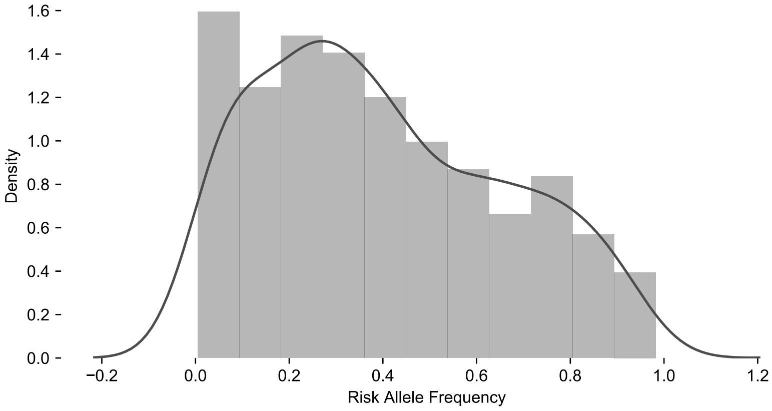


Figure S5: **Distribution of the risk allele frequency for breast cancer SNPs.**

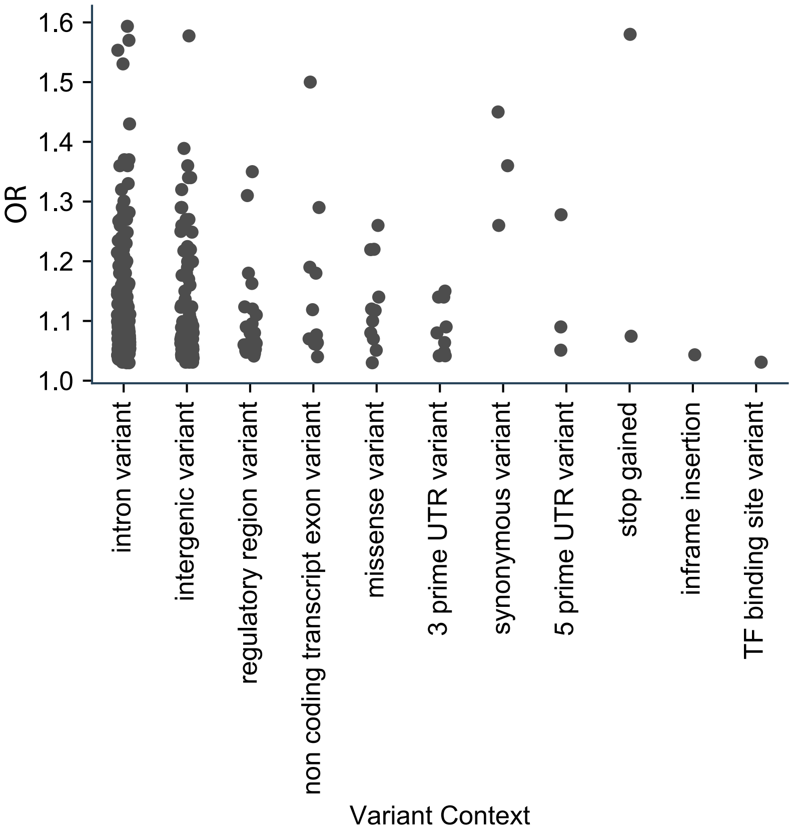


Figure S6: **Odds ratio (OR) distribution by variant type for breast cancer risk.** We report the maximum OR reported for each variant, or the one obtained by transforming the regression coefficient, β, as follows: $OR = exp(\beta)$.

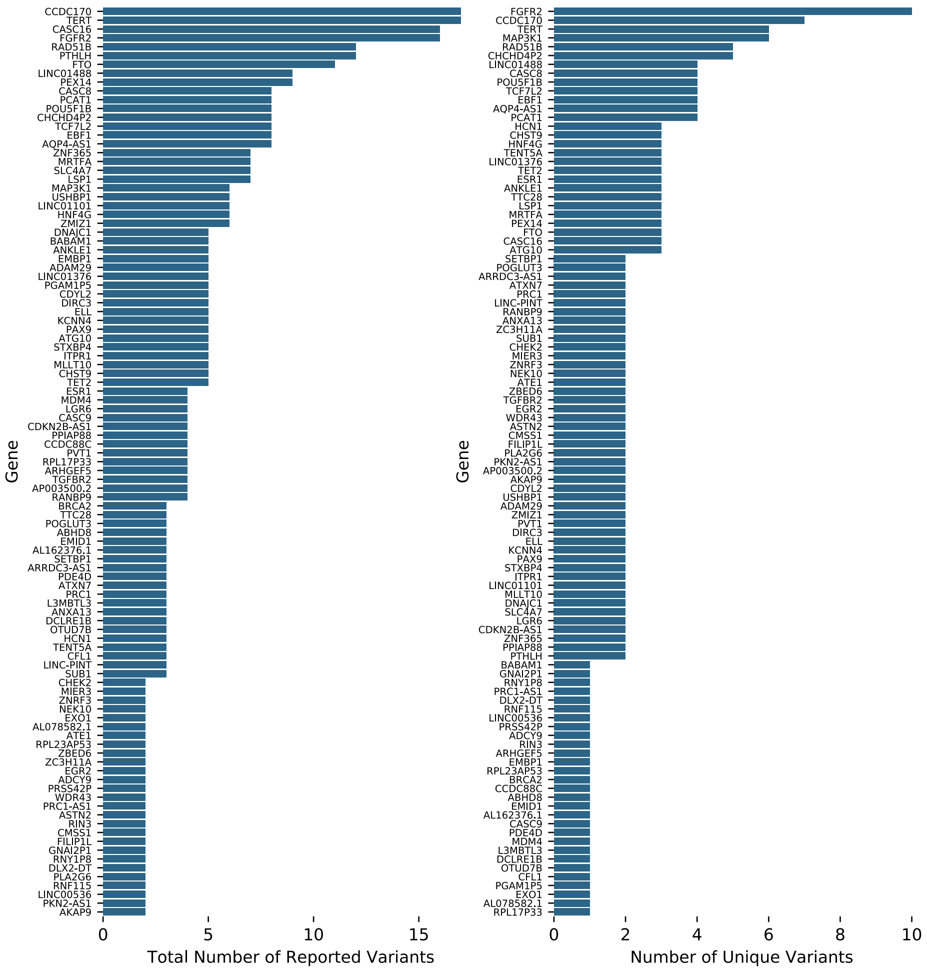


Figure S7: **Number of times a gene is reported in the catalog** On the left side, we show the number of variants reported by the GWAS catalog for each gene. We consider only genes with at least two occurrences in the catalog. On the right side, we report the number of unique variants for each gene.

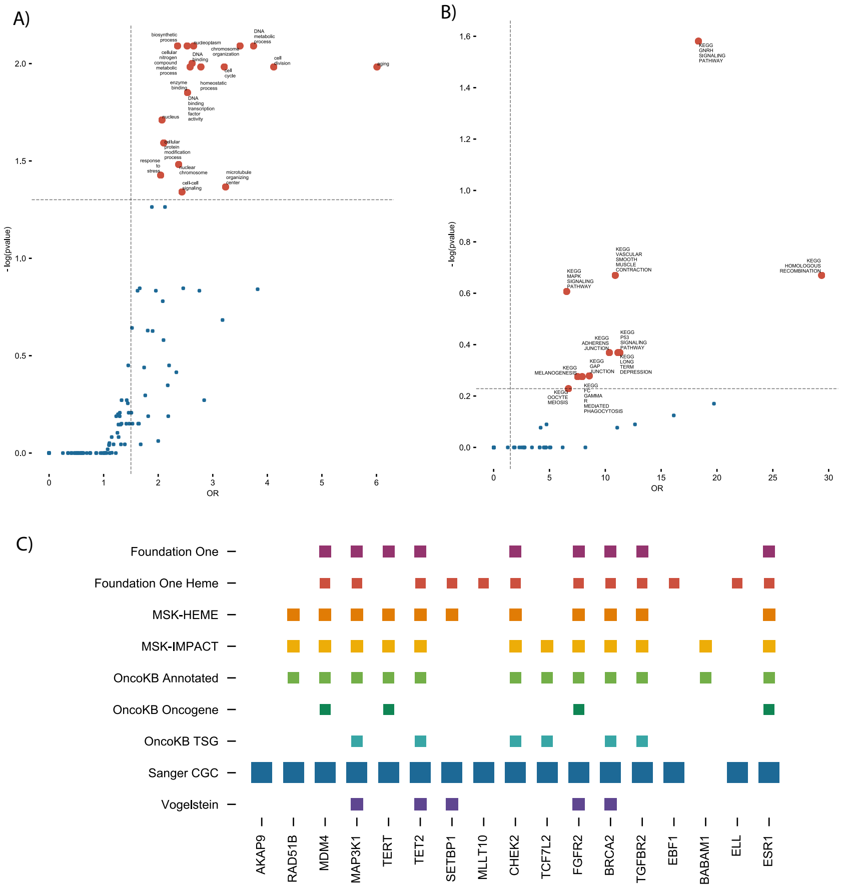


Figure S8: **Biological Characterisation** A) Pathways analysis with the principal gene ontologies. Term in red are significant. B) Pathways analysis with KEGG pathways. Only the top term (GnRH) is significant, whereas we report in red the top 10 terms regardless of the statistical significance level. C) Overlaps between all the genes and those reported in the major cancer driver genes annotations. The box size is proportional to the overlap between an annotation and the genes found in our analysis (see Supplementary Tables Pathways).

A.2 Supplementary materials for “The landscape of the heritable cancer genome”

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 η_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 η_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 $\ddot{h}_k^2 = \sum_{j=1}^{M_k} \hbar_j^2$, where $\ddot{h}_k^2$ is the output of HESS. With the notation we introduced in our study, $\ddot{h}_k^2/M_k = h_k^2/M$, where h_k^2 is the gene-level heritability estimated by BAGHERA. Both methods, however, test whether the local single SNP heritability, either h_k^2/M or $\ddot{h}_k^2/M_k$, is larger than the expected genome-wide heritability $\hbar_M^2$.

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, η, 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 ρ) and the corresponding p-value (see table below).

	HESS		BAGHERA				
ICD10	$h^2(se)$	Significant loci	$h^2(sd)$	Significant loci	Common loci	ρ	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 $\ddot{h}_k^2$ 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 $\ddot{h}_k^2$ estimates to BAGHERA's h_k^2, as $\ddot{h}_k^2/M_k \times M$. 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 η 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

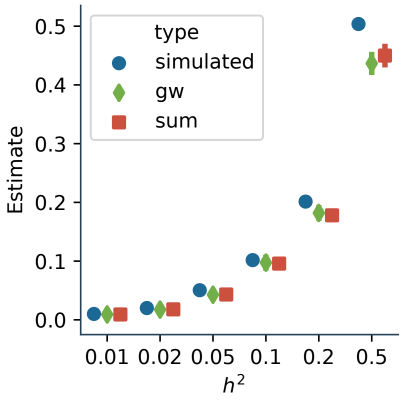
The UK Biobank provides GWAS results for multiple malignancies classified by patient self-reported cancer type at time of assessment. Here we show the results for this dataset using summary statistics computed by B. Neale et al. We found only 11 datasets with $\hat{\chi^2} > 1.01$ compared to the 17 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).

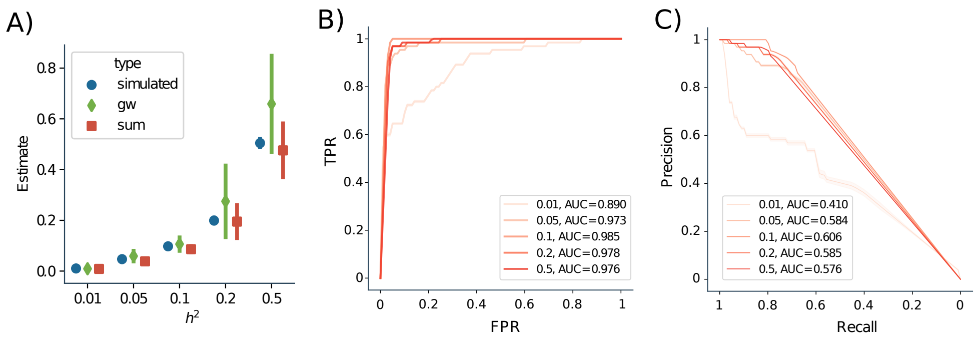
Overall, we find that quantitatively comparing the heritability loci results for self-reported and histologically classified cancers might be difficult. We then considered the Jaccard similarity coefficient computed between heritability genes for each pair of cancers (see Supplementary Figure 14). Here 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.

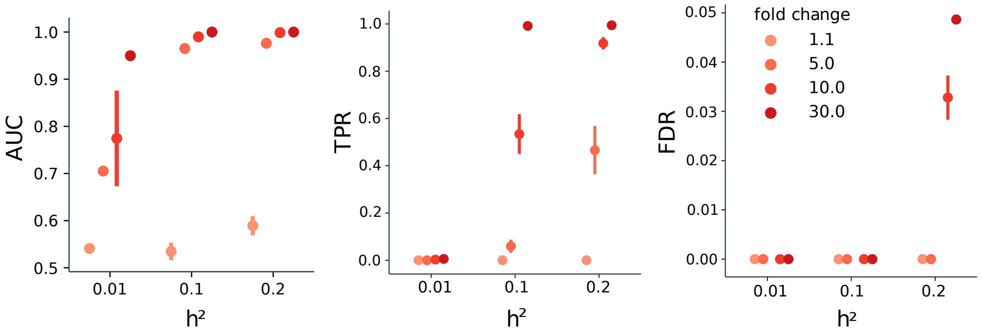
2 Supplementary Figures



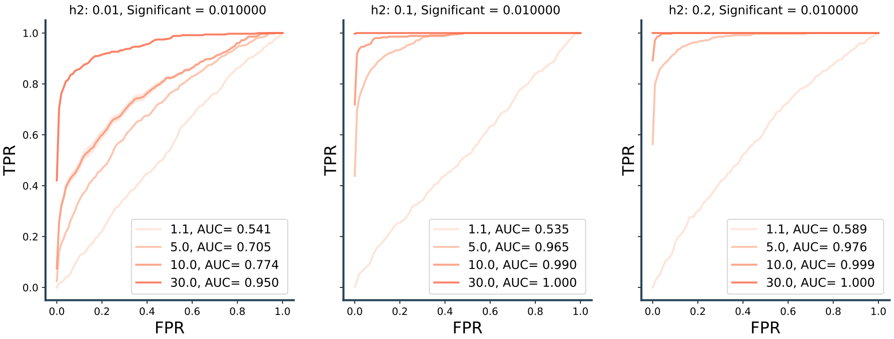
Supplementary Figure 1: **Performance on genome-wide heritability estimation for simulated dense effect datasets.** Genome-wide heritability estimates for dense effects. For each value of h^2, we plot the simulated heritability level, the genome-wide (gw) estimate, which is the median of the posterior 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.



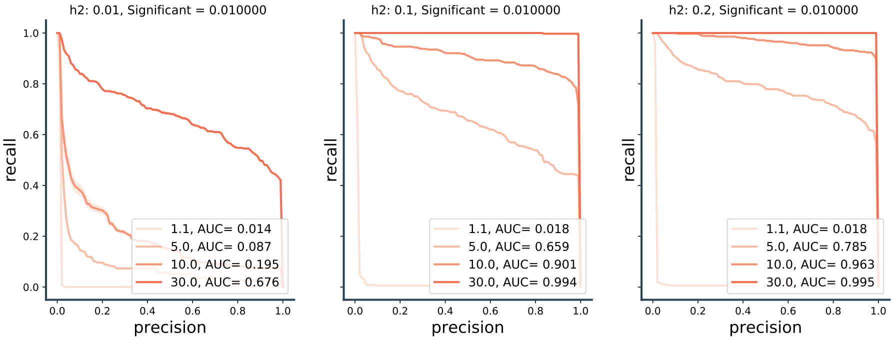
Supplementary Figure 2: **Performance on gene-level heritability estimation for simulated datasets.** A) Genome-wide heritability estimates for datasets with varying gene-level heritability. For each value of h^2, we plot the simulated heritability level, the genome-wide (gw) estimate, which is the median 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.



