Co-expression in tissue-specific gene networks links genes in cancer-susceptibility loci to known somatic driver genes

Urzúa-Traslaviña, Carlos G.; van Lieshout, Tijs; Boulogne, Floranne; Domanegg, Kevin; Zidan, Mahmoud; Bakker, Olivier B.; Claringbould, Annique; de Ridder, Jeroen; Zwart, Wilbert; Westra, Harm Jan

Published in:
BMC Medical Genomics

DOI:
10.1186/s12920-024-01941-4

Publication date:
2024

BMC Medical Genomics

RESEARCH

Open Access

Co-expression in tissue-specific gene networks links genes in cancer-susceptibility loci to known somatic driver genes

Carlos G. Urzúa-Traslaviña[1,2†], Tijs van Lieshout[1,2†], Floranne Boulogne[1,2†], Kevin Domanegg[1], Mahmoud Zidan[1], Olivier B. Bakker[3,4], Annique Claringbould[1,5], Jeroen de Ridder[2,6], Wilbert Zwart[2,7], Harm-Jan Westra[1,2], Patrick Deelen[1,2] and Lude Franke[1,2*]

Abstract

Background The genetic background of cancer remains complex and challenging to integrate. Many somatic mutations within genes are known to cause and drive cancer, while genome-wide association studies (GWAS) of cancer have revealed many germline risk factors associated with cancer. However, the overlap between known somatic driver genes and positional candidate genes from GWAS loci is surprisingly small. We hypothesised that genes from multiple independent cancer GWAS loci should show tissue-specific co-regulation patterns that converge on cancer-specific driver genes.

Results We studied recent well-powered GWAS of breast, prostate, colorectal and skin cancer by estimating co-expression between genes and subsequently prioritising genes that show significant co-expression with genes mapping within susceptibility loci from cancer GWAS. We observed that the prioritised genes were strongly enriched for cancer drivers defined by COSMIC, IntOGen and Dietlein et al. The enrichment of known cancer driver genes was most significant when using co-expression networks derived from non-cancer samples of the relevant tissue of origin.

Conclusion We show how genes within risk loci identified by cancer GWAS can be linked to known cancer driver genes through tissue-specific co-expression networks. This provides an important explanation for why seemingly unrelated sets of genes that harbour either germline risk factors or somatic mutations can eventually cause the same type of disease.

Keywords Cancer susceptibility genes, Cancer drivers, Tissue-specific, Gene networks, GWAS, recount3, Breast cancer, Prostate cancer, Colon cancer, Omnigenic

†Carlos G. Urzúa-Traslaviña, Tijs van Lieshout and Floranne Boulogne contributed equally to this work.

*Correspondence:
Lude Franke
l.h.franke@umcg.com
[1]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands
[2]Oncode Institute, Utrecht, The Netherlands
[3]Wellcome Sanger Institute, Human Genetics, Hinxton, UK
[4]Open Targets, Hinxton, UK
[5]EMBL Heidelberg, Structural and Computational Biology Unit, Heidelberg, Germany
[6]University Medical Center Utrecht, Utrecht, The Netherlands
[7]Division of Oncogenomics, Netherlands Cancer Institute, Amsterdam, The Netherlands

Background

Cancer continues to be a formidable health challenge, responsible for a substantial burden of morbidity and mortality worldwide [1]. Despite the many advances in cancer research, the molecular mechanisms that underlie cancer initiation and progression remain incompletely understood. It is now widely recognised that genetic alterations play a central role in the aetiology of most cancer types, enabling the acquisition of hallmark characteristics such as uncontrolled cell proliferation, evasion of cell death and the ability to invade and metastasise [2].

Two distinct types of genetic alterations have emerged as key players in cancer development: germline risk variants and somatic mutations. Germline risk variants are constitutional alterations present in every cell of an individual's body that are inherited from one or both parents. These mutations can confer an increased susceptibility to developing cancer [3, 4]. In contrast, somatic mutations arise in somatic cells and progressively accumulate over time, leading to the transformation of normal cells into their malignant counterparts. These kinds of mutations are the driving force behind the clonal expansion of malignant cells and the heterogeneity observed within tumours [5].

These two types of genetic alterations sometimes occur in the same gene suggesting a common mechanism for initiation (cancer susceptibility gene) and sustaining of cancer (cancer driver gene) [6]. Here we refer to 'cancer driver gene' as genes annotated to drive cancer by the COSMIC Cancer Gene Census, or with evidence of non-neutral somatic mutation frequencies by IntOGen or Dietlein et al. [5, 7, 8]. For example, *TP53* is often somatically mutated in several forms of cancer, but it also harbours rare germline variants that increase the risk of cancer development in Li-Fraumeni Syndrome patients [9]. Similarly, studies of rare germline coding variants in breast cancer have also found overlap with the cancer-driver genes derived from tumour data [10]. However, little overlap has been observed between cancer driver genes and genes that are influenced by more common germline variants. Hundreds of susceptibility variants, each conferring a small risk, have now been identified through genome-wide association studies (GWAS) of various cancers [11]. Yet only 5% of the genes implicated through cancer GWAS are established cancer driver genes [12].

These observations suggest that common germline cancer risk variants confer their risk by preferentially acting in a set of 'co-driver' genes that indirectly facilitate oncogenesis later in life [13]. In support of this idea, recent studies have found associations between germline variants and the tumour genome of patients with cancer [14, 15]. However, the question remains if and how the genes affected by common variants are influencing the activity of cancer drivers. As observed with rare germline variants, one possibility is that common germline variants influence the type of somatic mutational pattern of the tumour which then favours the activation of different cancer drivers [16]. Another possibility is that the genes affected by common variants are 'peripheral' genes which regulate a set of 'core' genes that mediate the cancer risk as suggested by the omnigenic model [17]. In previous work, Ratajczak et al. observed the omnigenic model to hold for the effect of common germline variants associated with a trait in 'peripheral' genes converging on 'core' genes that influence the given trait [18]. In line with this model, we hypothesised that some somatic cancer drivers may be regulated by the peripheral genes derived from cancer risk loci identified through GWAS.

We therefore aimed to test this hypothesis for four types of cancer—breast cancer, prostate cancer, colorectal cancer and skin cancer—for which GWAS have been conducted in a substantial number of cases. To achieve this aim, we first ascertained whether the genes inside these GWAS loci are suspected to be functionally related within a gene network derived from mRNA co-expression. We then determined if 'core' genes, genes that are highly linked to genes inside these cancer GWAS loci, are known somatic cancer driver genes. For this purpose, we applied Downstreamer: a gene prioritisation methodology that uses co-expression information to prioritise genes that are significantly co-expressed with genes in significant GWAS loci [19]. Downstreamer relies upon a gene-scorer, in this study PascalX, that estimates the extent that variants within and near each gene show an association signal in the GWAS.

We explored different gene co-expression networks to determine which regulatory context produces the strongest enrichment of somatic cancer driver genes. We considered co-expression networks generated from a mix of tissues (a multi-tissue network) and from specific tissues. Because cancer and adjacent healthy tissues may have different mRNA expression characteristics [20], we also studied whether networks derived exclusively from cancer or non-cancer samples affect the enrichment of driver genes.

We obtained the strongest enrichments of cancer driver genes when using networks derived from non-cancer tissue from the same origin as the cancer in question. Overall, our results suggest that genes in cancer GWAS loci seem to converge on downstream cancer driver genes for some cancers, suggesting that genes in GWAS loci may confer their risk by indirectly acting on those genes. Moreover, the prioritisation performed by Downstreamer points to other genes that could be considered novel cancer susceptibility genes.

Table 1 Overview of the four genome-wide association studies (GWAS) used in this study. GWAS ID is the identifier in the GWAS catalogue. The number of independent loci was determined using the Downstreamer methodology (see Methods). Number of cases is the number of European ancestry cases reported by the GWAS catalogue

Study	GWAS ID	Cancer	Number of independent loci	Number of cases
Schumacher et al., *Nat. Genet.* 2018	GCST006085	Prostate	335	79,148
Zhang et al., *Nat. Genet.* 2020	GCST010098	Breast	348	133,384
Sakaue et al., *Nat. Genet.* 2021	GCST90018921	Skin	123	25,928
Fernandez-Rozadilla et al., *Nat. Genet.* 2022	GCST90102485	Colorectal	99	100,204

Table 2 Cancer driver genes per tissue of origin. 'Total' refers to the total number of collected driver genes from the COSMIC GCG, IntOGen or Dietlein et al. annotated to be of relevance to the tissue of origin. 'Unique to tissue' refers to the total number of collected driver genes that are annotated only for the specified tissue of origin

Tissue of origin	Total	Unique to tissue
Breast	131	51
Prostate	106	49
Colon	142	71
Skin	177	94

Results

We collected summary statistics for 109 different GWAS of cancer traits and focused on those studies that identified at least 90 risk loci and had the highest number of cases among all the GWAS of the same cancer type (Table S1). After selection, four cancer GWAS remained that studied the following cancers: prostate, breast, colon and skin (Table 1).

