Distal gene regulation mediated by non-coding RNAs contributes to germline risk for breast and prostate cancer

Nolan Cole[1], Paige Lee[2], Tommer Schwarz[3,4], Pan Zhang[5], Matthew L. Freedman[6,7], Alexander Gusev[6], Sara Lindström[10,11], Michael J. Gandal[5,12,13], Bogdan Pasaniuc[4,12,13], Arjun Bhattacharya[4,14]*

1. Department of Statistics, Brigham Young University, Provo, UT, United States
2. Department of Statistics, University of California, Los Angeles, CA, United States
3. Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, United States
4. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, United States
5. Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
6. Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana Farber Cancer Institute, Boston, MA, USA
7. The Center for Cancer Genome Discovery, Dana Farber Cancer Institute, Boston, MA, USA
8. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
9. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
10. Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA
11. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
12. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
13. Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
14. Institute of Quantitative and Computational Biosciences, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

*Corresponding author: Arjun Bhattacharya (abtbhatt@ucla.edu)

ABSTRACT

Genome-wide association studies (GWAS) have identified numerous genetic loci associated with breast and prostate cancer risk, suggesting that germline genetic dysregulation influences tumorigenesis. However, the biological function underlying many genetic associations is not well-understood. Previous efforts to annotate loci focused on protein-coding genes (pcGenes) largely ignore non-coding RNAs (ncRNAs) which account for most transcriptional output in human cells and can regulate transcription of both pcGenes and other ncRNAs. Though the biological roles of most ncRNAs are not well-defined, many ncRNAs are involved in cancer development. Here, we explore one regulatory hypothesis: ncRNAs as *trans*-acting mediators of gene expression regulation in non-cancerous and tumor breast and prostate tissue. Using germline genetics as a causal anchor, we categorize distal (>1 Megabase) expression quantitative trait loci (eQTLs) of pcGenes significantly mediated by local-eQTLs of ncRNAs (within 1 Megabase). We find over 300 mediating ncRNAs and show the linked pcGenes are enriched for immunoregulatory and cellular organization pathways. By integrating eQTL and cancer GWAS results through colocalization and genetically-regulated expression analyses, we detect overlapping signals in nine known breast cancer loci and one known prostate cancer locus, and multiple novel genetic associations. Our results suggest a strong transcriptional impact of ncRNAs in breast and prostate tissue with implications for cancer etiology. More broadly, our framework can be systematically applied to functional genomic features to characterize genetic variants distally-regulating transcription through *trans*-mechanisms.

INTRODUCTION

Genome-wide associations studies (GWAS) of cancer risk have revealed risk-associated alleles at hundreds of distinct genetic loci, with breast and prostate cancer GWAS yielding the largest number of associations[1,2]. Through integration with transcriptomic and other functional genomics datasets, the proposed target genes for many of these risk loci have been found to overlap with coding regions of the genome[1–4]. However, many risk variants fall in non-coding regions of the genome and, for these variants, identifying the most likely mechanism is challenging. One proposed mechanism for GWAS-identified risk variants is *trans*-acting pathways: a GWAS variant could affect a regulatory feature, like a transcription factor, in proximity, which results in a cascade of effects on genes located far away from the GWAS variant. Particularly, one previous study identified GWAS risk variants for breast cancer that confer *trans*-effects through transcription factors, like *ESR1*, *MYC*, and *KLF4*[5]. Another potential regulatory feature by which GWAS variants in non-coding regions affect risk is mediation of *trans*-acting effects of genetic variants through non-coding RNAs (ncRNAs).

Although ncRNAs do not code for proteins, they account for nearly 60% of transcriptional output in human cells[6] and interact with a complex network of genes, transcripts, and proteins with widespread effects on cell biology[7]. Some ncRNAs, like microRNAs (miRNAs), target the 3' Un-Translated Region (UTR) of specific mRNA transcripts of genes, leading to their degradation, and naturally link to regulatory networks that include multiple ncRNAs and protein-coding genes (pcGenes). miRNAs and other regulatory ncRNAs are often targeted by circular or long ncRNAs to stabilize the transcripts within cells, in processes like sequestration. These interactions between ncRNAs and pcGenes are complex, supporting the hypothesis that ncRNAs have key roles in cellular pathways[7–9]. In fact, ncRNAs can serve as master regulators in cancer[10,11], like the long ncRNA *XIST*, which exerts oncogenic and metastatic effects in multiple cancer types[12]. Specific ncRNAs have been shown to leave the site of their transcription and operate in *trans*-

acting pathways by regulating gene expression at regions distal to their transcription start site, influencing nuclear structure and organization and regulating the behavior of RNA molecules[8,9]. However, regulatory impacts of ncRNAs on transcription of pcGenes are generally uncategorized, especially in a systematic fashion[8].