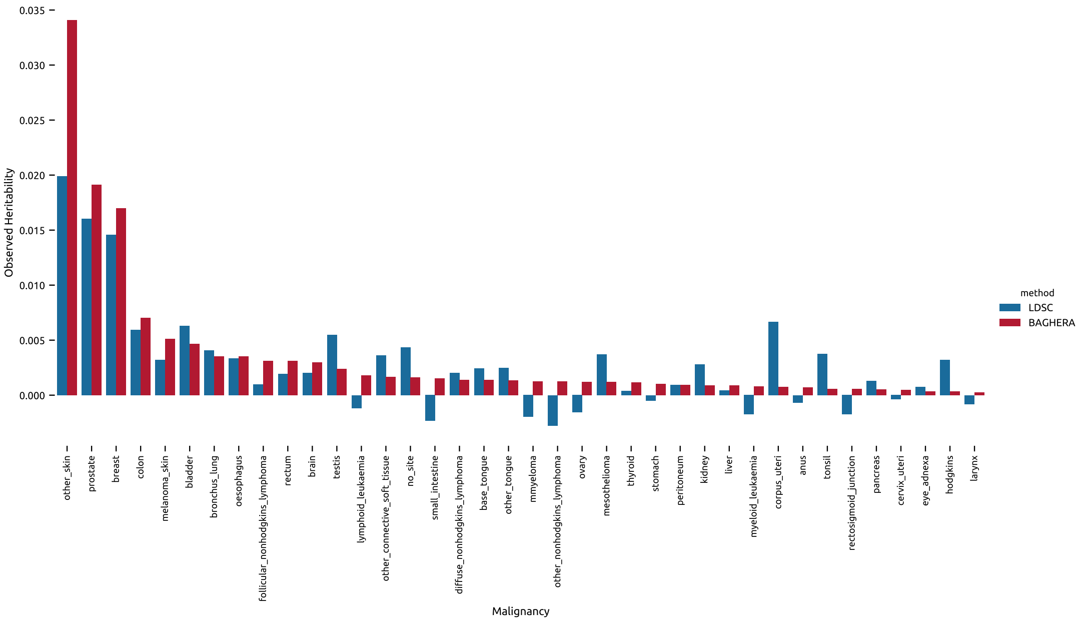
Supplementary Figure 3: **Performance on whole-genome simulated data.** Performance of BAGHERA for different levels of heritability h^2 (x-axes) and gene-level heritability enrichment (color coded). Here we show the AUCs of the ROC curves, the True Positive Rate (TPR) and False Discovery Rate (FDR) for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



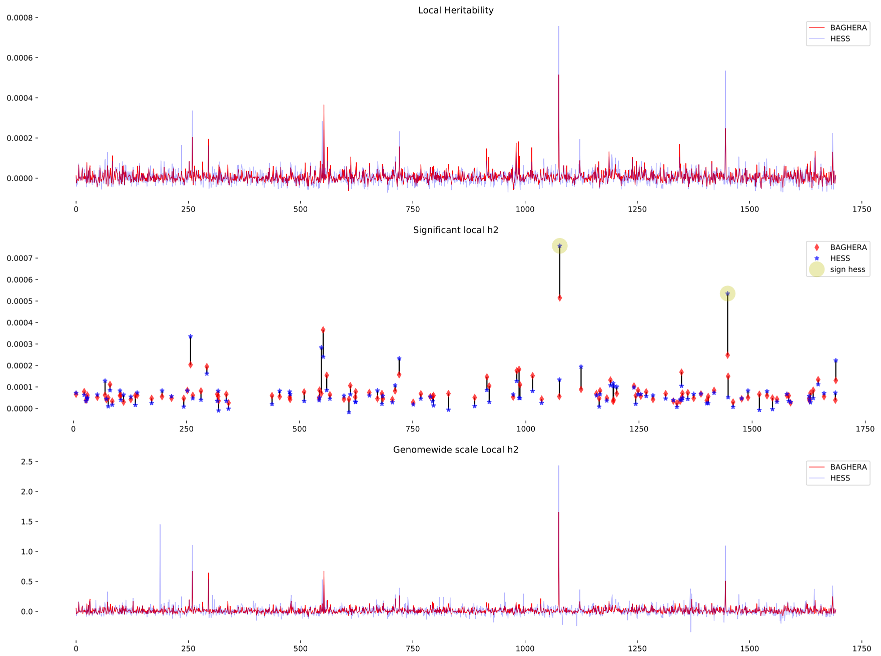
Supplementary Figure 4: **ROC curves for summary statistics simulations.** Receiver Operating Characteristic curve for data simulated from summary statistics. Fold changes,$f_c = \{1.1, 5, 10, 30\}$, are color-coded, while each column corresponds to different values of $h^2 = \{0.01, 0.1, 0.2\}$.



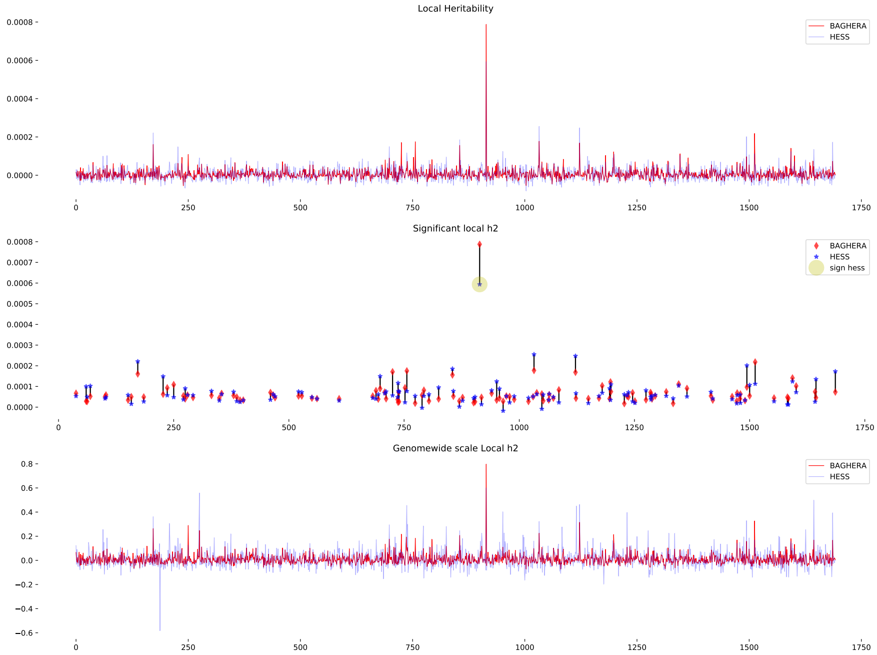
Supplementary Figure 5: **PR curves for summary statistics simulations.** Precision Recall curves for the data simulated from summary statistics. Fold changes,$f_c = \{1.1, 5, 10, 30\}$, are color-coded, while each column corresponds to different values of $h^2 = \{0.01, 0.1, 0.2\}$.



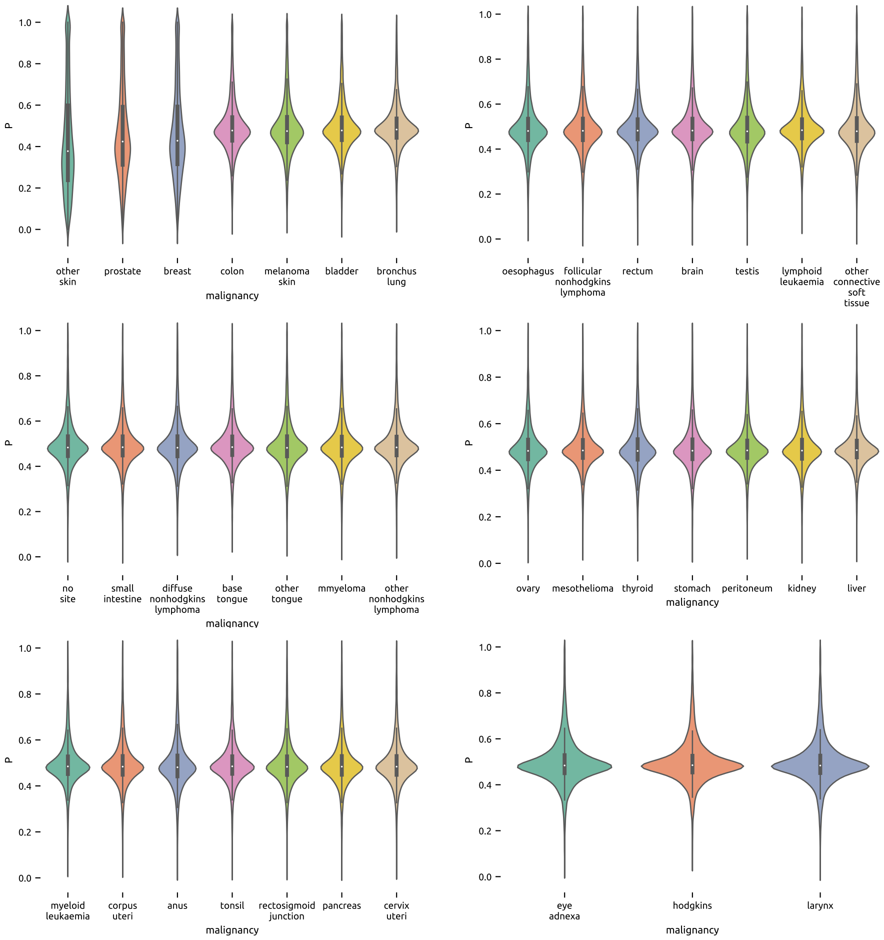
Supplementary Figure 6: **Comparison between LDSC and BAGHERA heritability estimates.** For each of the 38 malignancies (x-axis), we show the observed h^2 estimate (y-axis) for LDSC (blue) and BAGHERA (red).



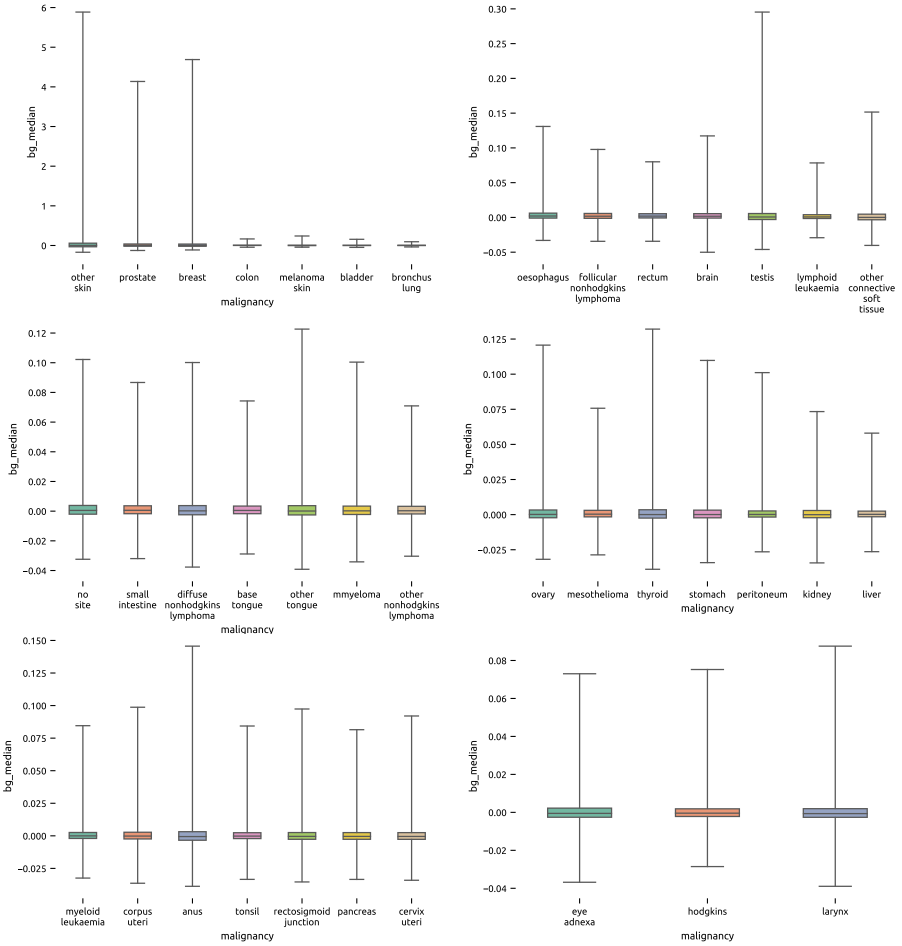
Supplementary Figure 7: **Comparison between BAGHERA and HESS local heritability estimates for breast cancer (C50).** The first panel shows HESS and BAGHERA values of local heritability $\ddot{h}_k^2$. The second panel reports the values of $\ddot{h}_k^2$, but it is limited the regions that are reported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



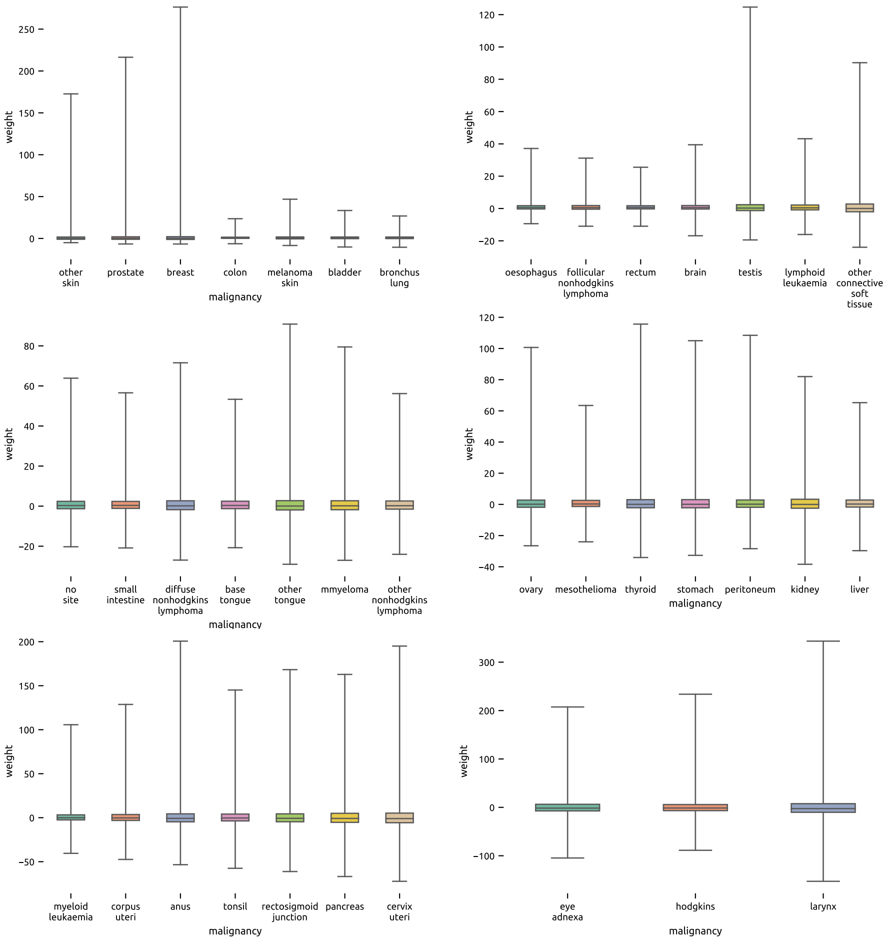
Supplementary Figure 8: **Comparison between BAGHERA and HESS local heritability estimates for prostate cancer (C61).** The first panel shows HESS and BAGHERA values of local heritability $\ddot{h}_k^2$. The second panel reports the values of $\ddot{h}_k^2$, but it is limited the regions that are deemed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



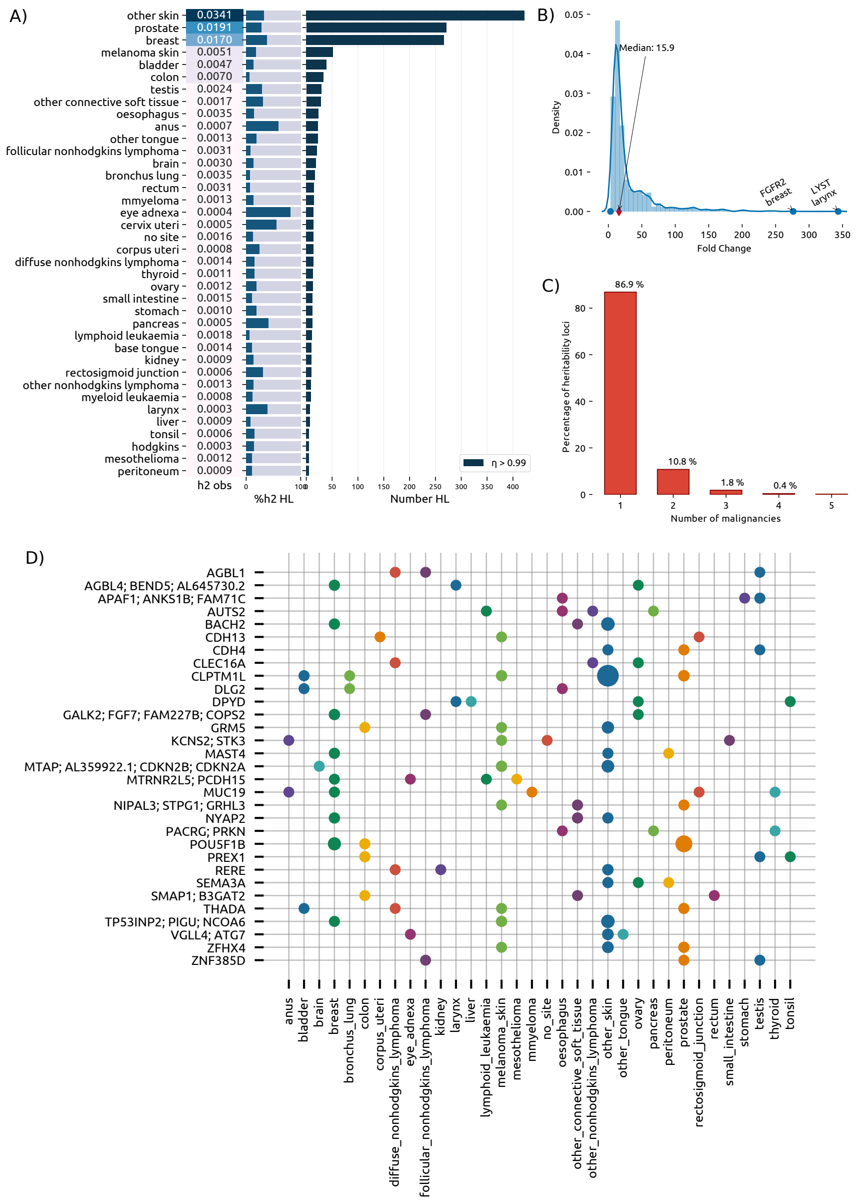
Supplementary Figure 9: **BAGHERA results - η distribution across 38 cancers in the UK Biobank.** For each dataset (x-axes), a violin plot shows the mass distribution of the indicator function η, which in the software implementation is named P (y-axes).