We next created a catalogue of confirmed and putative somatic cancer driver genes for a specific tissue of origin by combining genes annotated to drive cancer by the COSMIC CGC (Version 96, Tier 1 and Tier 2), or with evidence of non-neutral somatic mutation frequencies by IntOGen and Dietlein et al. [5, 7, 8] (Fig. S1, Table 2, Table S2). Using these resources, we evaluated if cancer drivers are highly coregulated with genes inside GWAS loci (Fig. 1).

Cancer-specific driver genes show limited enrichment for cancer GWAS signal

Previous reports have indicated that there is limited overlap between somatic cancer drivers and genes in cancer GWAS loci [12]. Because we are studying very recent GWAS that could have identified additional loci since the previous report, we first re-evaluated this overlap. To do so, we used PascalX [21] to estimate gene-level *p*-values from the GWAS summary statistics of the prostate, breast, skin and colon cancer studies using 19,922 protein-coding genes (Ensembl, release 94) and a window size of 25 kb around each gene (Table S3). The resulting gene-level *p*-values denote to what extent variants within and near these genes show an association signal in the GWAS. To test for the expected baseline enrichment, we looked at how many of the Bonferroni significant (p-value $< 2.51 \times 10^{-6}$) PascalX prioritised genes are known tissue-specific somatic cancer drivers. Among the Bonferroni significant hits, 9 out of the 344 (2.62%) genes prioritised for the breast cancer GWAS are known somatic tissue-specific cancer drivers (*FGFR2*, *MAP3K1*, *RAD51B, ESR1, CDKN2A, CASP8, TBX3, AKT1* and *PIK3R1*, Fisher exact *p*-value: 0.00041, median distance to nearest index variant = 34.9 Kb). For the prostate cancer GWAS, 5 out of the 382 (1.31%) genes prioritised by PascalX are known prostate cancer drivers (*KLK2, KLK3, SPEN, TMPRSS2* and *RNF43*, Fisher exact *p*-value: 0.05, median distance to nearest index variant = 6.2 Kb). In the PascalX prioritization for the skin cancer GWAS, 4 out of the 148 (2.70%) genes prioritised are known skin cancer drivers (*CASP8, TERT, CDKN2A* and *BCL2L12*, Fisher exact *p*-value: 0.04, median distance to nearest index variant = 40.8 Kb). No Bonferroni significant cancer drivers were found for the colorectal cancer GWAS in the PascalX analysis. We also tested (one-sided Wilcoxon rank-sum test) whether all cancer-specific driver genes generally show a more significant gene-level GWAS *p*-value compared to all other genes (Fig. S2). We observed significant enrichments for the breast (p-value: 1.13×10^{-5}) and colon cancer GWAS (p-value: 0.008), suggesting that beyond the Bonferroni significant hits there is an indication of enrichment for the breast cancer and colon cancer GWAS. These findings suggest that, in general, PascalX identifies some cancer driver genes using summary statistics of well-powered GWAS and that the amount of cancer drivers is consistent with the percentages reported in the literature [12]. However, many more somatic drivers associated with these cancers are not called as significant by PascalX.

Additional cancer driver genes can be linked to GWAS loci through co-expression networks

Following the low overlap between genes identified by PascalX from GWAS and somatic driver genes we next tested the hypothesis of whether genes identified by PascalX influence cancer drivers indirectly. For this, we employed Downstreamer, which uses a gene network (co-expression model derived from bulk mRNA transcription) together with gene level *p*-values derived by

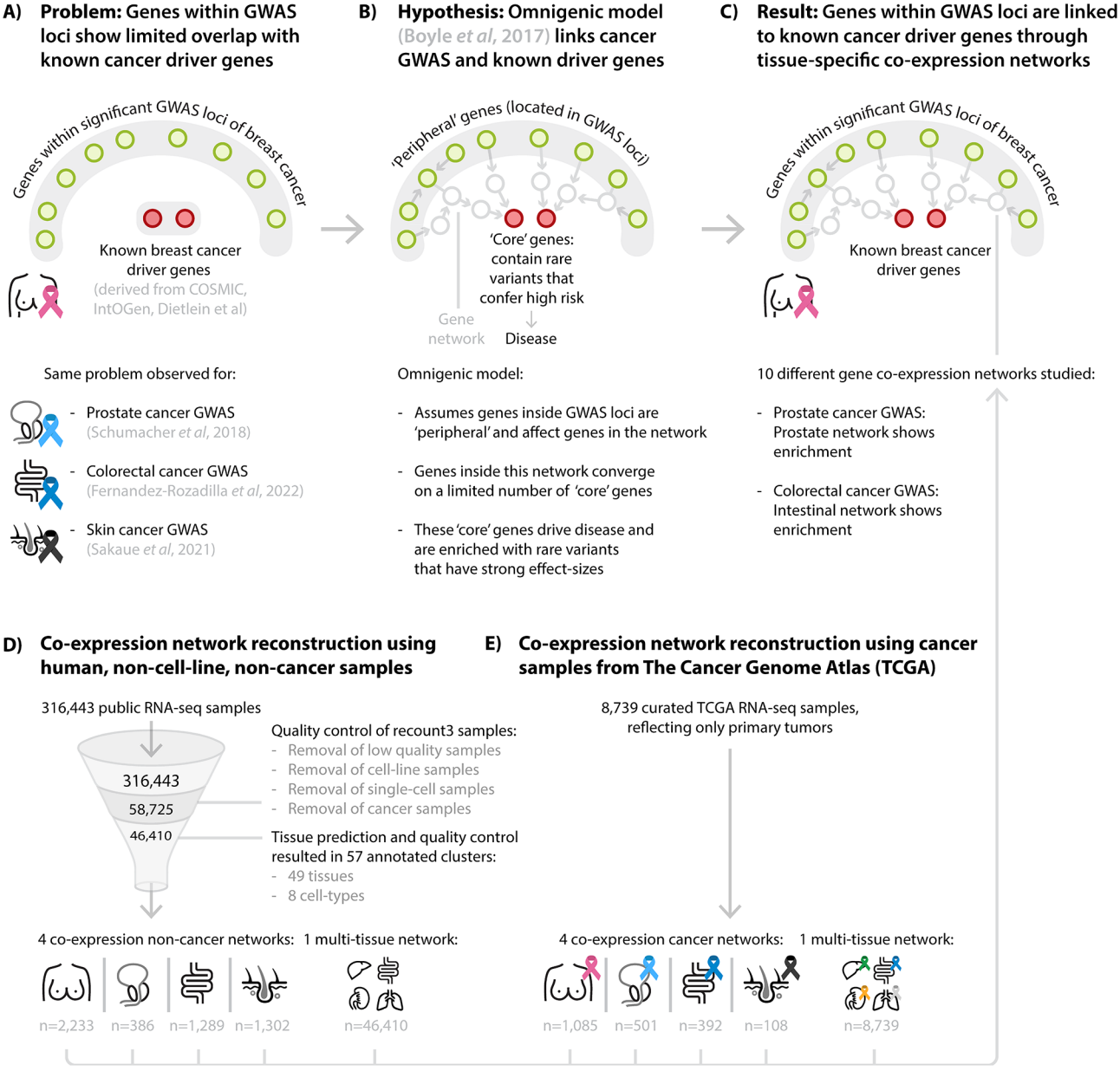

Fig. 1 Overview of application of gene-prioritisation methods on cancer GWAS summary statistics to identify cancer-type-specific somatic driver genes

PascalX. Downstreamer calculates a key gene prioritisation score that captures if a gene is significantly co-expressed with genes prioritised by PascalX from GWAS data [19].