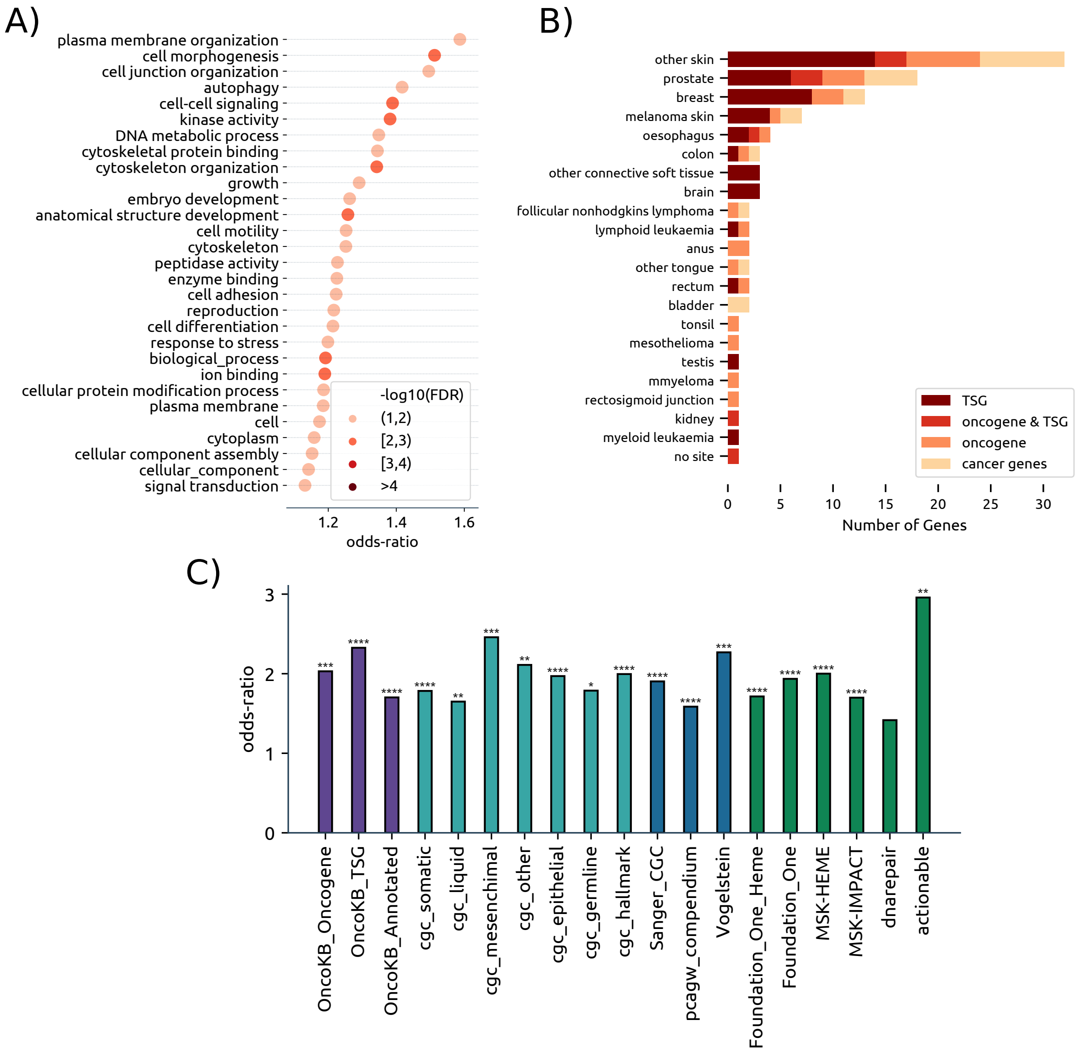
Supplementary Figure 10: **BAGHERA results - local heritability distribution across 38 cancers in the UK Biobank.** For each dataset (x-axes), we show the boxplot of the median h_k^2 for each gene, which in the software implementation is named bg median (y-axes).



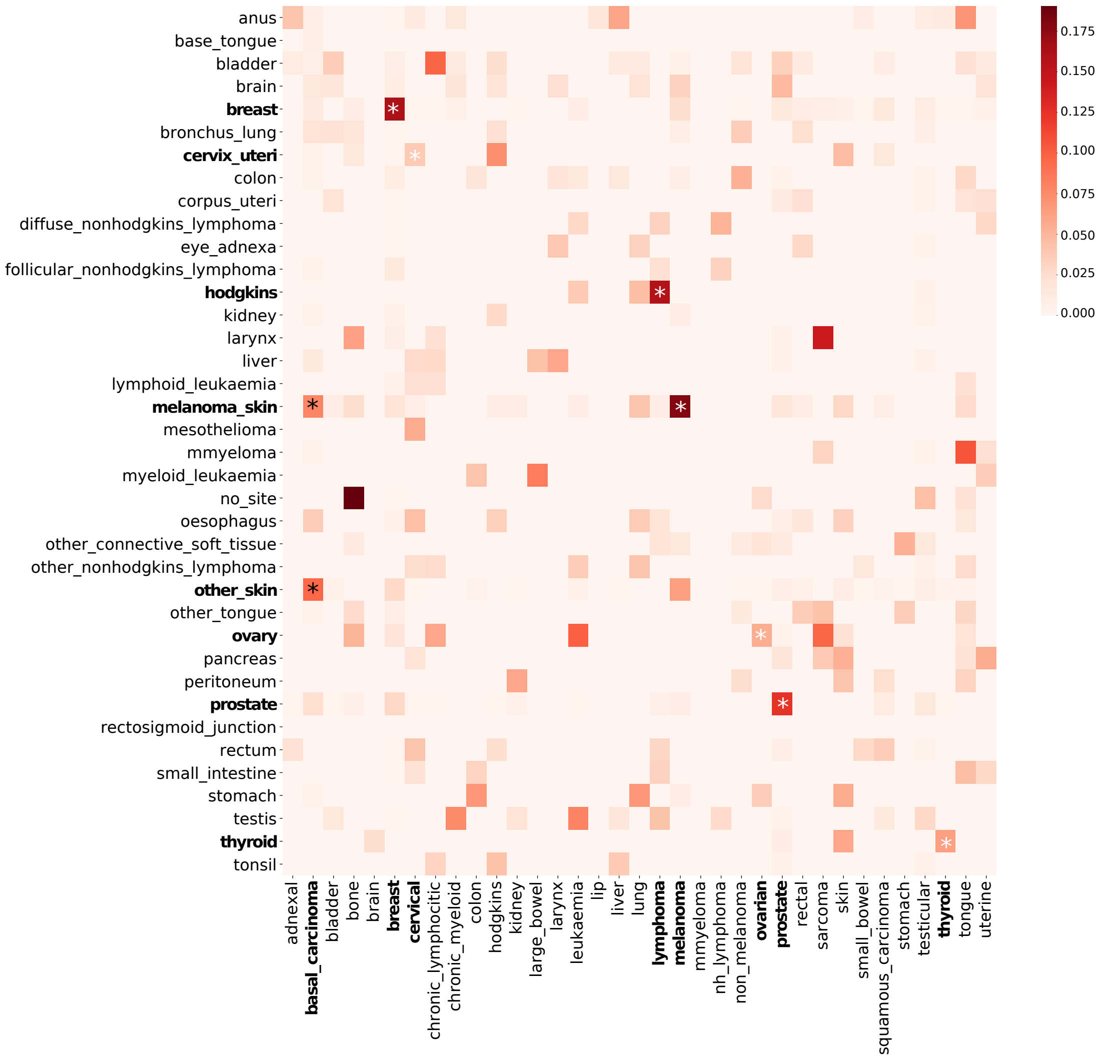
Supplementary Figure 11: **BAGHERA results overview: local heritability weights across** 38 **cancers in the UK Biobank**. For each analysed dataset (x-axes), we show the boxplot of the local heritability weights $w_k = (h_k^2 - h^2)/h^2$ for each gene. Please note that the fold change has the following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



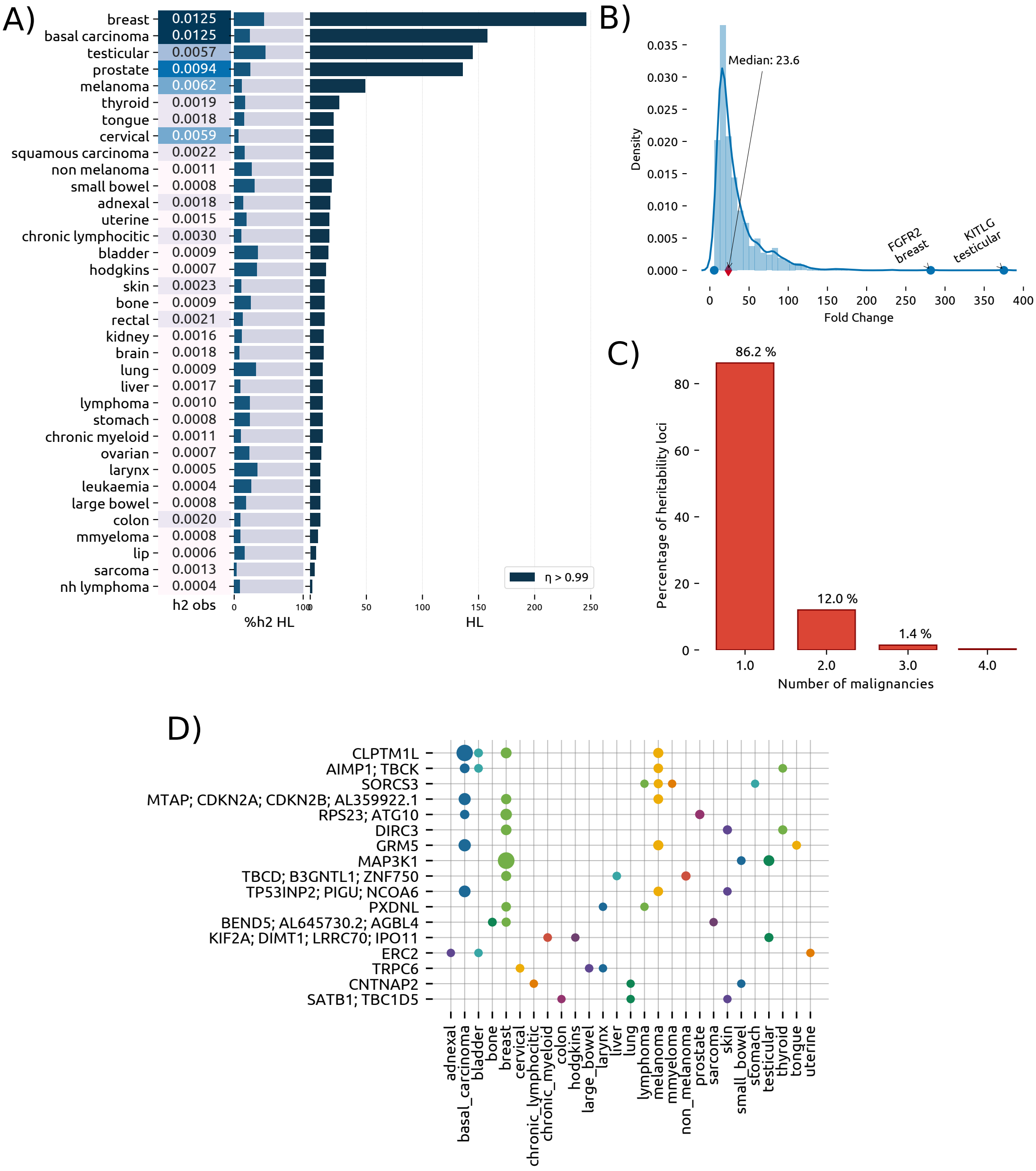
Supplementary Figure 12: **Heritability loci across 38 cancers in the UK Biobank.** A) For each malignancy we report the observed heritability (h^2_{SNP}, left box), the percentage of h^2_{SNP} explained by heritability loci (central barplot, dark blue is the percentage explained by HLs) and the number 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.



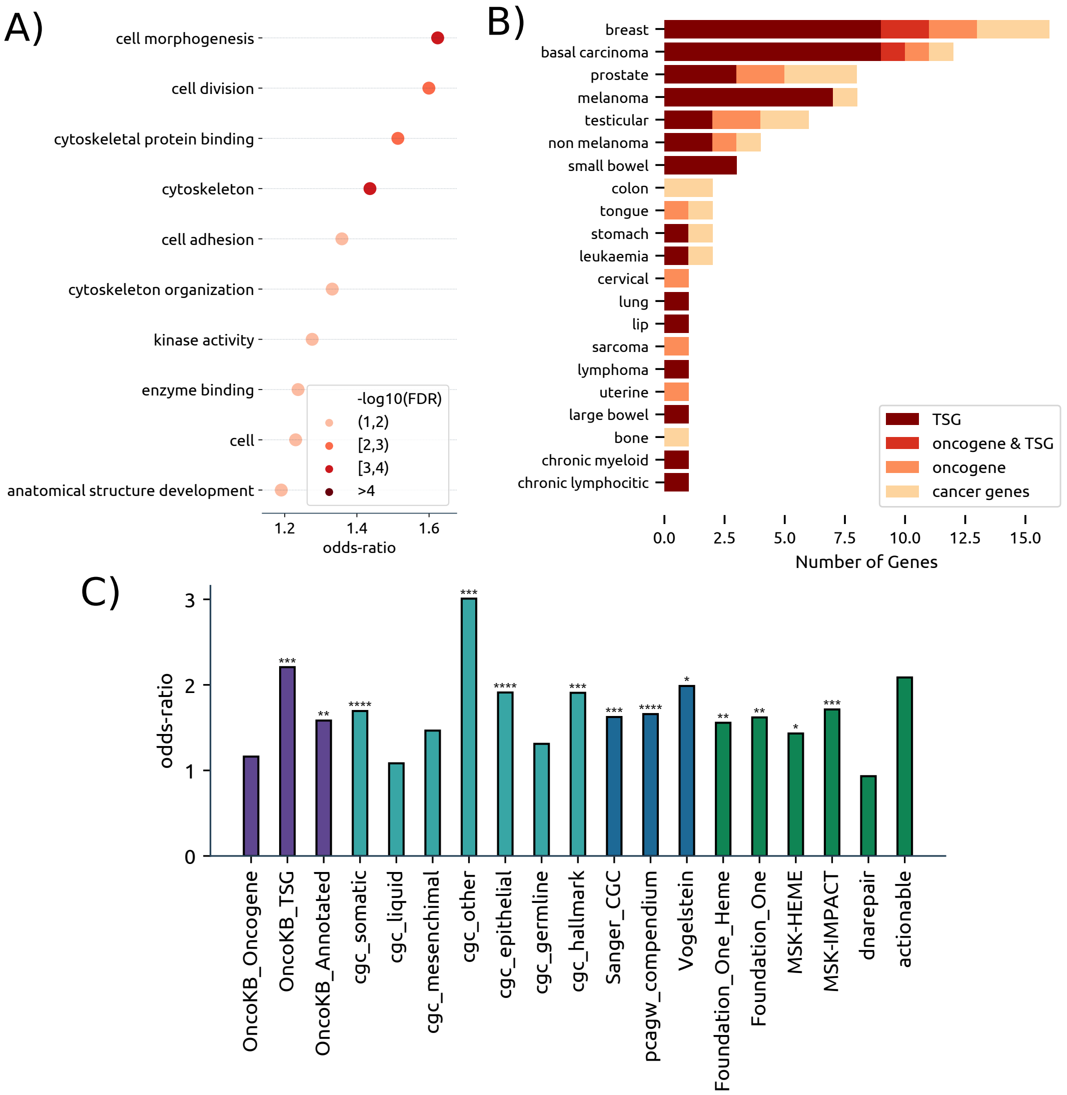
Supplementary Figure 13: **Functional characterization of cancer heritability genes across 38 cancers in the UK Biobank**. A) Gene Ontology enrichment analysis using Fisher's exact test. For each significant term, we report the odds-ratio (x-axis) and $-log_{10}$(FDR) (color gradients). B)Tumour suppressor 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 $p < 10^{-4}$.



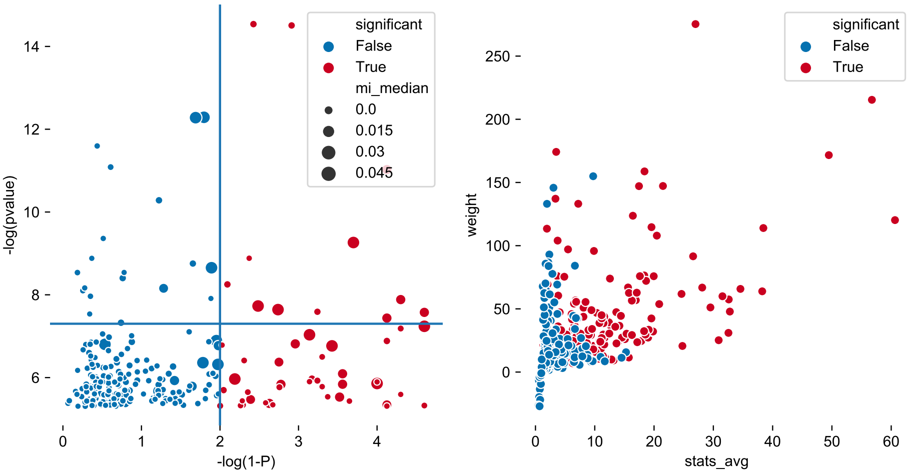
Supplementary Figure 14: **Jaccard similarity coefficient of heritability loci obtained from the 38 ICD10-classified datasets and the 35 self-reported cancers in the UKBB**. The heatmap shows the Jaccard similarity coefficient between significant genes of the histologically characterized dataset, y-axis, 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.



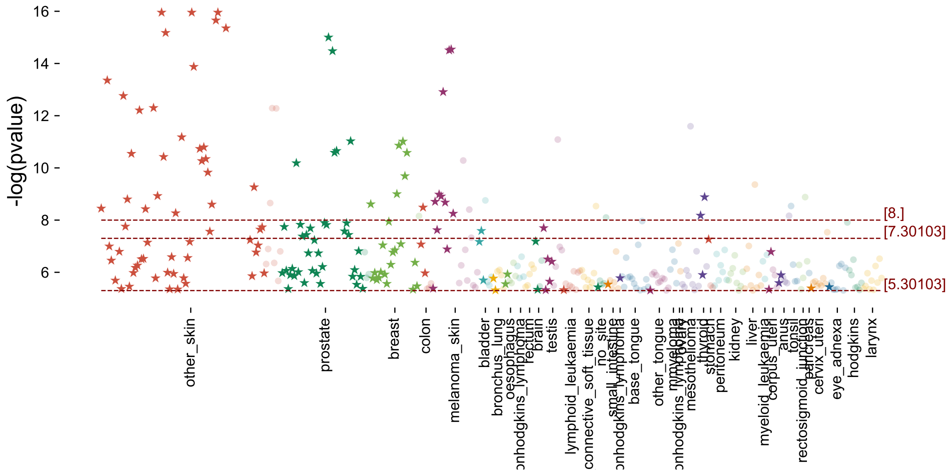
Supplementary Figure 15: **Heritability loci across 35 self-reported cancers in the UK Biobank** A) For each malignancy, we report the observed heritability (h^2_{SNP}, left box), the percentage of h^2_{SNP} explained by heritability loci (central barplot, dark blue is the percentage explained by HLs) 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 η (x-axis; in the software implementation η 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: 5×10^{-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 η 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 χ^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 η, 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 χ^2 of the SNPs considered in the GWAS analysis ($\hat{\chi}^2$), the genome-wide estimates of heritability, both on the observed (h^2_{SNP}) and the liability ($h^2_{SNP_L}$) scale, and the number of heritability loci (HL) reported by BAGHERA as significant for $\eta > 0.99$. 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/.

A.3 Supplementary materials for “PyGNA: a unified framework for geneset network analysis”

SUPPLEMENTARY MATERIALS

PyGNA: a unified framework for geneset network analysis

Viola Fanfani, Fabio Cassano and Giovanni Stracquadanio[*]

Contents

1 Supplementary materials

1.1 Parallel sampling performance

We tested how our sampler scales as a function of the number of cores allocated to PyGNA. We used the interaction network defined in [1] (13460 nodes, 138427 edges) and the generated genesets by taking random nodes from it; using a smaller network, allowed us to to minimize input/output overhead caused by reading HDF5 files. We then performed GNT analyses using both the module and random walk statistics, T_M and T_H, to test the performances when the statistic is estimated from a large matrix and when is evaluated only from the network structure. We performed our tests on genesets of size $[50, 100, 500]$ and by increasing number of cores $[1, 3, 6, 8]$ and permutations $[500, 1000, 10000]$ as shown in Fig. 1; experiments were performed on a Intel 3.2 GHz Intel Core i7 with 6 cores and 16Gb of RAM, running MacOS Mojave.

As expected, parallel sampling dramatically reduces the running time required to generate null distributions for the module test, although the maximum relative speedup was achieved when using 2 cores. For the T_H analysis, the most significant improvement was observed when running PyGNA on large genesets with more than thousands permutations. In general, for small genesets and a limited number of permutations, the cost of setting up the multiprocessing environment introduces a significant computational overhead; also, as expected, when allocating more than the number of available cores on the system, we observed no improvement or an increasing running time.

Taken together, we recommend using multiprocessing when large genesets are analysed or a large number of permutations are required to obtain a stable null distribution.

1.2 Stability of empirical null distributions

We determined experimentally the number of samples to be drawn to obtain a stable empirical null distribution for the GNT testing.

To do that, we used two real networks that we know have different densities and node degree distribution, namely the BioGRID network and smaller metabolic network reported by [1]. We then conducted our tests as follows: given a network G, we sample N_{gs} genes, which represent our tested geneset, and then apply GNT analysis with NoP number of permutations. For each scenario, we repeat the procedure R times, and record mean and standard deviation of the test statistic.

Here, we performed simulations for $N_{gs} = [50, 100, 200]$ and $NoP = [10, 100, 500, 1000]$ and $R = 10$ runs using total degree and RWR statistic for GNT testing, and RWR and shortest path for GNA testing. Experimental results show that 500 permutations are sufficient to obtain a stable null distribution, regardless of the geneset size (see Fig. 2, 3, 4, 5).

1.3 Geneset network association bootstrap procedures

We hereby explain how null distributions are generated for the geneset network association (GNA) tests, which give an estimate of the strength of interaction between two genesets, S_1, S_2. When we generate a null distribution by sampling two random genesets of size S_1 and S_2, we are performing a test under the null hypothesis of no difference between the strength of association observed for S_1 and S_2 and any two random genesets of equal size. Conversely, when one of the geneset is a Gene Ontology (GO) or pathway term T, it is advisable to be more conservative; in this case, we resample just the input geneset and keep the term T fixed, such that we perform a test under the null hypothesis that there is no difference in strength of interaction between the input geneset and any other random geneset of the same size as term T.

[*] Correspondence: giovanni.stracquadanio@ed.ac.uk
School of Biological Science, The University of Edinburgh, EH9 3BF Edinburgh, UK
Full list of author information is available at the end of the article

1.4 TCGA data retrieval and preprocessing

We downloaded six dataset from The Cancer Genome Atlas (TCGA) or the Genotype-Tissue EXpression (GTEX) repository using the TCGAbiolinks package [2]. Indeed, sometimes, TCGA data lack of control samples, in those cases we resorted to the data of the Recount2 project [3], that has reprocessed all TCGA and GTEX tissues. For each RNA-seq experiment we download the HTSeq- Counts and proceed to normalise them.

Details on the used datasets are reported in Table 1.

Then a differential expression analysis (DEA) is performed using the edgeR negative binomial generalized log-linear model and FDR correction is applied [4].

We then mapped EntrezID of significant genes to HUGO symbol using the R ORG.HS.EG.DB package, in order to have consistent nomenclatures with network data.

References

1. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., Barabási, A.L.: Uncovering disease-disease relationships through the incomplete interactome. Science **347**(6224), 841 (2015)
2. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., *et al.*: Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic acids research **44**(8), 71–71 (2015)
3. Collado-Torres, L., Nellore, A., Jaffe, A.E.: recount workflow: Accessing over 70,000 human rna-seq samples with bioconductor. F1000Research **6** (2017)
4. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)

TCGA Study	Cases	Tumor Tissue	Control Study	Controls	Control Tissue	Processing pipeline
BLCA	414	Bladder	TCGA	19	Bladder	GDC
BRCA	1102	Breast	TCGA	113	Breast	GDC
DLBC	48	Lymph nodes	GTEX	595	Blood	Recount
LAML	113	Bone marrow	GTEX	595	Blood	Recount
LUSC	502	Lung	TCGA	49	Lung	GDC
PRAD	498	Prostate	TCGA	52	Prostate	GDC

Table 1 RNA sequencing datasets used for the GNA analysis. For each dataset, we report the TCGA code, the number of cases, the tumor tissue, the study and the number of control samples, control tissue and the RNAseq processing pipeline used for quantification for both cases and controls.

2 Supplementary Figures

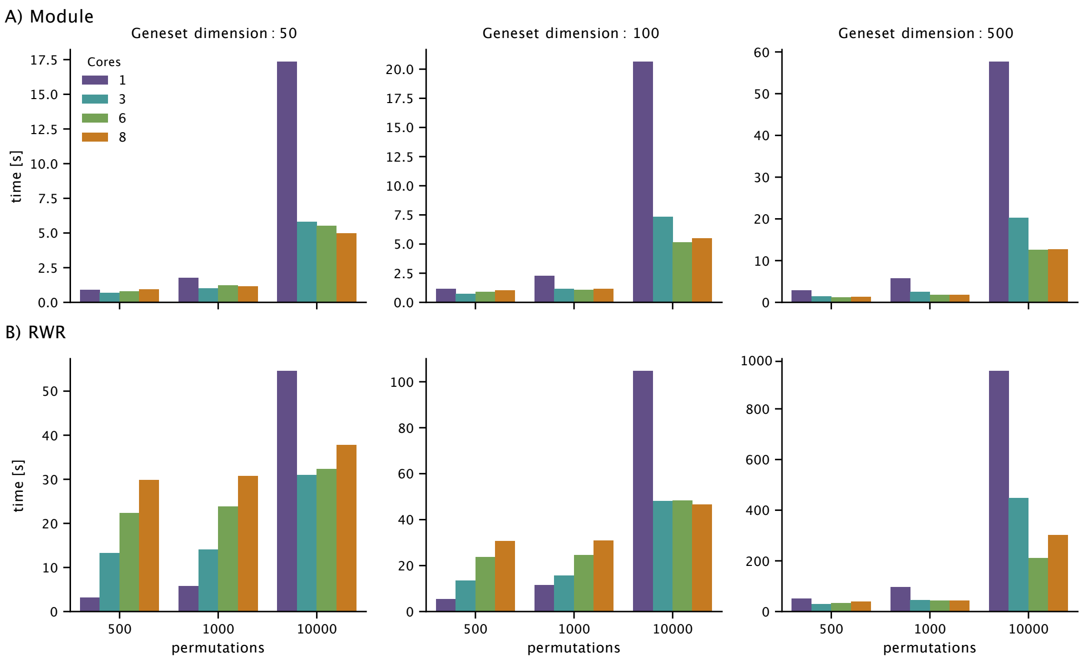


Figure 1 Running time of GNT tests using parallel sampling. We report the number of permutations on the x-axis, the average running time on the y-axis and different hues to denote different number of cores. A) Performing module GNT analysis is considerably faster on all configurations when using at least 2 cores. B) For the RWR GNT, instead, performance improvement is observed only for large genesets or large number of permutations; for few permutations and small genesets, setting up the multi-core architecture introduces a significant overhead not compensated by parallelizing the sampling process.

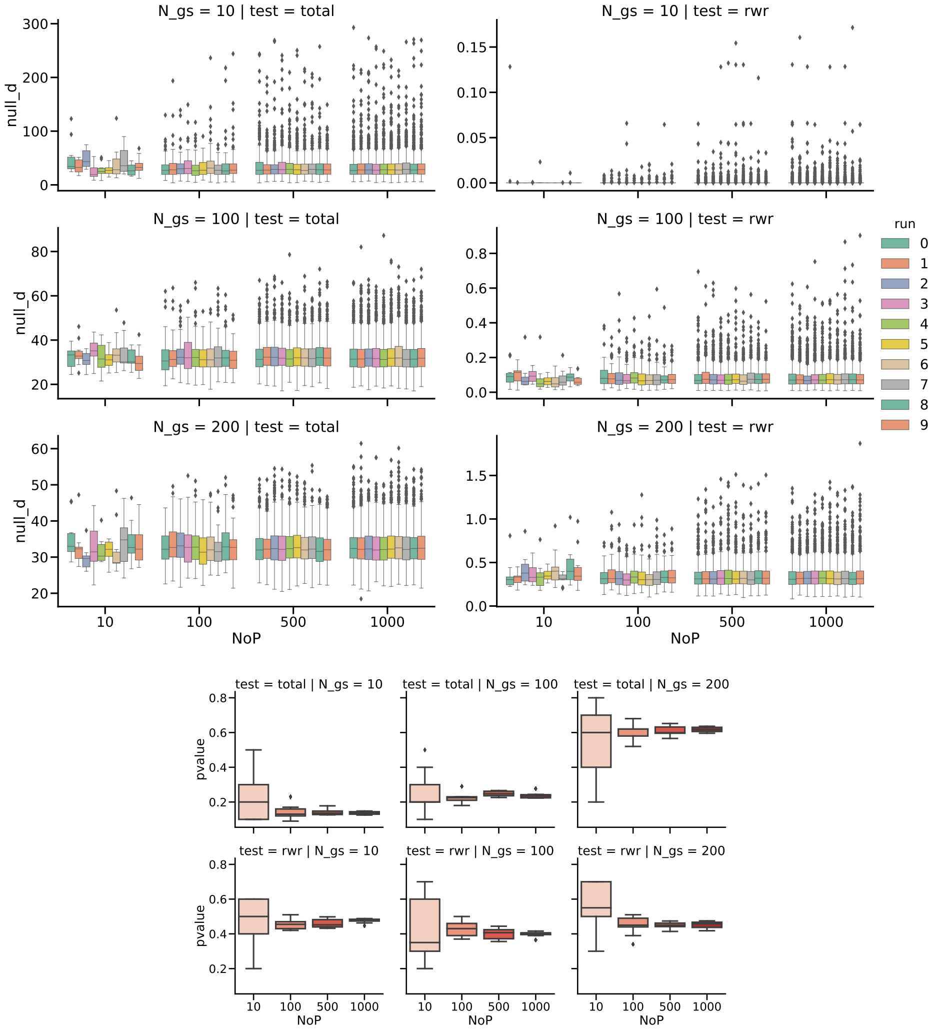


Figure 2 **Stability of empirical null distributions for GNT testing on the BioGRID network.** A) For each GNT test (columns) and each geneset size (N_gs, rows), we show the box plot of NoP samples of the the null distributions for each run. For a small number of samples, the distribution is relatively unstable, however with more than 100 samples the distributions are stabilized. B) For each GNT test (rows) and each geneset size (columns), we show the box plot of p-values for each run. As the number of permutations increases, the p-value stabilizes as well. Wider box plots reflect the fact that the same observed statistic has different significance levels, since the same geneset is tested for each run. However, for the same geneset we expect all p-values to be the same.

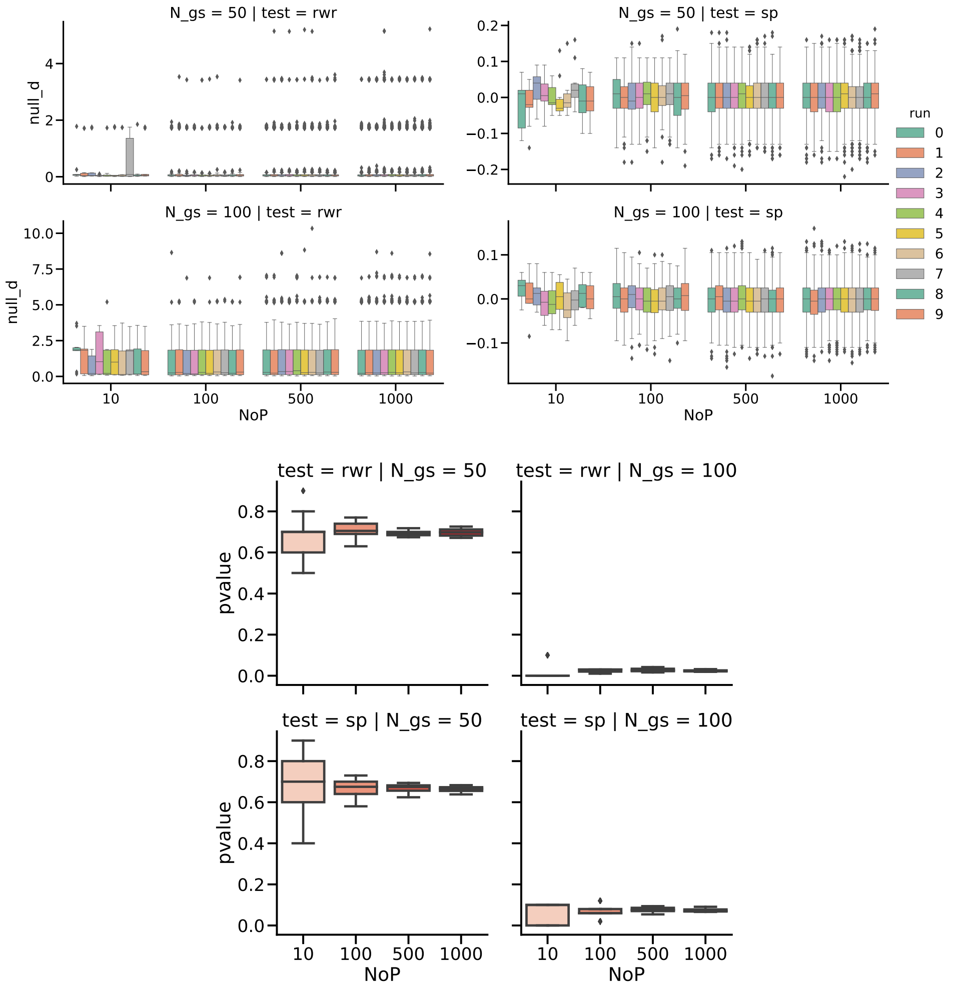


Figure 3 Stability of empirical null distributions for GNA testing on the BioGRID network. A) For each GNA test (columns) and each geneset size (N_gs, rows), we show the box plot of NoP samples of the the null distributions for each run. For a small number of samples, the distribution is relatively unstable, however with more than 100 samples the distributions are stabilized. B) For each GNT test (rows) and each geneset size (N_gs, columns), we show the box plot of p-values for each run. As the number of permutations increases, the p-value stabilizes as well. Wider box plots reflect the fact that the same observed statistic has different significance levels, since the same geneset is tested for each run. However, for the same geneset we expect all p-values to be the same.

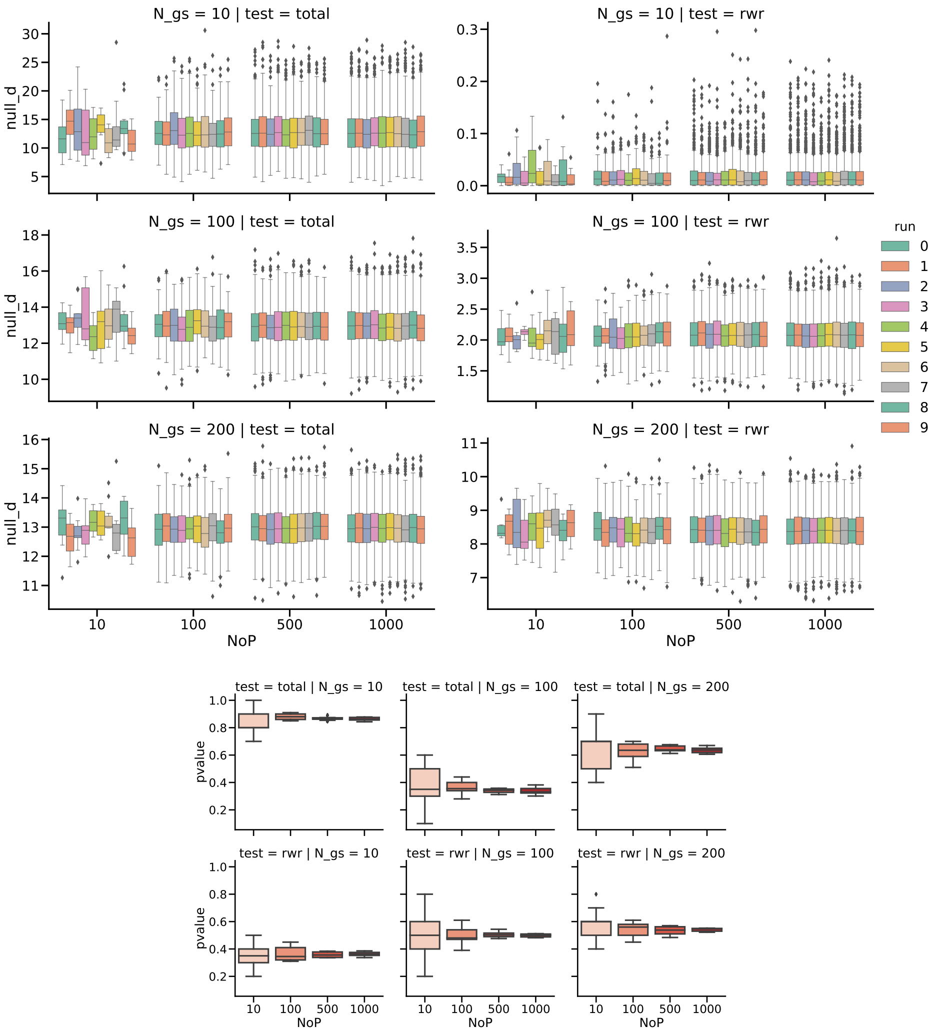


Figure 4 **Stability of empirical null distributions for GNT testing on the metabolic network.** A) For each GNT test (columns) and each geneset size (N_gs, rows), we show the box plot of NoP samples of the the null distributions for each run. For a small number of samples, the distribution is relatively unstable, however with more than 100 samples the distributions are stabilized. B) For each GNT test (rows) and each geneset size (N_gs, columns), we show the box plot of p-values for each run. As the number of permutations increases, the p-value stabilizes as well. Wider box plots reflect the fact that the same observed statistic has different significance levels, since the same geneset is tested for each run. However, for the same geneset we expect all p-values to be the same.