Co-expression networks derived from a mixture of tissues are derived from a comparatively larger number of samples and therefore may better model co-expression links between genes due to increased statistical power. However, deriving networks from tissue-specific subsets of samples may result in networks that better capture co-expression links relevant to the tissue. We therefore tested networks derived exclusively from the tissue of origin of the primary tumour or from a mixture of multiple tissues using publicly available data from the cancer genome atlas (TCGA) and recount3 [22] (Table 3) (See Methods for further details on network construction). Each co-expression network has a different number of components that capture the variation in transcriptional patterns which Downstreamer uses when performing the prioritizations [19]. Additionally, as cancer predisposition mechanisms may operate both before and after the onset of oncogenesis, we compared the results from networks derived from cancer tissue and non-cancer tissue. For every combination of multi-tissue and tissue-specific networks, we applied the Downstreamer methodology to the cancer GWAS (Table S4, Data S1).

Table 3 Co-regulation networks generated in this study

Co-regulation network	Resource	Components	Genes	Samples	Type
Non-cancer adipose-breast	recount3	223	33,935	2,233	Tissue-specific non-cancer
Non-cancer colon	recount3	139	35,623	1,289	Tissue-specific non-cancer
Non-cancer prostate	recount3	63	33,873	386	Tissue-specific non-cancer
Non-cancer skin	recount3	141	35,524	1,302	Tissue-specific non-cancer
Breast cancer	TCGA	161	35,523	1,085	Tissue-specific cancer
Colorectal cancer	TCGA	65	35,523	392	Tissue-specific cancer
Prostate cancer	TCGA	81	35,523	501	Tissue-specific cancer
Skin cancer	TCGA	27	35,523	108	Tissue-specific cancer
Multi-tissue non-cancer	recount3	848	28,942	46,410	Non-cancer multi-tissue
Multi-tissue cancer	TCGA	1,864	35,523	8,739	Cancer multi-tissue

We first describe the results of applying Downstreamer using the two multi-tissue networks: one derived exclusively from cancer (TCGA multi-cancer) and one derived from non-cancer tissue data (recount3 non-cancer multi-tissue; Table 3, Fig. S3).

Among the prioritised genes we identified known somatic driver genes passing the Bonferroni significance threshold ($p<0.05/19{,}922=2.51\times10^{-6}$). The result from the analysis with the cancer multi-tissue network and the prostate cancer GWAS analysis included the gene *CANT1* (p-value: 2.0×10^{-6}). *CANT1* is located 7.7 Mb from the nearest prostate cancer GWAS index variant. The large distance from the nearest index variant suggests that the GWAS signal near this gene is weak and explains why this gene was not genome wide significant through the PascalX approach (PascalX p-value: 0.043). *CANT1* encodes the androgen-regulated protein CANT1, which has previously been found to form a fusion transcript with *ETV4* in prostate cancer [23, 24]. Note that *ETV4* does not seem to be prioritised by PascalX (PascalX p-value: 0.54) or Downstreamer (p-value 0.64). Results from the analysis with the non-cancer multi-tissue network and the colon cancer GWAS included the genes *KMT2B, PTBP1* and *GEN1* (p-values 2.0×10^{-6}, 3.0×10^{-6} and 3.0×10^{-6}, respectively). Each of these genes are located at least 2.5 Mb away from the nearest index variant. *PTBP1* expression has previously been associated with invasiveness in colorectal cancer through alternative splicing of cortactin [25]. The analysis of the breast cancer GWAS with the same non-cancer multi-tissue network resulted in one Bonferroni significant gene *LRP1* (p-values 1.22×10^{-6}). *LRP1* repressed xenografts of triple negative breast cancer cell lines have been shown to decrease tumour growth and angiogenesis [26]. No Bonferroni significant hits were found for the skin cancer GWAS using both cancer and non-cancer multi-tissue networks. To determine if there is overrepresentation of cancer driver genes in the Downstreamer prioritisation scores, we performed a Wilcoxon rank-sum test of cancer specific driver genes versus all other genes. We observed improved enrichment p-values for known somatic cancer drivers in the gene prioritisations for the prostate GWAS when using the cancer multi-tissue network compared with the equivalent enrichment performed on PascalX results of the same GWAS (p-value: 7.76×10^{-4} vs. 0.08 one-sided Wilcoxon rank-sum test). We observed a smaller improvement in enrichment for the colon cancer GWAS when using the non-cancer multi-tissue network as compared to PascalX (p-value: 8.43×10^{-4} vs. 0.009).

We then created tissue-specific networks derived from cancer and non-cancer samples for prostate, breast, skin and colon tissues. In contrast to the previous multi-tissue networks, these are derived from a smaller subset of tissue specific samples to better capture tissue-specific co-expression. We applied Downstreamer to each cancer GWAS using the tissue-specific network corresponding to the tissue of origin (referred to hereafter as a 'matched network') (Fig. 2, Fig. S4). This analysis also resulted in Bonferroni significant hits: in the matched non-cancer tissue-specific network analysis, we prioritised three genes for breast cancer, *PHLPP1, PUM1* and *CCL28* (p-values: 5.93×10^{-7}, 1.55×10^{-6} and 2.01×10^{-6}, respectively), and two for colon cancer, *ZZEF1* and *BAZ2A* (p-values: 1.94×10^{-6} and 2.56×10^{-6}, respectively). For breast cancer, there is evidence that *CCL28* and *PHLPP1* may be drivers of breast cancer through the MAPK and AKT pathways, respectively [27, 28]. While *PUM1* and *BAZ2A* have, to our knowledge, not been implicated for the matching cancer types, they have been shown to be potential drivers in other cancers [29, 30].

Our selection of driver genes contains both confirmed and putative cancer drivers that have not been validated, we therefore repeated the enrichment analysis while restricting the genes to only the high confidence set defined as Tier 1 in the COSMIC Cancer Gene Census, removing all the COSMIC non-Tier 1, the IntOGen, and Dietlein sets of driver genes. We observed a less significant but comparable enrichment of Tier 1 breast cancer drivers giving support to the results observed in the full set of cancer drivers (Fig. S6). Overall, these enrichments further suggest how GWAS-associated genes may

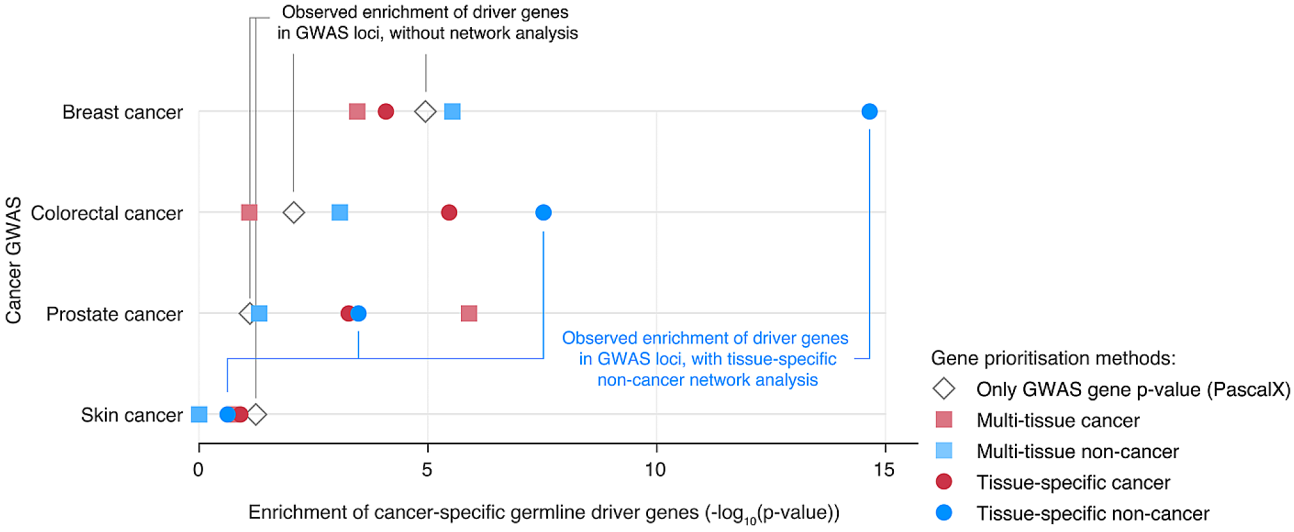

Fig. 2 Enrichment of cancer-specific somatic driver genes for different multi-tissue and tissue-specific gene-prioritisation methods applied to cancer GWAS summary statistics. A one-sided Wilcoxon rank-sum test was used to calculate the enrichment of cancer-specific driver genes in the list of prioritised genes (x-axis) for four cancer GWAS studies (y-axis). Open diamonds indicate gene enrichment values for the original GWAS, calculated with PascalX from the GWAS summary statistics. Coloured symbols indicate the enrichments for the different tissue networks, calculated using Downstreamer