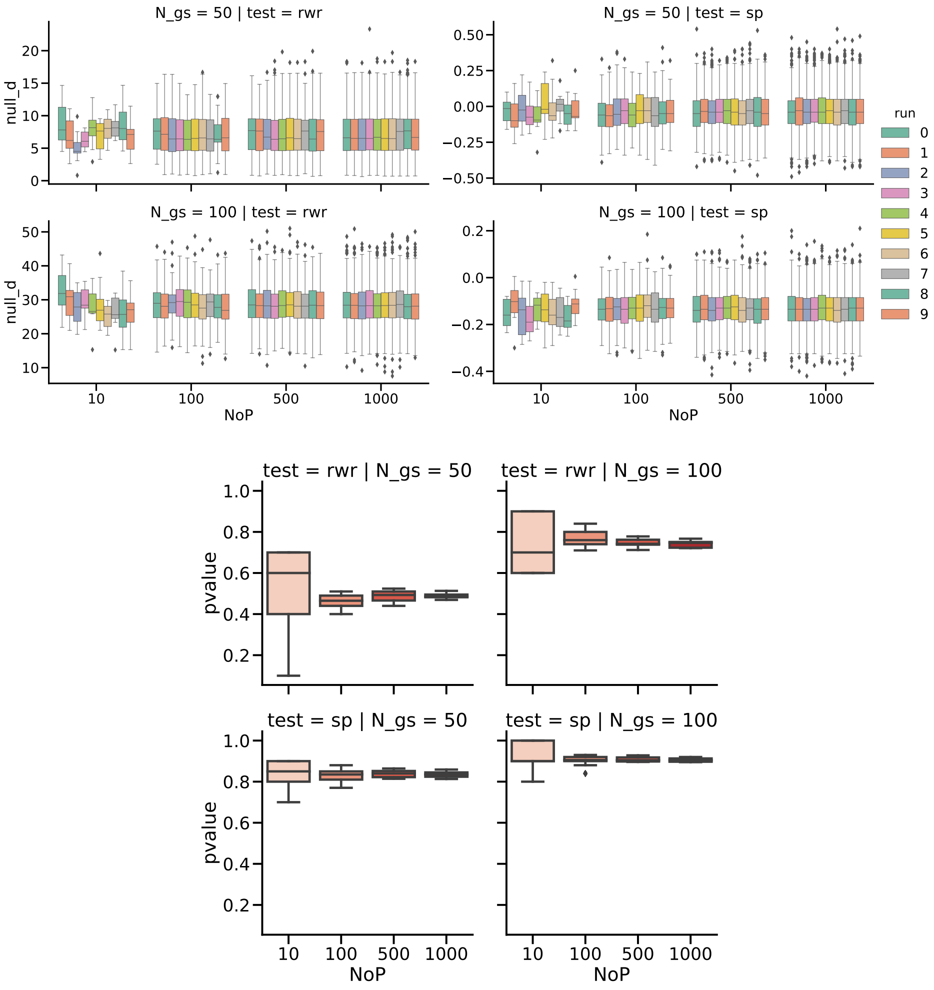
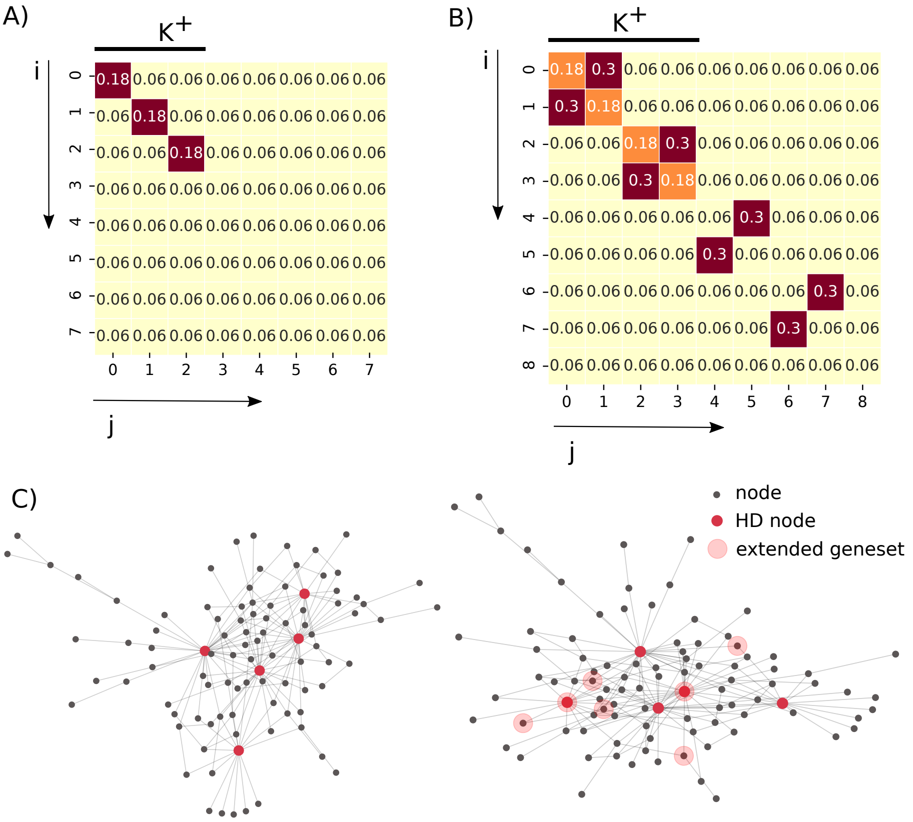


Figure 5 Stability of empirical null distributions for GNA testing on the metabolic network. A) For each GNA test (columns) and each geneset size (N_gs, rows), we show the box plot of NoP samples of the the null distributions for each run. For a small number of samples, the distribution is relatively unstable, however with more than 100 samples the distributions are stabilized. B) For each GNT test (rows) and each geneset size (N_gs, columns), we show the box plot of p-values for each run. As the number of permutations increases, the p-value stabilizes as well. Wider box plots reflect the fact that the same observed statistic has different significance levels, since the same geneset is tested for each run. However, for the same geneset we expect all p-values to be the same.



A)

i \ j	0	1	2	3	4	5	6	7
0	0.18	0.06	0.06	0.06	0.06	0.06	0.06	0.06
1	0.06	0.18	0.06	0.06	0.06	0.06	0.06	0.06
2	0.06	0.06	0.18	0.06	0.06	0.06	0.06	0.06
3	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
4	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
5	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
6	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
7	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06

B)

i \ j	0	1	2	3	4	5	6	7	8
0	0.18	0.3	0.06	0.06	0.06	0.06	0.06	0.06	0.06
1	0.3	0.18	0.06	0.06	0.06	0.06	0.06	0.06	0.06
2	0.06	0.06	0.18	0.3	0.06	0.06	0.06	0.06	0.06
3	0.06	0.06	0.3	0.18	0.06	0.06	0.06	0.06	0.06
4	0.06	0.06	0.06	0.06	0.06	0.3	0.06	0.06	0.06
5	0.06	0.06	0.06	0.06	0.3	0.06	0.06	0.06	0.06
6	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.3	0.06
7	0.06	0.06	0.06	0.06	0.06	0.06	0.3	0.06	0.06
8	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06

Figure 6 **Generation of synthetic networks** A) Example of a stochastic block model matrix for GNT testing, where each cell reports the value M_{ij}. In this case $p_0 = 0.06$, $\alpha = 3$ and $k^+ = 3$. For benchmarking, $i = 0, 1, 2$ would be considered positive examples, while $i = 4, 5, 6, 7$ would be used as negative ones. B) Example of a stochastic block model matrix for GNA testing, where each cell reports the value M_{ij}. In this case $p_0 = 0.06$, $\alpha = 3$ and $k^+ = 4$ and $\beta = 2$. For benchmarking, $\{0, 1\}, \{2, 3\}, \{4, 5\}, \{6, 7\}$ are used to generate mixture genesets. B) Example of the HDN network and geneset generation. First, a network with a number of HDNs (red dots) is created, while all the other nodes have p_0 probability of connection. Then, a geneset is created by taking at random a mixture of HDNs and background nodes (pink nodes).

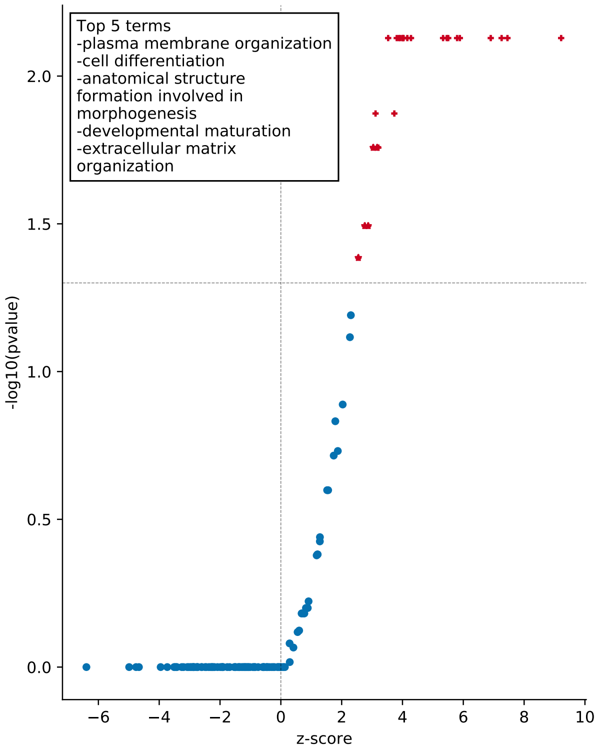


Figure 7 **Results of GNA test for the TCGA-BRCA dataset on the Gene Ontology (GO) slim dataset.** The x-axis reports the absolute value of the z-score of the GNA test statistic under a RWR interaction model, whereas the y-axis reports the $\log_{10}$ adjusted p-values for false discoveries using the Benjamini-Hochberg correction. Terms with adjusted p-value below 0.05 are reported as significant (red dots), with the top 5 marked with a star symbol.

A.4 Supplementary materials for "Discovering cancer driver genes and pathways using stochastic block model graph neural networks"

Discovering cancer genes and pathways using stochastic block model graph neural networks - Supplementary Materials

Viola Fanfani[1], Ramon Vinas Torne[2], Pietro Lio'[2], and Giovanni Stracquadanio *[1]

[1]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FD, United Kingdom

*Corresponding author. Email: giovanni.stracquadanio@ed.ac.uk

1 Methods

1.1 Simulated features

In our simulations, we studied the performance of our model using both uncorrelated and correlated gene features.

In the first case, we simulated features conditioned on the communities genes belongs to, but we did not account for correlation between different communities. Given a planted SBM structure, we simulated features for each community as random samples from a normal distribution; in the case of a hierarchical SBM, we only considered the blocks in the deepest layer. Specifically, given K communities, we first selected at random their mean from a uniform distribution $\mu_k =$ Uniform$(-5, 5)$, and we then sampled the features for each gene, $i \in k$, as $\mathbf{X}[i, :] = \mathcal{N}(\mu_k, 1)$. After obtaining the feature matrix $\mathbf{X}$, we set a fraction of genes N_{cancer} as cancer genes and multiplied their features by a weight factor W. With this strategy, features have probabilistically a distinct signal from the background.

We then used the Cholesky decomposition method to generate correlated features. This process, called features coloring, allows to impose a covariance matrix onto a stochastic process such that the final samples are correlated. Given $\mathbf{Y}$ a random variable of i.i.d. noise (uncorrelated) and $\mathbf{\Sigma}$ a covariance matrix, we want to find $\mathbf{X}$ s.t. its values are correlated conditioned on $\mathbf{\Sigma}$. Given a positive, semi-definite matrix $\mathbf{\Sigma}$, the Cholesky decomposition finds a lower triangular matrix $\mathbf{L}$ $s.t.$ $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T$. Correlated samples can then be found as:

$$\mathbf{X} = \mathbf{L}\mathbf{Y} \tag{1}$$

However, this approach requires the adjacency matrix to be positive semi-definite, which is not often the case. To overcome this limitation, we used the covariance matrix $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T$, which can be considered an edge correlation matrix.

1.2 Performance metrics

1.2.1 Blocks assignment and characterisation

One main advantage of the SBM-GNN model is the interpretability of the hidden layers. We remind that in the main manuscript we have described (Eq. 2-3) how the the membership matrix $\mathbf{Z}^{(s)}$ for the s-th SBM is learnt and then concatenated into the last layer. The block assignment matrix $\mathbf{Z}^{(s)}$, can be used to understand how the genes are assigned to different blocks, the relationship between different blocks and their characteristics.

Given the matrix $\mathbf{Z} = \mathbf{Z}^{(1)}|\mathbf{Z}^{(2)} \ldots |\mathbf{Z}^{(S)}$, each sub-matrix $\mathbf{Z}^{(s)}$ has dimensions $n \times k_s$, with n being the number of nodes and k_s the number of blocks in the s-th SBM, respectively. Each row represents the probability of the node to belong to one of the k_s blocks (soft assignment). The $\mathbf{Z}$ matrix is saved alongside the neural network parameters and can then be used to assess the community assignment performance with simulated data, when the background blocks are known, or to characterise the blocks obtained with cancer data.

Here we use a toy example to illustrate how $\mathbf{Z}$ is used in practive. Given an architecture with 2 parallel groups of respectively 2 and 4 blocks, the resulting matrix $\mathbf{Z}$ would be of the form:

	b2_0	**b2_1**	**b4_0**	**b4_1**	**b4_2**	**b4_3**
gA	0.1	**0.9**	0	0.3	**0.5**	0.2
gB	0.2	**0.8**	0	**0.9**	0.1	0
gC	0.3	**0.7**	0	0.3	**0.5**	0.2
gC	**0.8**	0.2	0	**0.9**	0.1	0
gD	**0.9**	0.2	0	0.3	**0.6**	0.1
gE	0.4	**0.6**	0	**0.6**	0.4	0
gF	**0.6**	0.4	0	0.1	**0.9**	0
gG	**0.7**	0.3	0	0.3	**0.5**	0.2
gH	0.1	**0.9**	0	**0.9**	0.1	0

where each entry is the probability that the node, rows gA,gB . . . gH, belongs to one of the blocks, b2_0, b2_1, . . . , b4_3. Here the block naming convention, e.g. b2_1, indicates both the group (SBM with two blocks) and the specific block within the group (second block). The softmax function is applied for each group (b2 or b4) such that all nodes are assigned to each parallel SBM.

Eventually, nodes are uniquely assigned to a single block for each group by picking the block with the largest memebership probability. In this example, we would obtain the assignment :

- b2_0: gC, gD, gF, gG
- b2_1: gA, gB, gE, gH ...

1.3 SBM-GNN hyperparameters

We trained our model with the following hyperparameters: learning rate:0.01, weight decay:1e-4, 16 hidden nodes (for the ϕ hidden layers), and 3 parallel SBM with 5,10,20 blocks (hidden nodes of ζ). Training was done over 15,000 epochs, with 80% of nodes in the training set and 10% in the test set. The same parameters were used with the simulated data, however we trained the model on 1,000 epochs, as they were sufficient to train SBM-GNN on a network with only 1000 genes.

1.4 UKIN and Hierarchical HotNet

We compared the performances of our model with those of two other state-of-the-art network-aware analysis methods: *using Knowledge In Networks* (uKIN) and *Hierarchical HotNet* (HHotNet).

uKIN and HHotNet are network inference methods for attributed networks, although they allow only single gene features; thus, we used Fisher's method to combine multiple datasets from PCAWG into a single score. uKIN also requires a set of known cancer driver genes to act as seeds for the guided random walks; similar to what proposed by the authors of the method, for each run (10 in total), we randomly sampled 30 cancer genes from the COSMIC dataset and used them as seeds. Testing is then done with the whole COSMIC geneset, but those employed for training, and with the PID labels, which are those deemed as cancer genes specific to this dataset. Results are presented in Fig. 5.

Conversely, Hierarchical Hotnet is a method to detect disrupted cancer subnetworks. It is then important to notice that the overall classification performance is probably worse than the one of UKIN as it is not directly designed to extract single cancer drivers but submodules of them. We ran Hierarchical Hotnet with the score randomisation strategy, 100 permutations, and extracted all the modules returned by it.

2 Figures

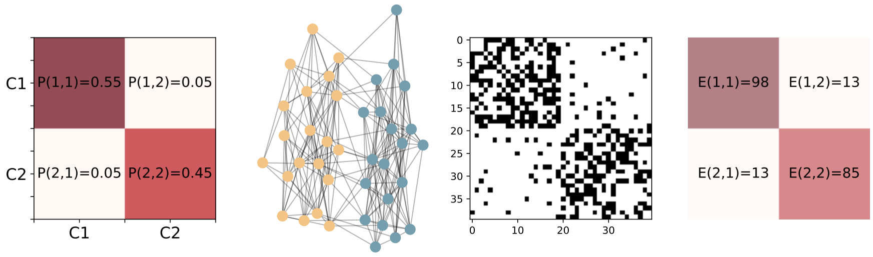


Figure 1: **Stochastic Block Model, network parameters and generation.** Here we present a simple stochastic block model of a network with two communities. We generate a network with 40 nodes, with 20 nodes assigned to community C_1, and the other 20 nodes to C_2. The probability of connection between nodes is defined by the SBM matrix shown on the right. Each element of the matrix defines the probability of connection within and between the blocks. Next to the community matrix, from left to right, we show a network generated from our SBM, the corresponding adjacency matrix, the number of observed links between each block on the left. In the network, different colors denote nodes in different communities, where the adjacency matrix is sorted by blocks. First, we can notice that a higher probability leads to more edges between the nodes; in this case we have an assortative network, where nodes within the same community are more connected than nodes between different communities.