Table 4 Enrichment of cancer-specific somatic driver genes for multi-tissue networks applied to cancer GWAS summary statistics

Matched tissue-specific network	*p*-value	Multi-tissue network	*p*-value
breast cancer	8.28×10^{-5}	multi-tissue cancer	3.51×10^{-4}
breast non-cancer	2.23×10^{-15}	multi-tissue non-cancer	2.94×10^{-6}
colon cancer	3.44×10^{-6}	multi-tissue cancer	0.008
colon non-cancer	2.98×10^{-8}	multi-tissue non-cancer	8.43×10^{-4}
prostate cancer	5.40×10^{-4}	multi-tissue cancer	1.27×10^{-6}
prostate non-cancer	3.29×10^{-4}	multi-tissue non-cancer	0.049
skin cancer	0.125	multi-tissue cancer	0.175
skin non-cancer	0.238	multi-tissue non-cancer	1.00

be converging to cancer-associated genes downstream within tissue-specific co-expression networks.

In order to bolster the credibility of our findings, we have applied PoPS, another gene prioritization methodology from Weeks et al., to the same set of summary statistics from the cancer GWAS and the same co-expression networks we used with our Downstreamer methodology [31]. Since PoPS relies on MAGMA as a gene-scorer we reprocessed our summary statistics. When applying PoPS to the MAGMA scores we observed highly concordant enrichments as compared to the Downstreamer analysis (Fig. S5).

When again performing Wilcoxon rank-sum test to test the enrichment of cancer specific driver genes versus other genes, we observed that the breast and colon cancer GWAS showed stronger and significant enrichments using the matched tissue-specific networks compared to multi-tissue networks, which did not consistently hold for prostate cancer or skin cancer (see *p*-values in Table 4).

Overall, the best performing combination was the breast cancer GWAS when using the adipose-breast co-regulation network (*p*-value: 2.23×10^{-15}), while the worst performing combination was the skin cancer GWAS, for which we observed no Bonferroni-significant enrichment (*p*-value$<$0.00125, calculated by dividing 0.05 by the 40 tested networks) using any of the matching networks. We additionally performed the same analysis using sets of genes with oncogene or tumour suppressor evidence as defined by the COSMIC CGC and IntOGen metadata. The enrichment observed for the breast cancer GWAS was higher for tumour suppressor genes while for the prostate cancer GWAS we observed higher enrichment for oncogenes. This suggests that both oncogenes and tumour suppressor genes may be downstream of the risk variants captured by cancer GWAS and that different tumour types may favor different types of driver genes (Fig. S7, Fig. S8).

Furthermore, for the breast cancer GWAS, we observed Bonferroni-significant enrichments (*p*-value$<$0.00125, calculated by dividing 0.05 by the 40 tested networks) for all other non-matched networks (Fig. 3). We reasoned this could be due to cancer driver genes that are shared with other cancer types and not specific to breast cancer. We therefore repeated the analysis using only breast cancer specific cancer driver genes and observed that the magnitude of non-matched enrichment was diminished, showing -$\log_{10}$(*p*-values) with a range of 3.5 to 10 for non-exclusive breast cancer driver genes compared to a range of 0.5 to 3 for exclusive breast cancer driver

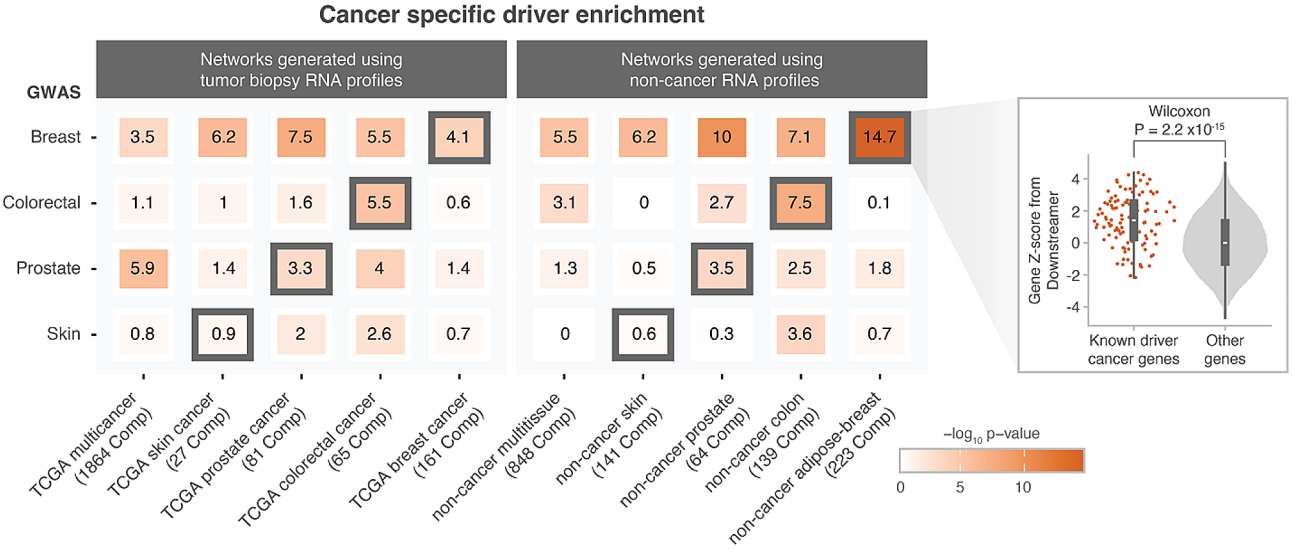

Fig. 3 Enrichment of cancer-specific somatic driver genes for all combinations of different multi-tissue and tissue-specific gene-prioritisation methods applied to cancer GWAS summary statistics. X-axis indicates the different tissue networks created based on cancer tissue data from TCGA and the different non-cancer tissue networks created based on tissue data from recount3. Y-axis indicates the different GWAS considered in our study. A one-sided Wilcoxon rank-sum test was used to calculate the enrichment of cancer-specific somatic driver genes in the list of prioritised genes. Rectangles with a black border correspond to the tissue-specific gene-prioritisation method that matches the tissue of origin for the cancer GWAS trait. The violin chart on the right highlights an example comparison for the breast cancer GWAS results in the non-cancer adipose-breast network

genes (Fig. S9). It could also be the case that the enrichment observed with non-matching networks is driven by the GWAS, independently of what network is used. We therefore compared the gene-prioritisation scores between all tested multi-tissue and tissue-specific networks. For the enrichments derived from the breast cancer GWAS and all tested co-expression networks no Pearson's correlation above 0.4 between any combination was observed (Fig. S10). This supports that the high enrichment in non-matching networks in breast cancer GWAS is not driven by the breast cancer GWAS. For the case of GWAS paired with non-matching networks we observed the most significant enrichment for the breast cancer GWAS with the prostate networks (cancer prostate network p-value: 3.10×10^{-8}, non-cancer prostate network p-value: 1.03×10^{-10}, Fig. 3). This observation is in concordance with a recent study that quantified shared heritability from multiple different cancer GWAS, where a positive genetic correlation was observed between breast and prostate cancer [32]. Together, these results suggest that enrichment in non-matched networks is driven by cancer driver genes that are shared between multiple cancer types or by shared heritability.

Relevant breast cancer driver genes are co-expressed with breast cancer GWAS associated-genes through a non-cancer adipose-breast network

As shown in Fig. 3, the combination of using a non-cancer adipose-breast network for the breast cancer GWAS resulted in the most significant observed enrichment in known cancer driver genes (p-value: 2.23×10^{-15}). This network is derived from adipose tissue samples in addition to breast cancer tissue samples because these samples of breast and adipose tissue were indistinguishable during the preparation of tissue-specific networks (Note S2).

To showcase specific examples of genes we investigated the known cancer driver genes that were prioritised by Downstreamer to be in the top 100 'core' genes.

For the non-cancer adipose-breast network that we applied to the breast cancer GWAS, Downstreamer prioritised seven known breast cancer driver genes among the top 100 genes: *CREBBP*, *TRPS1*, *ARID1A*, *ARID1B*, *XBP1*, *SPEN* and *NUMA1* (Fig. 4). *ARID1A* and *ARID1B* are well established Tier 1 COSMIC GCC breast cancer drivers and are located far away from the index variants in the corresponding GWAS (8.2 and 4.2 Mb respectively). *TRPS1* and *XBP1* have also been statistically implicated to likely be breast cancer driver genes based on analyses of whole tumour genome by Dietlein et al. [8]. . The *TRPS1* gene encodes a transcription factor of the GATA family that has preliminary evidence of potential driver role as a regulator of the downstream targets of the estrogen receptor α [33, 34]. *CREBBP*, *NUMA* and *SPEN* have not been linked to breast cancer by the COSMIC CGC but have been statistically determined to be breast cancer drivers by IntOGen. The protein encoded by the *SPEN* gene acts as an estrogen receptor cofactor and has been shown to have a tumour-suppressor role in regulating tumour growth, cell proliferation and

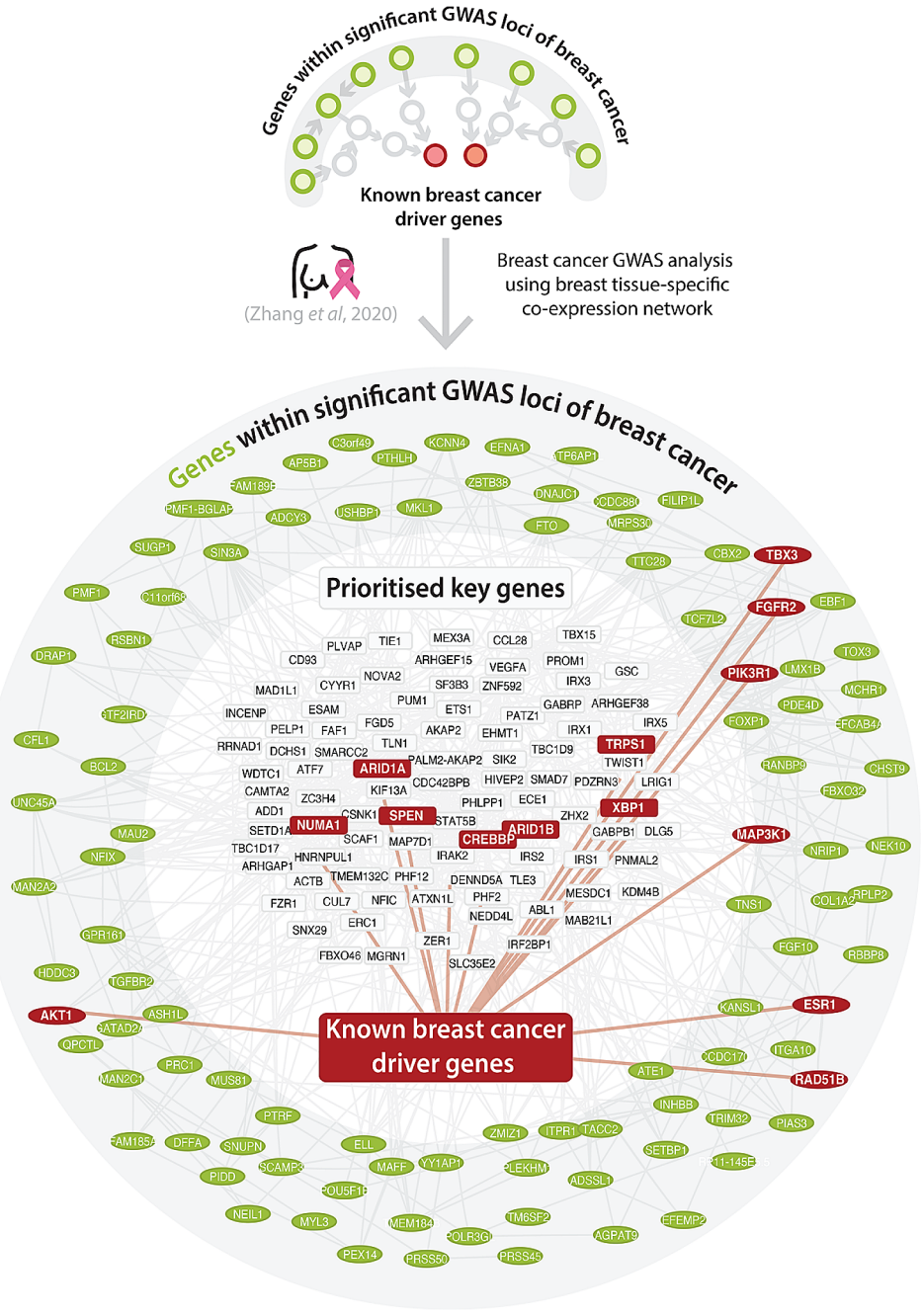

Fig. 4 Co-expression of the top 100 prioritised genes of the breast cancer GWAS using PascalX (outer circle) and a non-cancer cancer adipose-breast tissue-specific co-regulation model with Downstreamer (inner circle). Genes in the periphery are derived from the GWAS summary statistics using PascalX (potential peripheral genes; green nodes) while genes in the centre are derived with Downstreamer (potential 'core' genes; grey nodes). Edges are drawn when the co-regulation between genes has an absolute Z-score greater than two. Nodes with a red colour indicate known somatic cancer driver genes. Although some cancer drivers are prioritised by PascalX from GWAS summary statistics, the Downstreamer prioritisations point to new breast-specific cancer drivers that were not prioritised using only the GWAS data alone. For illustration purposes, the top 100 prioritised genes by Downstreamer and the top 100 PascalX genes prioritised genes by Downstreamer that are Bonferroni significant in their PascalX score are shown here

survival in ERα-expressing breast cancer cell lines [35]. In conclusion, these results indicate that a non-cancer adipose-breast network can be applied to breast cancer GWAS summary statistics to identify 'core' genes that are enriched for relevant somatically mutated driver genes.

To better understand which processes are shared between the genes that are prioritised by Downstreamer we performed pathway analysis. We performed pathway analysis for the non-cancer adipose-breast network on Downstreamer-prioritised genes (5%

FDR-significant) against all other tested genes. Using a Fisher's exact test on Gene Ontology (GO) biological process terms showed a Bonferroni-significant association (p-value$<4.04\times10^{-6}$, n-tests=1,236) with Downstreamer-prioritised genes for the GO terms "regulation of transcription by RNA polymerase II" (p-value: 1.27×10^{-7}), "positive regulation of transcription by RNA polymerase II" (p-value: 4.06×10^{-8}) and "negative regulation of transcription by RNA polymerase II" (p-value: 2.18×10^{-6}). Performing the same analysis against GO molecular function terms found Bonferroni-significant enrichments (p-value$<1.18\times10^{-5}$, calculated by dividing 0.05 by 4,243 GO molecular function terms) for "DNA-binding transcription factor activity, RNA polymerase II-specific" (p-value: 3.50×10^{-8}) and "transcription coactivator activity" (p-value: 1.27×10^{-6}). For GO cellular component terms, there were two Bonferroni-significant hits (p-value$<2.85\times10^{-5}$, calculated by dividing 0.05 by 1,757 GO cellular component terms): "chromatin" (p-value: 1.21×10^{-6}) and "cornified envelope" (p-value: 7.33×10^{-6}). Although these significant GO terms could be relevant for cancer, we are aware that they are not specific to cancer. All the other significant GO terms for all combinations of GWAS and networks tested can be found in the supplementary data (Data S2).

Genes that are co-expressed with cancer GWAS associated-genes are enriched for being loss-of-function intolerant

Within the omnigenic model hypothesis we would expect 'core' genes as prioritised by Downstreamer to be depleted for germline variants that lead to diminished protein function. For the prioritised genes, we compared their probability of being intolerant to protein truncating variants (pLI) in gnomAD is higher than for other genes [36]. Comparing the Downstreamer gene-prioritisation scores against the loss-of-function Z-scores provided by gnomAD using a Pearson's correlation test highlighted a significant, but small correlation for the colorectal cancer GWAS using the non-cancer multi-tissue network (Pearson's r: 0.12, p-value: 2.31×10^{-47}, n: 18,874 protein coding genes) and for the prostate cancer GWAS using the cancer multi-tissue network (Pearson's r: 0.13, p-value: 2.50×10^{-55}, n: 18,874 protein coding genes; Fig. S11). When using tissue-specific networks matched by the cancer's tissue of origin, the gene-prioritisation score was significantly correlated to loss-of-function Z-scores for all combinations of cancer types and networks except the non-cancer skin network after Bonferroni-correction (p-value$<1.25\times10^{-3}$, calculated by dividing 0.05 by the 40 tested networks; Fig. S12).