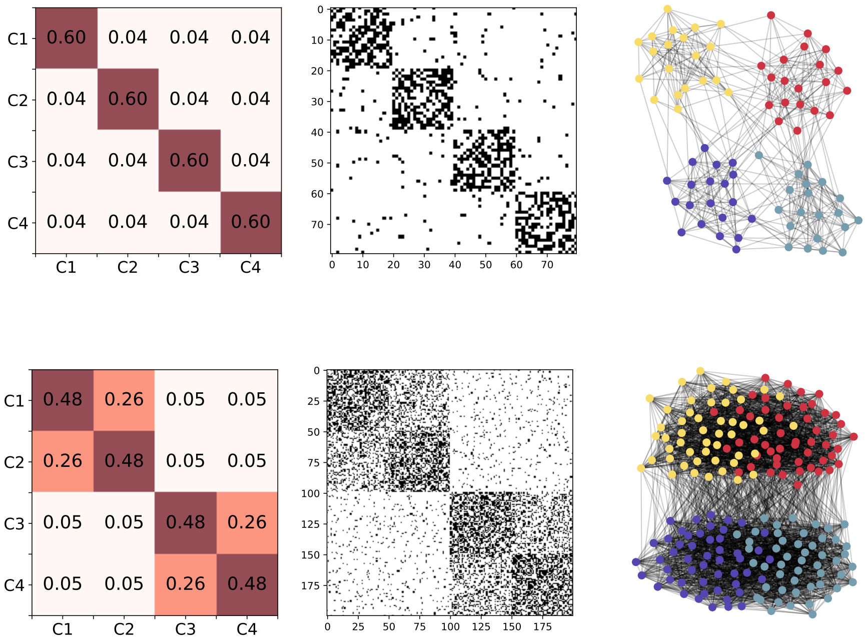


Figure 2: **SBM matrices for both disjoint and hierarchical communities** We show here two examples of how we simulate synthetic networks. First, top row, we simulate disjoint assortative communities with p_{ii} between 0.5 and 0.7 (values on the diagonal) and we add some background noise, by setting $\eta = 0.1$. On the right side we show a network generated from our SBM, with different colors to denote different communities, and the corresponding adjacency matrix. We then show a hierarchical SBM community matrix (bottom row). Blocks C1 and C2, and C3 and C4, are merged together to generate a hierarchical structure in the network. Connection probability $P_{i,j}$ is averaged across hierarchical levels.

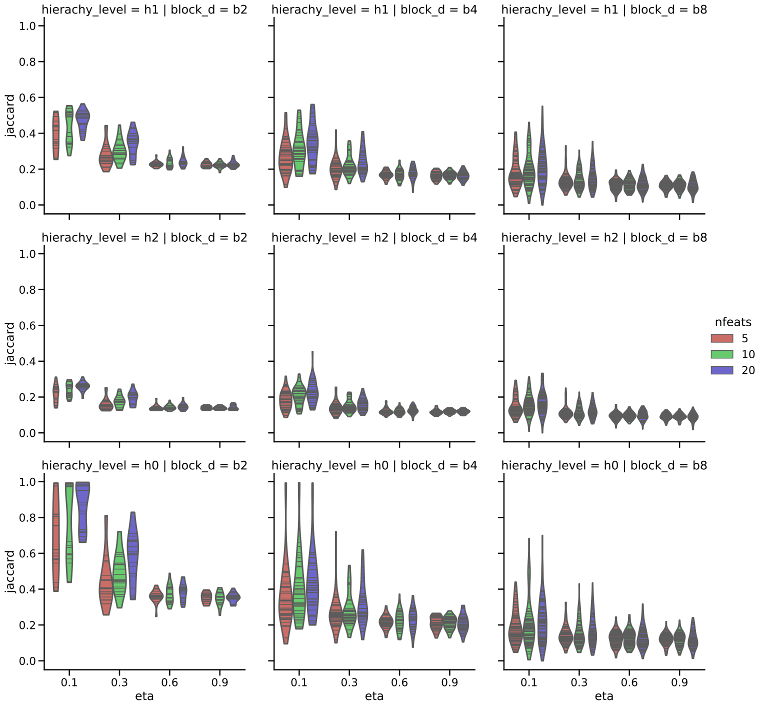


Figure 3: **Jaccard coefficient of colored features and 8 blocks**.Max Jaccard for each simulated dataset and run, for different values of η (x-axis), different number of features (color), hierarchy level (columns), block level (rows). We can notice that the detection for 2 blocks (h0 and blue dots) is good. For 4 and 8 blocks instead, it seems that SBM-GNN is not able to fully recover the fine structure, leaving some blocks empty, but properly recognising the others. We can also notice good performance for low noise (eta = 1) while expected randomness for random networks (eta = 0.9)

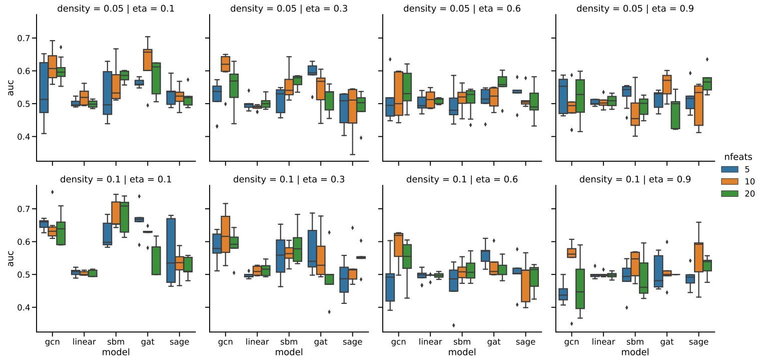


Figure 4: **Classification of colored features and 8 blocks**.AUC performance of different architectures (x-axis) for different η (columns), density (rows), number of features (colors).

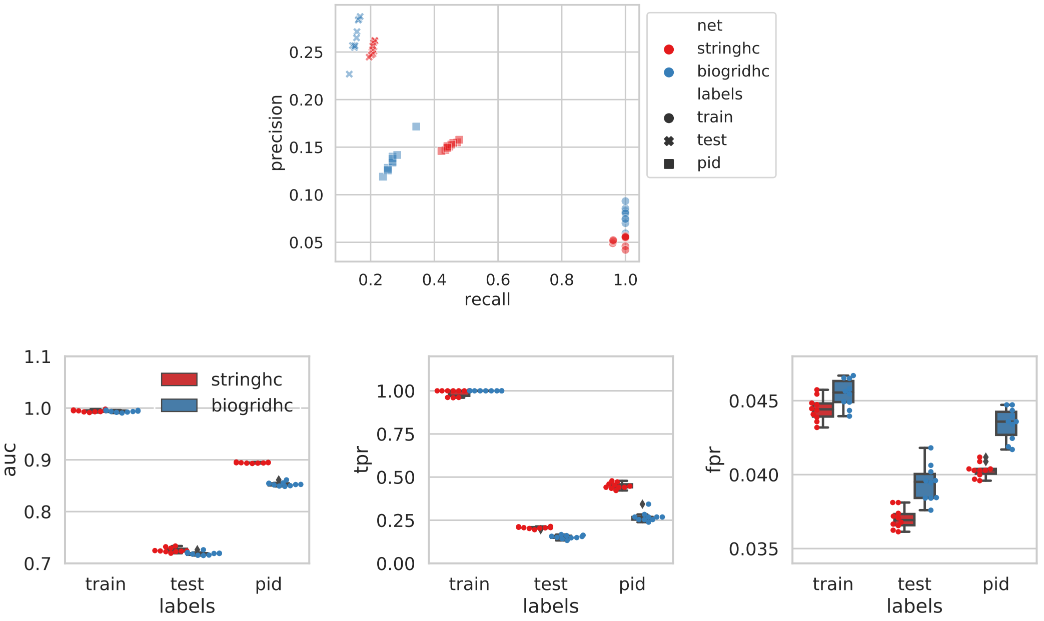


Figure 5: **UKIN performance.** A) Precision and recall for uKIN using different networks (colors) and on different labels, we cosider as significant the genes above the 90-th percentile. Train performance (round marker) is for the same genes used as seeds. The test performance (crosses) is that on the cosmic geneset, excluding the seed genes. The PID performance is on the original cancer genes of the dataset (square marker). B) AUC, TPR and FPR for the same data shown before.

B List of publications

This chapter contains a list of publications related to this thesis and the contributions to other projects.

Published

1. **Fanfani, V.**, Zatopkova, M., Harris, A.L., Pezzella, F., Stracquadanio, G., 2021. Dissecting the heritable risk of breast cancer: From statistical methods to susceptibility genes. Semin. Cancer Biol.
2. **Fanfani, V.**, Cassano, F., Stracquadanio, G., 2020. PyGNA: a unified framework for geneset network analysis. BMC Bioinformatics 21, 1–22.
3. **Fanfani, V.**, Citi, L., Harris, A.L., Pezzella, F., Stracquadanio, G., 2021. The landscape of the heritable cancer genome. Cancer Res. 81, 2588–2599.

Preprints

1. **Fanfani, V.**, Torne, R.V., Lio', P., Stracquadanio, G., 2021. Discovering cancer driver genes and pathways using stochastic block model graph neural networks. bioRxiv 2021.06.29.450342.

Contributions

1. Draberova, H., Janusova, S., Knizkova, D., Semberova, T., Pribikova, M., Ujevic, A., Harant, K., Knapkova, S., Hrdinka, M., **Fanfani, V.**, Stracquadanio, G., Drobek, A., Ruppova, K., Stepanek, O., Draber, P., 2020. Systematic analysis of the IL -17 receptor signalosome reveals a robust regulatory feedback loop. EMBO J. 39, e104202.

2. Zátopková, M., Ševčíková, T., **Fanfani, V.**, Chyra, Z., Rihova, L., Bezděková, R., Žihala, D., Growková, K., Filipova, J., Černá, L., Broskevičová, L., Kryukov, F., Minařík, J., Smejkalová, J., Maisnar, V., Harvanová, L., Pour, L., Jungova, A., Popková, T., Bago, J., Sithara, A.A., Hrdinka, M., Jelinek, T., Šimíček, M., Stracquadanio, G., and Hajek, R., 2021. Mutation landscape of multiple myeloma measurable residual disease: identification of targets for precision medicine. Blood Adv. (Accepted)

Bibliography

Abascal, Federico et al. (2021). “Somatic mutation landscapes at single-molecule resolution”. In: *Nature* 593.7859, pp. 405–410. ISSN: 14764687. DOI: 10.1038/s41586-021-03477-4.

Agrawal, Monica, Marinka Zitnik, and Jure Leskovec (2018). *Large-scale analysis of disease pathways in the human interactome*. Tech. rep. 212669, pp. 111–122. DOI: 10.1142/9789813235533_0011. arXiv: 1712.00843.

Aguet, François et al. (2020). “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369.6509, pp. 1318–1330. ISSN: 10959203. DOI: 10.1126/SCIENCE.AAZ1776.

Airoldi, Edoardo M. et al. (2009). “Mixed membership stochastic blockmodels”. In: *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference* 9, pp. 34–41. ISSN: 1532-4435. arXiv: 0705.4485.

Alexandrov, Ludmil B. et al. (2013). “Deciphering Signatures of Mutational Processes Operative in Human Cancer”. In: *Cell Reports* 3.1, pp. 246–259. ISSN: 22111247. DOI: 10.1016/j.celrep.2012.12.008.

Alpaydin, Ethem (2014). *Introduction to machine learning*. MIT press.

Alvarez, Mariano J. et al. (2018). “A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors”.

In: *Nature Genetics* 50.7, pp. 979–989. ISSN: 15461718. DOI: 10.1038/s41588-018-0138-4.

Anderson, David E. (1974). “Genetic study of breast cancer: Identification of a high risk group”. In: *Cancer* 34.4, pp. 1090–1097. ISSN: 10970142. DOI: 10.1002/1097-0142(197410)34:4<1090::AID-CNCR2820340419>3.0.CO;2-J.

Argelaguet, Ricard et al. (2018). “Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets”. In: *Molecular Systems Biology* 14.6, e8124. ISSN: 1744-4292. DOI: 10.15252/msb.20178124.

Aslam, Bilal et al. (2017). *Proteomics: Technologies and their applications*. DOI: 10.1093/chromsci/bmw167.

Avsec, Žiga et al. (2021). “Base-resolution models of transcription-factor binding reveal soft motif syntax”. In: *Nature Genetics* 53.3, pp. 354–366. ISSN: 15461718. DOI: 10.1038/s41588-021-00782-6.

Aytes, Alvaro et al. (2014). “Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy”. In: *Cancer Cell* 25.5, pp. 638–651. ISSN: 18783686. DOI: 10.1016/j.ccr.2014.03.017.

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio (2015). “Neural machine translation by jointly learning to align and translate”. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. arXiv: 1409.0473.

Bailey, Matthew H. et al. (2018). “Comprehensive Characterization of Cancer Driver Genes and Mutations”. In: *Cell* 173.2, 371–385.e18. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2018.02.060.

Baldassarre, Federico and Hossein Azizpour (2019). “Explainability Techniques for Graph Convolutional Networks”. In: arXiv: 1905.13686.

Berger, Bonnie, Jian Peng, and Mona Singh (2013). *Computational solutions for omics data*. DOI: 10.1038/nrg3433.

Bernard, Philip S. et al. (2009). “Supervised risk predictor of breast cancer based on intrinsic subtypes”. In: *Journal of Clinical Oncology* 27.8, pp. 1160–1167. ISSN: 0732183X. DOI: 10.1200/JCO.2008.18.1370.

Bersanelli, Matteo et al. (2016). “Methods for the integration of multi-omics data: Mathematical aspects”. In: *BMC Bioinformatics* 17.2, p. 15. ISSN: 14712105. DOI: 10.1186/s12859-015-0857-9.

Bothorel, Cecile et al. (2015). “Clustering attributed graphs: Models, measures and methods”. In: *Network Science* 3.3, pp. 408–444. ISSN: 20501250. DOI: 10.1017/nws.2015.9. arXiv: 1501.01676.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard (2017). “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. In: *Cell* 169.7, pp. 1177–1186. ISSN: 10974172. DOI: 10.1016/j.cell.2017.05.038.

Bruna, Joan et al. (2013). “Spectral Networks and Locally Connected Networks on Graphs”. In: *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. arXiv: 1312.6203.

Califano, Andrea et al. (2012). *Leveraging models of cell regulation and GWAS data in integrative network-based association studies*. DOI: 10.1038/ng.2355.

Campbell, Peter J. et al. (2010). “The patterns and dynamics of genomic instability in metastatic pancreatic cancer”. In: *Nature* 467.7319, pp. 1109–1113. ISSN: 00280836. DOI: 10.1038/nature09460.

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

Carbone, Michele et al. (2020). *Tumour predisposition and cancer syndromes as models to study gene–environment interactions*. DOI: 10.1038/s41568-020-0265-y.

Carter, Hannah et al. (2017). "Interaction landscape of inherited polymorphisms with somatic events in cancer". In: *Cancer Discovery* 7.4, pp. 410–423. ISSN: 21598290. DOI: 10.1158/2159-8290.CD-16-1045.

Chatr-Aryamontri, Andrew et al. (2017). "The BioGRID interaction database: 2017 update". In: *Nucleic Acids Research* 45.D1, pp. D369–D379. ISSN: 13624962. DOI: 10.1093/nar/gkw1102.

Chaudhary, Kumardeep et al. (2018). "Deep learning–based multi-omics integration robustly predicts survival in liver cancer". In: *Clinical Cancer Research* 24.6, pp. 1248–1259. ISSN: 15573265. DOI: 10.1158/1078-0432.CCR-17-0853.

Chen, Fei et al. (2019). "Analysis of heritability and genetic architecture of pancreatic cancer: A PANC4 study". In: *Cancer Epidemiology Biomarkers and Prevention* 28.7, pp. 1238–1245. ISSN: 10559965. DOI: 10.1158/1055-9965.EPI-18-1235.

Ching, Travers et al. (2018). "Opportunities and obstacles for deep learning in biology and medicine". In: *Journal of the Royal Society Interface* 15.141. ISSN: 17425662. DOI: 10.1098/rsif.2017.0387.

Cho, Ara et al. (2016). "MUFFINN: Cancer gene discovery via network analysis of somatic mutation data". en. In: *Genome Biology* 17.1. ISSN: 1474760X. DOI: 10.1186/s13059-016-0989-x.

Cieślik, Marcin and Arul M. Chinnaiyan (2018). *Cancer transcriptome profiling at the juncture of clinical translation*. DOI: 10.1038/nrg.2017.96.

Costanzo, Michael et al. (2010). "The genetic landscape of a cell". In: *Science* 327.5964, pp. 425–431. ISSN: 00368075. DOI: 10.1126/science.1180823.

Cowen, Lenore et al. (2017). “Network propagation: A universal amplifier of genetic associations”. In: *Nature Reviews Genetics* 18.9, pp. 551–562. ISSN: 14710064. DOI: 10.1038/nrg.2017.38.

Craig Venter, J. et al. (2001). “The sequence of the human genome”. In: *Science* 291.5507, pp. 1304–1351. ISSN: 00368075. DOI: 10.1126/science.1058040.

Dawson, Mark A. (2017). *The cancer epigenome: Concepts, challenges, and therapeutic opportunities*. DOI: 10.1126/science.aam7304.

Deritei, Dávid et al. (2019). “A feedback loop of conditionally stable circuits drives the cell cycle from checkpoint to checkpoint”. In: *Scientific Reports* 9.1, pp. 1–19. ISSN: 20452322. DOI: 10.1038/s41598-019-52725-1.

Di Giovannantonio, Matteo et al. (2020). “Heritable genetic variants in key cancer genes link cancer risk with anthropometric traits”. In: *Journal of Medical Genetics* 58.6, pp. 392–399. ISSN: 14686244. DOI: 10.1136/jmedgenet-2019-106799.

Dimitrakopoulos, Christos et al. (2019). “Identification and Validation of a Biomarker Signature in Patients with Resectable Pancreatic Cancer via Genome-Wide Screening for Functional Genetic Variants”. In: *JAMA Surgery* 154.6, e190484–e190484. ISSN: 21686254. DOI: 10.1001/jamasurg.2019.0484.