Overall, these results suggest that genes associated to cancer traits by GWAS may be converging on regulated cancer drivers downstream within tissue-specific gene networks.

Discussion

In this study we sought to explain the limited overlap observed between genes identified through cancer risk GWAS and known somatic cancer driver genes. We collected well-powered GWAS summary statistics for four cancer traits and evaluated different types of co-expression networks to ascertain if somatic cancer driver genes are potentially regulated by genes inside GWAS cancer risk-loci. We propose that this holds for breast cancer, prostate cancer and colon cancer, particularly when using gene networks derived from same-tissue mRNA expression data (Fig. 3). These results indicate that tissue-specific gene networks matched on the tissue of origin of the tumour can be used to link common germline variants that are associated with cancer risk to somatic driver genes. We believe this has several implications.

First, although the omnigenic model has been described in the context of germline variation between common and rare variants, our observations suggest that the omnigenic model hypothesis might also apply between common germline variants and rare somatic variation. Genes inside cancer GWAS risk loci show relatively little overlap with known somatic cancer driver genes. However, we show that some cancer drivers are strongly co-regulated with genes in GWAS risk loci. Additionally, the lack of prioritization of some well-established cancer drivers warrants future investigation: although it is not yet clear why this happens, one potential explanation could be that some of these drivers could (1) Play a role at a later stage where somatic variants effects dominate over germline variants and therefore show little co-expression with genes in GWAS loci (2) Be less likely to independently initiate oncogenesis than other more highly prioritised drivers. One example is *KRAS*, an oncogene mutated in 40% of colorectal cancers [37]. *KRAS* is not prioritised by either PascalX or Downstreamer. Meanwhile *APC* shows a trend for PascalX (p-value 0.001 rank: 500 out of 18,874 protein coding genes) suggesting that *APC* inactivation can initiate oncogenesis earlier than *KRAS* activation in line with the current understanding of colorectal cancer oncogenesis [38].

Second, our approach provides a way to use germline risk variants and gene networks to identify potential novel cancer susceptibility genes, some of which may also have a cancer driver role. We show that prioritised genes are enriched for known somatic driver genes, although we recognise that not all the genes prioritised by our methodology can easily be placed in the context of these cancer types, and follow-up research is warranted to clarify upon their putative roles. This complements the strategies to search for cancer susceptibility genes, which are usually derived exclusively from cancer GWAS results and currently showing low overlap with cancer drivers

[12]. Future cancer subtype-specific analyses using the Downstreamer methodology might confirm if cancer drivers prioritised by Downstreamer correspond to the drivers observed in that specific cancer type. This can be useful for studying the origin of the different subtypes of cancer and support the use of polygenic scores for patient stratification.

Third, we also observed that cancer drivers are co-regulated with genes inside cancer-GWAS-associated loci in the context of a regulatory network derived from non-cancer tissues. This might suggest that cancer driver dysregulation could also occur before oncogenesis rewrites the co-expression relationships between genes. This is in line with the recent observation that non-cancer tissues harbour clonal populations of cells that have cancer driver mutations that endow these clones with a replicative fitness advantage [39].

Despite the lack of enrichment for cancer tissue derived networks, we cannot conclude that such networks are irrelevant for interpreting cancer GWAS since the lower observed enrichment in the cancer tissue derived networks might be due to technical factors. Firstly, the quality and number of cancer samples available to generate regulatory networks is lower than the corresponding set of non-cancer samples which potentially impacts power. Secondly, since expression of genes in cancer is highly heterogeneous, preparing cancer regulatory networks based simply on tissue of origin may not be ideal. For the same reason, the heterogeneity of gene expression in cancer samples increases the challenge of creating co-differentially expressed networks (differences in expression profiles between cancer and non-cancer samples). Care needs to be taken to ensure that the high heterogeneity does not result in a lack of power to identify co-expression differences between cancer and non-cancer samples. Future analyses with networks derived from more cancer samples, considering their heterogeneity, could more confidently determine if the heritability of cancer is explained by processes that occur in the context of the altered transcription of tumours.

Our study has several limitations. Although well-powered, the GWAS of cancer traits used in this study were performed based on a clinical cancer definition that may encompass multiple oncogenesis pathways. Moreover, current estimates suggest that current cancer GWAS only explain a fraction of heritability [40], therefore this analysis can still benefit from future GWAS with bigger cohorts and with well represented cancer subtypes. Additionally, current methodologies to link associated variants within GWAS loci to genes are also not definitive and do not take tissue type into account, and future improvements of these methods could help validate links between variants and genes for different tissues.

Despite sufficient statistical power, our co-expression networks can also further be improved upon. In the case of the creation of the adipose-breast co-expression network we were not able to reliably distinguish between the predicted tissue-type annotation of RNA-seq samples of adipose and breast tissue. Future work would benefit from the creation of a co-expression network that separates the adipose tissues from the other type of tissues present within the breast.

The majority of the RNA-Seq profiles used for calculating the co-expression networks are from samples with European ancestry. This matches with the ancestry of the samples used for the GWAS. Future studies with a larger number of RNA-Seq profiles based on non-European samples are likely valuable to take maximum advantage of GWAS performed on individuals of non-European descent.

In this study we limited ourselves to only prioritizing genes with a significant positive prioritization score, which is due to positive co-expression relationships of a gene with positional candidate genes inside the GWAS loci. Future work is required to allow our methodology to prioritise genes that show significant negative expression correlation with positional candidate genes inside the GWAS loci as well.

While our study demonstrates the potential of co-expression networks to link cancer driver genes to cancer GWAS genes, many questions remain. For instance, we are currently uncertain if this method can be used to identify driver genes for cancers that lack identified somatic driver genes. Additionally, we currently use co-expression networks derived from tumour tissue, however co-expression of genes in other cancer related cells such as immune cells could also be relevant for interpreting the cancer risk captured by GWAS.

As our unsupervised methodology does not rely on prior knowledge on functional gene-to-gene annotations, future work on the addition of a complementary, supervised methodology with additional information could help prioritise potential candidate cancer driver genes.

All in all, by using currently available methods, we show enrichment of cancer drivers in the Downstreamer prioritisations of core genes of some cancer-risk GWAS. In the future, with improved methodology and data, we expect that the identification of core genes for GWAS of different cancer types could lead to the identification of new cancer-susceptibility genes and cancer drivers.

Conclusions

Cancer risk-variants identified through GWAS are an important resource that could serve as independent validation of the observations made when studying somatic mutations in tumour tissue. For at least some cancer tissues, common cancer risk-variants can be linked with

somatic cancer driver genes using co-expression networks. This knowledge can contribute to our understanding of how cancer risk-variants contribute to the initiation and progression of cancers.

Methods

GWAS data

We downloaded publicly available harmonised summary statistics for 109 GWAS of cancer traits from the GWAS catalogue and from the BCAC official website (accession date: 18 October 2022) [41, 42]. In total, 24 different cancer traits were represented by at least one study. For each cancer, we first determined the number of independent loci in each GWAS using the TOP_HITS mode of Downstreamer 1.32. We then selected summary statistics with at least 90 independent loci as to assure for the use of GWAS with high enough statistical power. For every cancer trait, we selected the GWAS summary statistics derived from the study with the highest number of samples, which resulted in a final selection of four GWAS: prostate cancer, breast cancer, colorectal cancer and skin cancer (Table S1).

Gene eigenvectors for Downstreamer

Downstreamer uses eigenvectors with gene loadings to prioritise genes. We used two sources of RNA-seq data to obtain these expression eigenvectors: recount3 and TCGA.

recount3 (multi-tissue)

The recount3 resource that we use to create our multi-tissue and tissue-specific networks was collected by Wilks et al. [22]. This dataset contains 316,443 human RNA-seq samples that have been uniformly processed, technical covariate-corrected and quantified. In order to limit batch effects from the integration of different expression studies within this dataset we performed additional quality control (QC) and predicted if the samples are primary tissue, cell line or cancerous (see Note S1 for additional details). This allowed us to select 58,725 samples expected to be non-cancer primary tissues. We then predicted the tissues of origin for samples lacking annotations (Note S2). The additional QC per tissue reduced the total number of samples to 46,410, covering 57 different tissues or cell types. We then selected the breast (n=2,233), colon (n=1,289), prostate (n=386) and skin (n=1,302) samples for the tissue-specific networks relevant to the cancers we studied.