Duncan, L. et al. (2019). “Analysis of polygenic risk score usage and performance in diverse human populations”. In: *Nature Communications* 10.1, pp. 1–9. ISSN: 20411723. DOI: 10.1038/s41467-019-11112-0.

Fagny, Maud et al. (2020). “Nongenic cancer-risk SNPs affect oncogenes, tumour-suppressor genes, and immune function”. In: *British Journal of Cancer* 122.4, pp. 569–577. ISSN: 15321827. DOI: 10.1038/s41416-019-0614-3.

Ferlay, J et al. (2020). *Global Cancer Observatory: Cancer Today.* Lyon, France.

Finucane, Hilary K. et al. (2015). “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature Genetics* 47.11, pp. 1228–1235. ISSN: 1061-4036. DOI: 10.1038/ng.3404. arXiv: 15334406.

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

Foulkes, William D. (2008). “Inherited Susceptibility to Common Cancers”. In: *New England Journal of Medicine* 359.20, pp. 2143–2153. ISSN: 0028-4793. DOI: 10.1056/nejmra0802968.

Foulkes, William D., Bartha Maria Knoppers, and Clare Turnbull (2016). “Population genetic testing for cancer susceptibility: founder mutations to genomes”. In: *Nature Reviews Clinical Oncology* 13.1, pp. 41–54. ISSN: 1759-4774. DOI: 10.1038/nrclinonc.2015.173.

Frankish, Adam et al. (Dec. 2020). “GENCODE 2021”. In: *Nucleic Acids Research* 49.D1, pp. D916–D923. ISSN: 0305-1048.

Fritsche, Lars G et al. (2021). “On Cross-ancestry Cancer Polygenic Risk Scores”. In: *medRxiv*, p. 2021.02.24.21252351. DOI: 10.1101/2021.02.24.21252351.

Futschik, Matthias E., Gautam Chaurasia, and Hanspeter Herzel (Jan. 2007). “Comparison of human protein–protein interaction maps”. In: *Bioinformatics* 23.5, pp. 605–611. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl683. eprint: https://academic.oup.com/bioinformatics/article-pdf/23/5/605/16861213/btl683.pdf.

Gallagher, Michael D. and Alice S. Chen-Plotkin (2018). “The Post-GWAS Era: From Association to Function”. In: *American Journal of Human Genetics* 102.5, pp. 717–730. ISSN: 15376605. DOI: 10.1016/j.ajhg.2018.04.002.

Geijn, Bryce van de et al. (2021). “Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability”. In: *Human Molecular Genetics* 29.7, pp. 1057–1067. ISSN: 14602083. DOI: 10.1093/HMG/DDZ226.

Gerstung, Moritz et al. (2020). “The evolutionary history of 2,658 cancers”. In: *Nature* 578.7793, pp. 122–128. ISSN: 14764687. DOI: 10.1038/s41586-019-1907-7.

Ghasemian, Amir et al. (2020). “Stacking models for nearly optimal link prediction in complex networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 117.38, pp. 23393–23400. ISSN: 10916490. DOI: 10.1073/pnas.1914950117. arXiv: 1909.07578.

Golan, David, Eric S. Lander, and Saharon Rosset (2014). “Measuring missing heritability: Inferring the contribution of common variants”. In: *Proceedings of the National Academy of Sciences* 111.49, E5272–E5281. ISSN: 0027-8424. DOI: 10.1073/pnas.1419064111. arXiv: NIHMS150003.

Goyal, Palash and Emilio Ferrara (2018). “Graph embedding techniques, applications, and performance: A survey”. In: *Knowledge-Based Systems* 151, pp. 78–94. ISSN: 09507051. DOI: 10.1016/j.knosys.2018.03.022. arXiv: 1705.02801.

Grover, Aditya and Jure Leskovec (2016). “Node2vec: Scalable feature learning for networks”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu, pp. 855–864. ISSN: 2154-817X. DOI: 10.1145/2939672.2939754. arXiv: 1607.00653.

Gusev, Alexander et al. (2016). “Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation”. In: *Nature Communications* 7.1, pp. 1–13. ISSN: 20411723. DOI: 10.1038/ncomms10979.

Gysi, Deisy Morselli et al. (2021). “Network medicine framework for identifying drug-repurposing opportunities for COVID-19”. In: *Proceedings of the Na-*

tional Academy of Sciences of the United States of America 118.19. ISSN: 10916490. DOI: 10.1073/pnas.2025581118. arXiv: 2004.07229.

Haenig, Christian et al. (2020). “Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains”. In: *Cell Reports* 32.7, p. 108050. ISSN: 2211-1247.

Hahn, William C. et al. (2021). *An expanded universe of cancer targets*. DOI: 10.1016/j.cell.2021.02.020.

Haibe-Kains, Benjamin et al. (2020). *Transparency and reproducibility in artificial intelligence*. DOI: 10.1038/s41586-020-2766-y.

Hamilton, William L, Rex Ying, and Jure Leskovec (2017). *Inductive Representation Learning on Large Graphs*. Tech. rep.

Hanahan, Douglas and Robert A. Weinberg (2011). “Hallmarks of cancer: The next generation”. In: *Cell* 144.5, pp. 646–674. ISSN: 00928674. DOI: 10.1016/j.cell.2011.02.013.

Hansen, Tommy and Fabio Vandin (2016). “Finding Mutated Subnetworks Associated with Survival in Cancer”. In: pp. 1–26. arXiv: 1604.02467.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). “The Elements of Statistical Learning”. In: *Elements* 1, pp. 337–387. ISSN: 03436993. DOI: 10.1007/b94608. arXiv: 1010.3003.

He, Zongzhen et al. (2017). “Network based stratification of major cancers by integrating somatic mutation and gene expression data”. In: *PLoS ONE* 12.5, e0177662. ISSN: 19326203. DOI: 10.1371/journal.pone.0177662.

Henaff, Mikael, Joan Bruna, and Yann LeCun (2015). “Deep Convolutional Networks on Graph-Structured Data”. In: arXiv: 1506.05163.

Hoadley, Katherine A. et al. (2018). “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer”. In: *Cell* 173.2, 291–304.e6. ISSN: 10974172. DOI: 10.1016/j.cell.2018.03.022.

Hofree, Matan et al. (2013). “Network-based stratification of tumor mutations”. In: *Nature Methods* 10.11, pp. 1108–1115. ISSN: 1548-7091. DOI: 10.1038/nmeth.2651.

Holland, Paul W, Kathryn Blackmond, and Samuel Leinhardt (1983). *STOCHASTIC BLOCKMODELS: FIRST STEPS * Educational Testing Seroice ***. Tech. rep., pp. 9–137.

Horn, Heiko et al. (2018). “NetSig: Network-based discovery from cancer genomes”. In: *Nature Methods* 15.1, pp. 61–66. ISSN: 15487105. DOI: 10.1038/nmeth.4514.

Hristov, Borislav H., Bernard Chazelle, and Mona Singh (2020). “uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes”. In: *Cell Systems* 10.6, 470–479.e3. ISSN: 24054720. DOI: 10.1016/j.cels.2020.05.008.

Hristov, Borislav H. and Mona Singh (2017). “Network-Based Coverage of Mutational Profiles Reveals Cancer Genes”. In: *Cell Systems* 5.3, 221–229.e4. ISSN: 24054720. DOI: 10.1016/j.cels.2017.09.003.

Huang, Hailiang et al. (2011). “Gene-Based tests of association”. In: *PLoS Genetics* 7.7, e1002177. ISSN: 15537390. DOI: 10.1371/journal.pgen.1002177.

Hutchinson, Anna, Jennifer Asimit, and Chris Wallace (Aug. 2020). “Fine-mapping genetic associations”. In: *Human Molecular Genetics* 29.R1, R81–R88. ISSN: 0964-6906.

Huttlin, Edward L. et al. (2015). “The BioPlex Network: A Systematic Exploration of the Human Interactome”. In: *Cell* 162.2, pp. 425–440. ISSN: 10974172. DOI: 10.1016/j.cell.2015.06.043.

Huttlin, Edward L. et al. (2021). “Dual proteome-scale networks reveal cell-specific remodeling of the human interactome”. In: *Cell* 184.11, 3022–3040.e28. ISSN: 00928674. DOI: 10.1016/j.cell.2021.04.011.

Ihnatova, Ivana, Vlad Popovici, and Eva Budinska (2018). “A critical comparison of topology-based pathway analysis methods”. In: *PLOS ONE* 13.1. Ed. by Xia Li, e0191154. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0191154.

Jassal, Bijay et al. (2020). “The reactome pathway knowledgebase”. In: *Nucleic Acids Research* 48.D1, pp. D498–D503. ISSN: 13624962. DOI: 10.1093/nar/gkz1031.

Jeffers, John R. et al. (2021). “The common germline TP53-R337H mutation is hypomorphic and confers incomplete penetrance and late tumor onset in a mouse model A C”. In: *Cancer Research* 81.9, pp. 2442–2456. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-20-1750.

Jeggari, Ashwini and Andrey Alexeyenko (2017). “NEArender: An R package for functional interpretation of ’omics’ data via network enrichment analysis”. In: *BMC Bioinformatics* 18. ISSN: 14712105. DOI: 10.1186/s12859-017-1534-y.

Jia, Peilin and Zhongming Zhao (2014). “VarWalker: Personalized Mutation Network Analysis of Putative Cancer Genes from Next-Generation Sequencing Data”. In: *PLoS Computational Biology* 10.2, p. 1003460. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003460.

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

Jiao, Shuo et al. (2014). “Estimating the heritability of colorectal cancer”. In: *Human Molecular Genetics* 23.14, pp. 3898–3905. ISSN: 14602083. DOI: 10.1093/hmg/ddu087.

Kamoun, Aurélie et al. (2016). “Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas”. In: *Nature Communications* 7.1, pp. 1–11. ISSN: 20411723. DOI: 10.1038/ncomms11263.

Kanehisa, Minoru and Susumu Goto (2000). *KEGG: Kyoto Encyclopedia of Genes and Genomes*. DOI: 10.1093/nar/28.1.27.

Karrer, Brian and M. E.J. Newman (2011). "Stochastic blockmodels and community structure in networks". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 83.1. ISSN: 15393755. DOI: 10.1103/PhysRevE.83.016107. arXiv: 1008.3926.

Kavran, Andrew and Aaron Clauset (2020). "Denoising Large-Scale Biological Data Using Network Filters". In: DOI: 10.1101/2020.03.12.989244.

Khalighi, Sirvan, Salendra Singh, and Vinay Varadan (2020). *Untangling a complex web: Computational analyses of tumor molecular profiles to decode driver mechanisms*. DOI: 10.1016/j.jgg.2020.11.001.

Khatri, Purvesh, Marina Sirota, and Atul J. Butte (2012). *Ten years of pathway analysis: Current approaches and outstanding challenges*. Ed. by Christos A. Ouzounis. DOI: 10.1371/journal.pcbi.1002375.

Kichaev, Gleb et al. (2019). "Leveraging Polygenic Functional Enrichment to Improve GWAS Power". In: *American Journal of Human Genetics* 104.1, pp. 65–75. ISSN: 15376605. DOI: 10.1016/j.ajhg.2018.11.008.

Kipf, Thomas N and Max Welling (2016). *Variational Graph Auto-Encoders. A latent variable model for graph-structured data*. Tech. rep. arXiv: 1611.07308v1.

Kovács, István A. et al. (2019). "Network-based prediction of protein interactions". In: *Nature Communications* 10.1, pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-019-09177-y.

Kuenzi, Brent M. and Trey Ideker (2020). "A census of pathway maps in cancer systems biology". In: *Nature Reviews Cancer* 20.4, pp. 233–246. ISSN: 14741768. DOI: 10.1038/s41568-020-0240-7.

Kuijjer, Marieke Lydia et al. (2019). “Estimating Sample-Specific Regulatory Networks”. In: *iScience* 14, pp. 226–240. ISSN: 25890042. DOI: 10.1016/j.isci.2019.03.021. arXiv: 1505.06440.

Lander, Eric S. et al. (2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822, pp. 860–921. ISSN: 00280836. DOI: 10.1038/35057062.

Larremore, Daniel B., Aaron Clauset, and Caroline O. Buckee (2013). “A Network Approach to Analyzing Highly Recombinant Malaria Parasite Genes”. In: *PLoS Computational Biology* 9.10. Ed. by Rustom Antia, e1003268. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1003268. arXiv: 1308.5254.

Lawrenson, Kate et al. (2015). “Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer”. In: *Nature Communications* 6.1, pp. 1–14. ISSN: 20411723. DOI: 10.1038/ncomms9234.

Lee, Clement and Darren J. Wilkinson (2019). “A review of stochastic block models and extensions for graph clustering”. In: *Applied Network Science* 4.1, pp. 1–50. ISSN: 23648228. DOI: 10.1007/s41109-019-0232-2. arXiv: 1903.00114.

Lee, Sang Hong et al. (2011). “Estimating missing heritability for disease from genome-wide association studies”. In: *American Journal of Human Genetics* 88.3, pp. 294–305. ISSN: 00029297. DOI: 10.1016/j.ajhg.2011.02.002.

Lehner, Ben and Andrew G. Fraser (2004). “A first-draft human protein-interaction map.” In: *Genome biology* 5.9, p. 63. ISSN: 14656914. DOI: 10.1186/gb-2004-5-9-r63.

Leiserson, Mark D.M. et al. (2015). “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes”. In: *Nature Genetics* 47.2, pp. 106–114. ISSN: 15461718. DOI: 10.1038/ng.3168. arXiv: 15334406.

Li, Michelle M and Marinka Zitnik (2021). "Deep Contextual Learners for Protein Networks". In: arXiv: 2106.02246.

Li, Ruoyu et al. (2018). "Adaptive graph convolutional neural networks". In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 3546–3553. arXiv: 1801.03226.

Li, Yang I. et al. (2016). "RNA splicing is a primary link between genetic variation and disease". In: *Science* 352.6285, pp. 600–604. ISSN: 10959203. DOI: 10.1126/science.aad9417.

Liao, Yuxing et al. (2019). "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs". In: *Nucleic Acids Research* 47.W1, W199–W205. ISSN: 0305-1048. DOI: 10.1093/nar/gkz401.

Lindström, Sara et al. (2017). "Quantifying the Genetic Correlation between Multiple Cancer Types". In: *Cancer Epidemiology Biomarkers & Prevention* 26.9, pp. 1427–1435. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-17-0211.

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

Liu, Yuxi et al. (2021). "Somatic mutational profiles and germline polygenic risk scores in human cancer 1". In: *bioRxiv*, p. 2021.01.28.428663. DOI: 10.1101/2021.01.28.428663.

Lopes-Ramos, Camila M. et al. (2018). "Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism". In: *Cancer Research* 78.19, pp. 5538–5547. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-18-0454.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2".

In: *Genome Biology* 15.12, p. 550. ISSN: 1474760X. DOI: 10.1186/s13059-014-0550-8.

Lowe, Rohan et al. (2017). “Transcriptomics technologies”. In: *PLoS Computational Biology* 13.5, e1005457. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005457.

Luck, Katja et al. (2020). “A reference map of the human binary protein interactome”. In: *Nature* 580.7803, pp. 402–408. ISSN: 14764687. DOI: 10.1038/s41586-020-2188-x.

Lynch, Henry T. and Albert de la Chapelle (2003). “Hereditary Colorectal Cancer”. In: *New England Journal of Medicine* 348.10. Ed. by Alan E. Guttmacher and Francis S. Collins, pp. 919–932. ISSN: 0028-4793. DOI: 10.1056/NEJMra012242.

Ma, Jing, Ali Shojaie, and George Michailidis (2019). “A comparative study of topology-based pathway enrichment analysis methods”. In: *BMC Bioinformatics* 20.1, p. 546. ISSN: 14712105. DOI: 10.1186/s12859-019-3146-1.

MacArthur, Jacqueline et al. (2017). “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)”. en. In: *Nucleic Acids Research* 45.D1, pp. D896–D901. ISSN: 13624962. DOI: 10.1093/nar/gkw1133.

Mancuso, Nicholas et al. (2015). “The contribution of rare variation to prostate cancer heritability”. In: *Nature Genetics* 48.1, pp. 30–35. ISSN: 1061-4036. DOI: 10.1038/ng.3446.

Margolin, Adam A. et al. (2006). “ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context”. In: *BMC Bioinformatics* 7.SUPPL.1. ISSN: 14712105. DOI: 10.1186/1471-2105-7-S1-S7. arXiv: 0410037 [q-bio].

Mars, Nina et al. (2020). “Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers”. In: *Nature Medicine* 26.4, pp. 549–557. ISSN: 1546170X. DOI: 10.1038/s41591-020-0800-0.

Martincorena, Iñigo et al. (2015). “High burden and pervasive positive selection of somatic mutations in normal human skin”. In: *Science* 348.6237, pp. 880–886. ISSN: 10959203. DOI: 10.1126/science.aaa6806.

Martínez-Jiménez, Francisco et al. (2020). *A compendium of mutational cancer driver genes*. DOI: 10.1038/s41568-020-0290-x.

Mavaddat, Nasim et al. (2019). “Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes”. In: *American Journal of Human Genetics* 104.1, pp. 21–34. ISSN: 15376605. DOI: 10.1016/j.ajhg.2018.11.002.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. In: *Nucleic Acids Research* 40.10, pp. 4288–4297. ISSN: 03051048. DOI: 10.1093/nar/gks042.

McKinney, Scott Mayer et al. (2020). “International evaluation of an AI system for breast cancer screening”. In: *Nature* 577.7788, pp. 89–94. ISSN: 14764687. DOI: 10.1038/s41586-019-1799-6.

Menche, Jörg et al. (2015). “Uncovering disease-disease relationships through the incomplete interactome”. In: *Science* 347.6224, p. 841. ISSN: 10959203. DOI: 10.1126/science.1257601. arXiv: 15334406.

Meyerson, Matthew, Stacey Gabriel, and Gad Getz (2010). *Advances in understanding cancer genomes through second-generation sequencing*. DOI: 10.1038/nrg2841.

Mezlini, Aziz M and Anna Goldenberg (2017). “Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases”. In: *PLoS Computational Biology* 13.10. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005580.

Miki, Yoshio et al. (1994). “A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1”. In: *Science* 266.5182, pp. 66–71. ISSN: 00368075. DOI: 10.1126/science.7545954.

Moore, Luiza et al. (2020). *The mutational landscape of human somatic and germline cells*.