For the multi-tissue recount3 expression matrix (n=46,410) and the tissue-specific matrices, we used singular value decomposition on the per-gene-scaled expression data to obtain the eigenvectors with the gene loadings. For the full matrix, we selected the components that jointly explain 85% of the variance (n=848). For the tissues, we confined ourselves to components that explain 80% of the variance and have an eigenvalue of at least 1: breast (n=223), colon (n=139), prostate (n=63) and skin (n=141).

TCGA

We used the TCGA [43] data as quantified by the recount3 project [22] for the multi-cancer network and the cancer-specific networks. We performed the same QC as for the recount3 data (Note S1) but did not need to predict the tissue or cancer status because the TCGA data is extensively annotated. We confined ourselves to samples that passed QC and are annotated as `primary tumors` (n=8,739). Genes expressed in <50% of the selected samples were excluded. The samples were jointly variance stabilising transformed (VST) using tumour origin as the confounder. We then applied the same covariate correction that we applied to the recount3 data (Note S2). From this jointly normalised data, we extracted the tumour-specific subsets relevant to the cancers we studied: breast (n=1,085), colorectal (n=392), prostate (n=501) and skin (n=108).

In the same manner as for recount3, we calculated the eigenvectors using an explained variance threshold of 85% for primary tumour TCGA data, resulting in (n=1,864) components, and an 80% threshold for the tumour subsets: breast (n=161), colorectal (n=65), prostate (n=81) and skin (n=27).

Cancer-specific drivers

Cancer drivers were obtained from the COSMIC Cancer Gene Census [7], the IntOGen catalogue of gene drivers [5] and the drivers identified in Dietlein et al. [8]. The IntOGen catalogue of gene drivers (Release 2020.02.01) was obtained from the IntOGen official website [41]. The Dietlein et al. collection of coding and noncoding gene drivers was obtained from the supplementary tables of the original publication [8]. The collection of cancer genes from COSMIC (Version 96) was obtained from the official website in May 2022 [7]. We considered a gene to be a cancer-specific driver when it was called as a driver gene for that cancer type in at least one collection (Table S2). The gene symbols of these collections were then mapped to Ensembl IDs using the mapIDs formula of the R package "org.Hs.eg.db". We then manually curated the tissue-specificity of cancer drivers by inspecting the CANCER_TYPE column for IntOGen, the table title in Dietlein et al. and the 'Tumor Types (Somatic)' column for COSMIC CGC Genes were annotated to preferentially have oncogene or tumour suppressor activity based on their COSMIC CGC ("oncogene", "TSG"), or IntOGen ("Act", "LoF") labels.

Downstreamer

Determining linkage disequilibrium covariates

When performing a standard analysis using Downstreamer, results can be confounded by the linkage disequilibrium (LD) of genomic regions [19]. To overcome this, an LD covariate was prepared using European LD scores https://github.com/bulik/ldsc [44] (Table S5). For the LD covariate, we calculated the mean LD score of the variants with a minor allele frequency above 5% that were within 25 kb upstream and 25 kb downstream of protein-coding gene transcription start sites on autosomal chromosomes. Importantly, these scores were not significantly enriched for cancer driver genes (one-sided Wilcoxon rank-sum test, p-value=0.88), making this correction appropriate for studying cancer traits.

Component prioritisation

For each GWAS summary statistic and each co-expression network, we applied Downstreamer v1.32 to prioritise the components, with the following parameters: *genepruningR*: 0.8, *referenceGenotypes*: European non-Finnish 1000 genomes samples, *variantCorrelation*: 0.95, *window*: 25,000, *permutations*: 100,000, *permutationFDR*: 100, *permutationGeneCorrelations*: 10,000, *permutationsRescue*: 10,000,000, *geneCorrelationWindow: -1*, *permutationPathwayEnrichment*: 10,000, covariates: (see "Determining linkage disequilibrium covariates" section above) and the additional flags "--excludeHla –forceNormalGenePvalues --saveExcel --regress-gene-lengths". This resulted in a set of p-values and betas from the generalised least squares (GLS) regression model for every component.

Gene prioritisation

The gene-level prioritisation (G) was then derived by the following matrix vector product: $G=CB$, where C is the genes x FDR 5% significant components co-regulation matrix and B is the column-vector of component betas obtained from the GLS regression model. The vector G represents the Downstreamer prioritisations scores (Table S4).

PoPS

In a similar procedure as Downstreamer, we applied PoPS v0.2 for all combination of GWAS summary statistics and co-expression networks as described by Weeks et al. [31]. Firstly, we limited the inputs to protein coding genes in ensemble version 94 that were also present in the co-expression networks (used as the feature matrix for PoPS), the gene annotations file and the gene locations file (as used by MAGMA). The GWAS summary statistics were converted to gene-level p-values using MAGMA v1.10 as described in the documentation, using the 1000 genomes SNP location file provided by MAGMA. MAGMA was configured with the flag "--gene model snp-wise=mean" and the the flag "--pval" set to the number of cases of the GWAS [45]. We converted our co-expression networks to PoPS feature matrix format using the "Munge features" step of PoPS configured with the "--max_cols" flag set to 2000 (higher than the highest number of components observed in our co-expression networks) to ensure the creation of a single feature matrix per co-expression network. Finally, PoPs was run with the default parameters on the feature matrices and MAGMA output with the flag "--num_feature_chunks 1".

Pathway analysis

To better understand which diseases are associated with the top prioritised genes, we performed a pathway analysis of GO terms. We analysed all the combinations of different multi-tissue and tissue-specific gene-prioritisation methods applied to the cancer GWAS summary statistics separately. For this analysis, we focused on the 5%-FDR-significant genes from PascalX with the top 50 most significant prioritisation scores from the Downstreamer analysis. We downloaded the molecular function, cellular component and biological process databases (accession date/version: 1 June 2020) and filtered out any GO term with fewer than 10 annotated genes [46, 47]. We determined significant GO terms by performing a two-sided Fisher's exact test on the 5% FDR significance of Downstreamer genes versus inclusion of genes in GO terms.

Gene constraint analysis

To identify further characteristics of the genes prioritised by Downstreamer, we performed statistical tests on the Downstreamer gene score versus the probability of a gene being loss-of-function intolerant. For this statistical test, we performed a Pearson's correlation test on loss-of-function Z-scores provided by gnomAD [36]. A positive loss-of-function Z-score suggests intolerance to loss-of-function variants. Negative loss-of-function Z-scores indicate genes that have more loss-of-function variants than expected.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-024-01941-4.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5
Supplementary Material 6
Supplementary Material 7
Supplementary Material 8

Acknowledgements
We would like to thank Kate Mc Intyre for editorial assistance. We appreciate the UG Center for Information Technology and the UMCG Genomics Coordination Center, and their sponsors BBMRI-NL & TarGet, for the storage and computing infrastructure.The breast cancer genome-wide association analyses for BCAC and CIMBA were supported by Cancer Research UK (PPRPGM-Nov20\100002, C1287/A10118, C1287/A16563, C1287/A10710, C12292/A20861, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565) and the Gray Foundation, The National Institutes of Health (CA128978, X01HG007492 - the DRIVE consortium), the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, the Québec Breast Cancer Foundation, the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), the European Union's Horizon 2020 Research and Innovation Programme (634935 and 633784), the Post-Cancer GWAS initiative (U19 CA148537, CA148065 and CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund. All studies and funders are listed in Zhang H et al. (Nat Genet, 2020).

Author contributions
Conceptualisation: C.G.U.T., T.L., F.B., P.D., L.F. Data curation: C.G.U.T., K.D., M.Z., P.D. Formal Analysis C.G.U.T., T.L., F.B., P.D. Funding acquisition: P.D., L.F. Investigation: C.G.U.T., T.L., F.B., K.D., M.Z., P.D. Methodology: C.G.U.T., T.L., F.B., O.B.B., A.C., P.D., L.F. Software: C.G.U.T., T.L. F.B., P.D. Supervision: C.G.U.T., P.D., L.F. Visualisation: C.G.U.T., T.L., P.D., L.F. Writing – original draft: C.G.U.T., T.L., F.B., P.D., L.F. Writing – review & editing: W.Z., J.R., H.J.W. Roles as defined by: CRediT (Contributor Roles Taxonomy).