Muir, Paul et al. (2016). “The real cost of sequencing: Scaling computation to keep pace with data generation”. In: *Genome Biology* 17.1, pp. 1–9. ISSN: 1474760X. DOI: 10.1186/s13059-016-0917-0.

Newman, M. E.J. and Aaron Clauset (2016). “Structure and inference in annotated networks”. In: *Nature Communications* 7, pp. 1–16. ISSN: 20411723. DOI: 10.1038/ncomms11863. arXiv: 1507.04001.

Nurk, Sergey et al. (2021). “The complete sequence of a human genome”. In: *bioRxiv*, p. 2021.05.26.445798. DOI: 10.1101/2021.05.26.445798.

O’Connor, Luke J. et al. (2019). “Extreme Polygenicity of Complex Traits Is Explained by Negative Selection”. In: *American Journal of Human Genetics* 105.3, pp. 456–476. ISSN: 15376605. DOI: 10.1016/j.ajhg.2019.07.003.

Ozturk, Kivilcim et al. (2018). *The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine*. DOI: 10.1016/j.jmb.2018.06.016.

Padi, Megha and John Quackenbush (2018). “Detecting phenotype-driven transitions in regulatory network structure”. In: *npj Systems Biology and Applications* 4.1, pp. 1–12. ISSN: 20567189. DOI: 10.1038/s41540-018-0052-5.

Paull, Evan O. et al. (2013). “Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)”. In: *Bioinformatics* 29.21, pp. 2757–2764. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt471.

Peixoto, Tiago P. (2015). “Model selection and hypothesis testing for large-scale network models with overlapping groups”. In: *Physical Review X* 5.1,

p. 011033. ISSN: 21603308. DOI: 10.1103/PhysRevX.5.011033. arXiv: 1409.3059.

Peng, Xiaoqing et al. (July 2016). "Protein–protein interactions: detection, reliability assessment and applications". In: *Briefings in Bioinformatics* 18.5, pp. 798–819. ISSN: 1467-5463. DOI: 10.1093/bib/bbw066. eprint: https://academic.oup.com/bib/article-pdf/18/5/798/25581142/bbw066.pdf.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena (2014). "DeepWalk: Online learning of social representations". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710. DOI: 10.1145/2623330.2623732. arXiv: 1403.6652.

Pickrell, Joseph K (2014). "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits". In: *American Journal of Human Genetics* 94.4, pp. 559–573. ISSN: 15376605. DOI: 10.1016/j.ajhg.2014.03.004. arXiv: 1311.4843.

Pillich, Rudolf T. et al. (2017). "NDEx: A community resource for sharing and publishing of biological networks". In: *Methods in Molecular Biology*. Vol. 1558. Humana Press Inc., pp. 271–301. DOI: 10.1007/978-1-4939-6783-4_13.

Pimentel, Harold et al. (2017). "Differential analysis of RNA-seq incorporating quantification uncertainty". In: *Nature Methods* 14.7, pp. 687–690. ISSN: 15487105. DOI: 10.1038/nmeth.4324.

Pleasance, Erin D. et al. (2010). "A comprehensive catalogue of somatic mutations from a human cancer genome". In: *Nature* 463.7278, pp. 191–196. ISSN: 00280836. DOI: 10.1038/nature08658.

Qing, Tao et al. (2020). "Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden". In: *Nature Communications* 11.1, pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-020-16293-7.

Rahman, Nazneen (2014). *Realizing the promise of cancer predisposition genes*. DOI: 10.1038/nature12981.

Ramroop, Johnny R., Madelyn M. Gerber, and Amanda Ewart Toland (2019). *Germline Variants Impact Somatic Events during Tumorigenesis*. DOI: 10.1016/j.tig.2019.04.005.

Rapley, Elizabeth A. et al. (2009). "A genome-wide association study of testicular germ cell tumor". In: *Nature Genetics* 41.7, pp. 807–810. ISSN: 10614036. DOI: 10.1038/ng.394.

Ravasi, T. et al. (2010). "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man". In: *Cell* 140.5, pp. 744–752. ISSN: 00928674. DOI: 10.1016/j.cell.2010.01.044.

Reel, Parminder S. et al. (2021). *Using machine learning approaches for multi-omics data analysis: A review*. DOI: 10.1016/j.biotechadv.2021.107739.

Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder (2015). *High-Throughput Sequencing Technologies*. DOI: 10.1016/j.molcel.2015.05.004.

Reyes-Aldasoro, Constantino Carlos (2017). "The proportion of cancer-related entries in PubMed has increased considerably; Is cancer truly "the Emperor of All Maladies"?" In: *PLoS ONE* 12.3. ISSN: 19326203. DOI: 10.1371/journal.pone.0173671.

Reyna, Matthew A, Mark D.M. M Leiserson, and Benjamin J Raphael (2018). "Hierarchical HotNet: Identifying hierarchies of altered subnetworks". In: *Bioinformatics*. Vol. 34. 17. Oxford University Press, pp. i972–i980. DOI: 10.1093/bioinformatics/bty613.

Reyna, Matthew A. et al. (2020). "Pathway and network analysis of more than 2500 whole cancer genomes". In: *Nature Communications* 11.1, p. 16. ISSN: 20411723. DOI: 10.1038/s41467-020-14367-0.

Rhee, Sungmin, Seokjun Seo, and Sun Kim (2018). "Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification". In: *IJCAI International Joint Conference on Artificial*

Intelligence. Vol. 2018-July, pp. 3527–3534. ISBN: 9780999241127. DOI: 10.24963/ijcai.2018/490. arXiv: 1711.05859.

Rolland, Thomas et al. (2014). “A proteome-scale map of the human interactome network”. In: *Cell* 159.5, pp. 1212–1226. ISSN: 10974172. DOI: 10.1016/j.cell.2014.10.050.

Rozenblatt-Rosen, Orit et al. (2020). *The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution*. DOI: 10.1016/j.cell.2020.03.053.

Rual, Jean François et al. (2005). “Towards a proteome-scale map of the human protein-protein interaction network”. In: *Nature* 437.7062, pp. 1173–1178. ISSN: 00280836. DOI: 10.1038/nature04209.

Ruark, Elise et al. (2013). “Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14”. In: *Nature Genetics* 45.6, pp. 686–689. ISSN: 10614036. DOI: 10.1038/ng.2635.

Ruffalo, Matthew, Mehmet Koyutürk, and Roded Sharan (2015). “Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer”. In: *PLoS Computational Biology* 11.12. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004595.

Ruiz, Camilo, Marinka Zitnik, and Jure Leskovec (2021). “Identification of disease treatment mechanisms through the multiscale interactome”. In: *Nature Communications* 12.1, pp. 1–15. ISSN: 20411723. DOI: 10.1038/s41467-021-21770-8.

Sampson, Joshua N et al. (2015). “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, djv279. ISSN: 1460-2105. DOI: 10.1093/jnci/djv279.

Scarselli, Franco et al. (2009). “The graph neural network model”. In: *IEEE Transactions on Neural Networks* 20.1, pp. 61–80. ISSN: 10459227. DOI: 10.1109/TNN.2008.2005605.

Schulte-Sasse, Roman et al. (2021). “Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms”. In: *Nature Machine Intelligence* 3.6, pp. 513–526. ISSN: 25225839. DOI: 10.1038/s42256-021-00325-y.

Shi, Huwenbo, Gleb Kichaev, and Bogdan Pasaniuc (2016). “Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data”. In: *The American Journal of Human Genetics* 99.1, pp. 139–153. ISSN: 00029297. DOI: 10.1016/j.ajhg.2016.05.013.

Shi, Huwenbo et al. (2017). “Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits”. In: *The American Journal of Human Genetics* 101.5, pp. 737–751. ISSN: 00029297. DOI: 10.1016/j.ajhg.2017.09.022.

Silverbush, Dana and Roded Sharan (2019). “A systematic approach to orient the human protein–protein interaction network”. In: *Nature Communications* 10.1, pp. 1–9. ISSN: 20411723. DOI: 10.1038/s41467-019-10887-6.

Silverbush, Dana et al. (2019). “Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules”. In: *Cell Systems* 8.5, 456–466.e5. ISSN: 24054720. DOI: 10.1016/j.cels.2019.04.005.

Sondka, Zbyslaw et al. (2018). “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.” In: *Nature reviews. Cancer* 18.11, pp. 696–705. ISSN: 1474-1768. DOI: 10.1038/s41568-018-0060-1.

Southey, Melissa C. et al. (2016). “PALB2, CHEK2 and ATM rare variants and cancer risk: Data from COGS”. In: *Journal of Medical Genetics* 53.12, pp. 800–811. ISSN: 14686244. DOI: 10.1136/jmedgenet-2016-103839.

Stanley, Natalie et al. (2019). “Stochastic block models with multiple continuous attributes”. In: *Applied Network Science* 4.1, pp. 1–22. ISSN: 23648228. DOI: 10.1007/s41109-019-0170-z. arXiv: 1803.02726.

Stark, C. et al. (2006). “BioGRID: a general repository for interaction datasets”. In: *Nucleic Acids Research* 34.90001, pp. D535–D539. ISSN: 0305-1048. DOI: 10.1093/nar/gkj109.

Stephens, Zachary D. et al. (2015). “Big data: Astronomical or genomical?” In: *PLoS Biology* 13.7, e1002195. ISSN: 15457885. DOI: 10.1371/journal.pbio.1002195.

Stracquadanio, Giovanni et al. (2016). “The importance of p53 pathway genetics in inherited and somatic cancer genomes”. In: *Nature Reviews Cancer* 16.4, pp. 251–265. ISSN: 14741768. DOI: 10.1038/nrc.2016.15.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal (2009). *The cancer genome*. DOI: 10.1038/nature07943.

Stuart, Tim and Rahul Satija (2019). *Integrative single-cell analysis*. DOI: 10.1038/s41576-019-0093-7.

Subramanian, Aravind et al. (2005). “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. In: *Genetics* 102.43, pp. 15545–15550.

Sud, Amit, Ben Kinnersley, and Richard S. Houlston (2017). “Genome-wide association studies of cancer: current insights and future perspectives”. In: *Nature reviews. Cancer* 17.11, pp. 692–704. ISSN: 14741768. DOI: 10.1038/nrc.2017.82.

Sudlow, Cathie et al. (2015). “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLOS Medicine* 12.3, e1001779. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001779.

Sung, Hyuna et al. (2021). “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians*, caac.21660. ISSN: 0007-9235. DOI: 10.3322/caac.21660.

Surakhy, Mirvat et al. (2020). “A common polymorphism in the retinoic acid pathway modifies adrenocortical carcinoma age-dependent incidence”. In: *British Journal of Cancer* 122.8, pp. 1231–1241. ISSN: 15321827. DOI: 10.1038/s41416-020-0764-3.

Szklarczyk, Damian et al. (2019). “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic Acids Research* 47.D1, pp. D607–D613. ISSN: 0305-1048. DOI: 10.1093/nar/gky1131.

Taylor-Weiner, Amaro et al. (2019). “Scaling computational genomics to millions of individuals with GPUs”. In: *Genome Biology* 20.1, pp. 1–5. ISSN: 1474760X. DOI: 10.1186/s13059-019-1836-7.

Tuncbag, Nurcan et al. (2016). *Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package*. DOI: 10.1371/journal.pcbi.1004879.

U.S. Department of Health and Human Services (2016). *14th Report on carcinogens*.

Vandin, Fabio, Eli Upfal, and Benjamin J Raphael (2011). “Algorithms for detecting significantly mutated pathways in cancer”. In: *Journal of Computational Biology*. Vol. 18. 3, pp. 507–522. DOI: 10.1089/cmb.2010.0265.

Vanunu, Oron et al. (2010). “Associating genes and protein complexes with disease via network propagation”. In: *PLoS Computational Biology* 6.1. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000641.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. Neural information processing systems foundation, pp. 5999–6009. arXiv: 1706.03762.

Veličković, Petar et al. (2018). “Graph attention networks”. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv: 1710.10903.

Venkatesan, Kavitha et al. (2009). “An empirical framework for binary interactome mapping”. In: *Nature Methods* 6.1, pp. 83–90. ISSN: 15487091. DOI: 10.1038/nmeth.1280.

Vidal, Marc, Michael E. Cusick, and Albert László Barabási (2011). *Interactome networks and human disease*. DOI: 10.1016/j.cell.2011.02.016.

Visscher, Peter M. et al. (2017). “10 Years of GWAS Discovery: Biology, Function, and Translation”. In: *The American Journal of Human Genetics* 101.1, pp. 5–22. ISSN: 00029297. DOI: 10.1016/j.ajhg.2017.06.005.

Vogelstein, Bert et al. (2013). “Cancer genome landscapes.” In: *Science (New York, N.Y.)* 340.6127, pp. 1546–58. ISSN: 1095-9203. DOI: 10.1126/science.1235122.

Von Mering, Christian et al. (2002). “Comparative assessment of large-scale data sets of protein–protein interactions”. In: *Nature* 417.6887, pp. 399–403.

Vosoughi, Aram et al. (2020). “Common germline-somatic variant interactions in advanced urothelial cancer”. In: *Nature Communications* 11.1. ISSN: 20411723. DOI: 10.1038/s41467-020-19971-8.

Wainberg, Michael et al. (2019). “Opportunities and challenges for transcriptome-wide association studies”. In: *Nature Genetics* 51.4, pp. 592–599. ISSN: 15461718. DOI: 10.1038/s41588-019-0385-z.

Wainberg, Michael et al. (2021). “A genome-wide atlas of co-essential modules assigns function to uncharacterized genes”. In: *Nature Genetics* 53.5, pp. 638–649. ISSN: 15461718. DOI: 10.1038/s41588-021-00840-z.

Watanabe, Kyoko et al. (2017). “Functional mapping and annotation of genetic associations with FUMA”. In: *Nature communications* 8.1, pp. 1–11.

Weissbrod, Omer et al. (2020). “Functionally informed fine-mapping and polygenic localization of complex trait heritability”. In: *Nature Genetics* 52.12, pp. 1355–1363. ISSN: 15461718. DOI: 10.1038/s41588-020-00735-5.

Wen, Xiaoquan, Roger Pique-Regi, and Francesca Luca (2017). “Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization”. In: *PLoS Genetics* 13.3, e1006646. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006646.

Whitington, Thomas et al. (2016). “Gene regulatory mechanisms underpinning prostate cancer susceptibility”. In: *Nature Genetics* 48.4, pp. 387–397. ISSN: 15461718. DOI: 10.1038/ng.3523.

WHO (2020). *Global Health Estimates 2019: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019.* Tech. rep. Geneva: World Health Organization.

Wooster, R et al. (1994). “Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13”. In: *Science* 265.5181, pp. 2088–2090. ISSN: 0036-8075. DOI: 10.1126/science.8091231.

Wooster, Richard et al. (1995). “Identification of the breast cancer susceptibility gene BRCA2”. In: *Nature* 378.6559, pp. 789–792. ISSN: 00280836. DOI: 10.1038/378789a0.

Wu, Zonghan et al. (2019). *A Comprehensive Survey on Graph Neural Networks*. Tech. rep. X, pp. 1–22. arXiv: 1901.00596v3.

Xu, Keyulu et al. (2019). “How powerful are graph neural networks?” In: *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR. arXiv: 1810.00826.

Yang, Jian et al. (2011). “Genome partitioning of genetic variation for complex traits using common SNPs.” In: *Nature genetics* 43.6, pp. 519–25. ISSN: 1546-1718. DOI: 10.1038/ng.823.

Yao, Douglas W. et al. (2020). “Quantifying genetic effects on disease mediated by assayed gene expression levels”. In: *Nature Genetics* 52.6, pp. 626–633. ISSN: 15461718. DOI: 10.1038/s41588-020-0625-2.

Yardımcı, Galip Gürkan et al. (2019). “Measuring the reproducibility and quality of Hi-C data”. In: *Genome biology* 20.1, pp. 1–19.

You, Jiaxuan, Rex Ying, and Jure Leskovec (2020). “Design Space for Graph Neural Networks”. In: ISSN: 10495258. arXiv: 2011.08843.

Yuan, Han et al. (2019). “BindSpace decodes transcription factor binding signals by large-scale sequence embedding”. In: *Nature Methods* 16.9, pp. 858–861. ISSN: 15487105. DOI: 10.1038/s41592-019-0511-y.

Zhang, Jiani et al. (2018). “GaAN: Gated attention networks for learning on large and spatiotemporal graphs”. In: *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. Vol. 1. Association For Uncertainty in Artificial Intelligence (AUAI), pp. 339–349. ISBN: 9781510871601. arXiv: 1803.07294.

Zhang, Ping et al. (2021). “Germline and somatic genetic variants in the p53 pathway interact to affect cancer risk, progression, and drug response”. In: *Cancer Research* 81.7, pp. 1667–1680. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-20-0177.

Zhang, Yan Dora et al. (2020). “Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers”. In: *Nature*

Communications 11.1. ISSN: 20411723. DOI: 10.1038/s41467-020-16483-3.

Zheng, Fan et al. (2021). “Interpretation of cancer mutations using a multiscale map of protein systems”. In: *Science* 374.6563, eabf3067.

Zhou, Jian and Olga G Troyanskaya (2015). “Predicting effects of noncoding variants with deep learning-based sequence model”. In: *Nature Methods* 12.10, pp. 931–934. ISSN: 15487105. DOI: 10.1038/nmeth.3547.

Zhou, Jie et al. (2020). “Graph neural networks: A review of methods and applications”. In: *AI Open* 1, pp. 57–81. ISSN: 26666510. DOI: 10.1016/j.aiopen.2021.01.001. arXiv: 1812.08434.

Zhu, Bin et al. (2016). “An investigation of the association of genetic susceptibility risk with somatic mutation burden in breast cancer”. In: *British Journal of Cancer* 115.6, pp. 752–760. ISSN: 15321827. DOI: 10.1038/bjc.2016.223.

Zitnik, Marinka, Monica Agrawal, and Jure Leskovec (2018). “Modeling polypharmacy side effects with graph convolutional networks”. In: *Bioinformatics*. Vol. 34. 13. Oxford Academic, pp. i457–i466. DOI: 10.1093/bioinformatics/bty294. arXiv: 1802.00543.

Zitnik, Marinka et al. (2019). “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. In: *Information Fusion* 50, pp. 71–91. ISSN: 15662535. DOI: 10.1016/j.inffus.2018.09.012. arXiv: 1807.00123.