Funding
P.D. is supported by a Dutch Research Council (NWO) ZonMW-VENI grant (no. 9150161910057). L.F. is supported by a grant from the NWO (ZonMW-VICI 09150182010019 to L.F.), a European Research Council Starting Grant (grant agreement 637640 (ImmRisk)), and through a Senior Investigator Grant from the Oncode Institute and a grant from Saxum Volutum (Pericode).

Data availability
All GWAS summary statistics and RNA-seq data are publicly available. For further information, please see Note S1, Note S2 and Table S1. Information about cancer driver genes can be found in Table S2. PascalX results can be found in Table S3, and Downstreamer results can be found in Table S4.

Declarations

Ethics approval
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 May 2024 / Accepted: 18 June 2024
Published online: 15 July 2024

References

1. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: an overview. Int J Cancer. 2021;149(4):778–89.
2. Hanahan D. Hallmarks of Cancer: New dimensions. Cancer Discov. 2022;12(1):31–46.
3. Garber JE, Offit K. Hereditary cancer predisposition syndromes. J Clin Oncol off J Am Soc Clin Oncol. 2005;23(2):276–92.
4. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nat Reviews Cancer Nat Publishing Group. 2017;17:692–704.
5. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72.
6. Chernoff J. The two-hit theory hits 50. Mol Biol Cell. 2021;32(22):rt1.
7. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705.
8. Dietlein F, Wang AB, Fagre C, Tang A, Besselink NJM, Cuppen E, et al. Genome-wide analysis of somatic noncoding mutation patterns in cancer. Science. 2022;376(6589):eabg5601.
9. Malkin D. Li-fraumeni syndrome. Genes Cancer. 2011;2(4):475–84.
10. Wilcox N, Dumont M, González-Neira A, Carvalho S, Joly Beauparlant C, Crotti M et al. Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk. Nat Genet. 2023;1–5.
11. Porta-Pardo E, Valencia A, Godzik A. Understanding oncogenicity of cancer driver genes and mutations in the cancer genomics era. Febs Lett. 2020;594(24):4233–46.
12. Fanfani V, Citi L, Harris AL, Pezzella F, Stracquadanio G. The Landscape of the Heritable Cancer Genome. Cancer Res. 2021;81(10):2588–99.
13. Agarwal D, Nowak C, Zhang NR, Pusztai L, Hatzis C. Functional germline variants as potential co-oncogenes. Npj Breast Cancer. 2017;3(1):1–4.
14. Eckel-Passow JE, Decker PA, Kosel ML, Kollmeyer TM, Molinaro AM, Rice T, et al. Using germline variants to estimate glioma and subtype risks. Neuro-Oncol. 2019;21(4):451–61.
15. Carter H, Marty R, Hofree M, Gross A, Jensen J, Fisch KM, et al. Interaction landscape of inherited polymorphisms with somatic events in cancer. Cancer Discov. 2017;7(4):410–23.
16. Vali-Pour M, Lehner B, Supek F. The impact of rare germline variants on human somatic mutation processes. Nat Commun. 2022;13(1):3724.
17. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169(7):1177–86.
18. Ratajczak F, Joblin M, Hildebrandt M, Ringsquandl M, Falter-Braun P, Heinig M. Speos: an ensemble graph representation learning framework to predict core gene candidates for complex diseases. Nat Commun. 2023;14(1):7206.
19. Bakker OB, Claringbould A, Westra HJ, Wiersma H, Boulogne F, Võsa U et al. Linking common and rare disease genetics through gene regulatory networks [Internet]. 2021 Oct [cited 2022 Jan 7] p. 2021.10.21.21265342. https://www.medrxiv.org/content/https://doi.org/10.1101/2021.10.21.21265342v2.
20. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun. 2017;8:1077.
21. Krefl D, Brandulas Cammarata A, Bergmann S. PascalX: a Python library for GWAS gene and pathway enrichment tests. Bioinformatics. 2023;39(5):btad296.
22. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. Genome Biol. 2021;22(1):1–40.
23. Gerhardt J, Steinbrech C, Büchi O, Behnke S, Bohnert A, Fritzsche F, et al. The androgen-regulated calcium-activated nucleotidase 1 (CANT1) is commonly overexpressed in prostate cancer and is tumor-biologically relevant in vitro. Am J Pathol. 2011;178(4):1847–60.
24. Hermans KG, Bressers AA, van der Korput HA, Dits NF, Jenster G, Trapman J. Two unique novel prostate-specific and androgen-regulated fusion partners of ETV4 in prostate cancer. Cancer Res. 2008;68(9):3094–8.
25. Wang Z na, Liu D, Yin B, Ju W, yi, Qiu H, zhong, Xiao Y, et al. High expression of PTBP1 promote invasion of colorectal cancer by alternative splicing of cortactin. Oncotarget. 2017;8(22):36185–202.
26. Campion O, Thevenard Devy J, Billottet C, Schneider C, Etique N, Dupuy JW, et al. LRP-1 matricellular receptor involvement in Triple negative breast Cancer Tumor Angiogenesis. Biomedicines. 2021;9(10):1430.
27. Haque MA, Abdelaziz M, Puteri MU, Vo Nguyen TT, Kudo K, Watanabe Y, et al. PMEPA1/TMEPAI is a unique tumorigenic activator of AKT promoting proteasomal degradation of PHLPP1 in Triple-negative breast Cancer cells. Cancers. 2021;13(19):4934.
28. Yang XL, Liu KY, Lin FJ, Shi HM, Ou ZL. CCL28 promotes breast cancer growth and metastasis through MAPK-mediated cellular anti-apoptosis and pro-metastasis. Oncol Rep. 2017;38(3):1393–401.

29. Gong Y, Liu Z, Yuan Y, Yang Z, Zhang J, Lu Q, et al. PUMILIO proteins promote colorectal cancer growth via suppressing p21. Nat Commun. 2022;13(1):1627.
30. Gu L, Frommel SC, Oakes CC, Simon R, Grupp K, Gerig CY, et al. BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence. Nat Genet. 2015;47(1):22–30.
31. Weeks EM, Ulirsch JC, Cheng NY, Trippe BL, Fine RS, Miao J, et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nat Genet. 2023;55(8):1267–76.
32. Sato G, Shirai Y, Namba S, Edahiro R, Sonehara K, Hata T, et al. Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis. Nat Commun. 2023;14(1):3671.
33. Yang J, Liu X, Huang Y, He L, Zhang W, Ren J, et al. TRPS1 drives heterochromatic origin refiring and cancer genome evolution. Cell Rep. 2021;34(10):108814.
34. Scott TG, Sathyan KM, Gioeli D, Guertin MJ. TRPS1 modulates chromatin accessibility to regulate estrogen receptor (ER) binding and ER target gene expression in luminal breast cancer cells. BioRxiv Prepr Serv Biol. 2023;2023.07.03.547524.
35. Légaré S, Cavallone L, Mamo A, Chabot C, Sirois I, Magliocco A, et al. The estrogen receptor cofactor SPEN functions as a tumor suppressor and candidate biomarker of drug responsiveness in hormone-dependent breast cancers. Cancer Res. 2015;75(20):4351–63.
36. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
37. Meng M, Zhong K, Jiang T, Liu Z, Kwan HY, Su T. The current understanding on the impact of KRAS on colorectal cancer. Biomed Pharmacother. 2021;140:111717.
38. Levine AJ, Jenkins NA, Copeland NG. The roles of initiating truncal mutations in human cancers: the order of mutations and Tumor Cell type matters. Cancer Cell. 2019;35(1):10–5.
39. Martincorena I. Somatic mutation and clonal expansions in human tissues. Genome Med. 2019;11(1):35.
40. Zhang YD, Hurson AN, Zhang H, Choudhury PP, Easton DF, Milne RL, et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. Nat Commun. 2020;11:3353.
41. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–12.
42. Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet. 2020;52(6):572–81.
43. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
44. Bulik-Sullivan BK, Loh PR, 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(3):291–5.
45. De Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. Tang H, editor. PLOS Comput Biol. 2015;11(4):e1004219.
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
47. The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224(1):iyad031.

Publisher's Note