ncRNAs have shown associations with the onset and progression of different cancers, are enriched in multiple tumor types, and are even therapeutic targets, as they act as regulators of genes in important tumorigenic or progressive networks[10]. Profiling and deep sequencing of ncRNAs have shown that perturbing ncRNA biogenesis affects amplification, deletion, and normal epigenetic and transcriptional regulation[10,11,13,14]. Many ncRNAs act as oncogenes or antagonize tumor suppressors, but as most ncRNA mechanisms in cancer tumorigenesis or progression have been categorized on a case-by-case basis[13,14], mechanistic impacts of ncRNAs have not been explored systematically. Bioinformatics analyses that leverage high-throughput genomics have investigated the role of ncRNAs through mainly computational target prediction or differential expression analyses. Although these computational methods have improved our understanding of the role of ncRNAs in cancer, they have major limitations, including computational feasibility and functional translation of sequence similarity methods[15] and reverse causality for differential expression analyses (i.e. differential expression more likely reflects consequences of disease)[16].

One systematic approach to identifying potential *trans*-mechanisms of regulation is to use genetic variants as causal anchors. A prevailing thought is that distal expression quantitative trait loci (eQTLs) of genes, where the genetic variant is far away from the gene (more than 1 Megabase, or Mb), are often themselves local-, or *cis*-acting, QTLs of a regulatory feature[17–22]. We emphasize that the modifiers “local” and “distal” refer merely to distances in the genome (i.e., within or outside 1 Mb), whereas *cis*- and *trans*-acting refer to the biological mechanism

(i.e., direct or indirect interaction). Molecular features, like ncRNAs, that have potential *trans*-acting, or indirect, regulatory effects can be identified, by either SNP-level eQTL and subsequent rigorous mediation analysis[17,21,23] or by transcript-level mapping by using methods borrowing from two-step regression[24]. Not only can these methods point to distal mechanisms between ncRNAs and pcGenes, but they can point to genetic variants associated with disease etiology that potentially have distal effects in the transcriptome.

We employ these methods to systematically map distal-eQTLs of pcGenes that are potentially mediated by local-eQTLs of ncRNAs in non-cancerous and tumor prostate and breast tissue, using data from the Genotype Tissue-Expression (GTEx) Project[25] and The Cancer Genome Atlas (TCGA)[26], respectively. We then employ colocalization[27] and genetically-regulated expression analysis[28] to identify potential overlaps in eQTLs and GWAS signals for overall prostate[1] and both overall and molecular subtype-specific breast cancer risk[2]. In total, our work shows the widespread transcriptomic impact of genetically-mediated portion of ncRNAs and that this impact has key associations with cancer etiology. This approach provides a rigorous framework to not only categorize functional hypotheses of distal regulatory effects of ncRNAs but also other regulatory molecular features.

RESULTS

Overview of methods

In this work, we uncover hidden mechanisms contributing to genetic risk for breast and prostate cancer mediated by ncRNAs, systematically exploring one particular regulatory hypothesis: distal mediation of pcGene expression regulation in non-cancerous and tumor breast and prostate tissue. Specifically, the goal of our analysis is to identify distal-eQTLs of protein-coding genes (pcGenes) in non-cancerous and tumor samples, in both breast and prostate tissue, that are significantly mediated by ncRNAs local to these distal-eQTLs. Here, we define a local-eQTL

as a variant within 1 Mb of the gene body and a distal-eQTL as a variant outside the 1 Mb window of the gene body. eQTLs outside this 1 Mb window are unlikely to have direct effects on the promoters or enhancers of the gene and are more likely to have *trans*-acting mechanisms[17,18]. We approach this analysis with two complementary methods (**Methods, Figure 1, Supplemental Figure S1**): (1) formal mediation analysis[17,21,23] and (2) gene-based association testing (GBAT), followed by fine-mapping[24,29]. In formal mediation analysis, for a given pcGene, we first identify distal-eQTLs of the pcGene that are also local-eQTLs of ncRNAs. We then test, through permutation, the magnitude of the total mediation effect (TME) of the distal-eQTL on the pcGene through a set of local-ncRNAs. In GBAT, we first estimate, through leave-one-out cross-validation, the genetically-regulated expression (GReX) of an ncRNA. We then estimate the association of the GReX of the ncRNA with distal-pcGenes. Lastly, to identify a set of SNPs that best explains the distal association, we employ fine-mapping to designate a 90% credible set of distal-eQTLs. We employ European-ancestry patient-derived samples of non-cancerous prostate (N = 186) and breast (N = 337) tissue from GTEx[25] and prostate (N = 349) and breast (N = 437) tumor tissue from TCGA[26]. Finally, to detect any overlap between ncRNA-mediated distal-eQTLs of pcGenes and GWAS signal for cancer risk, we conduct colocalization tests and analyses of the GReX of ncRNAs.

ncRNAs have widespread distal transcriptional associations in breast and prostate tissue

Distal-eQTL mapping through mediation analysis

We first conducted distal-eQTL mapping (SNP and gene are more than 1 Mb away) through ncRNA mediation in GTEx[25] and TCGA[26]. Here, as in previous distal-eQTL mapping work[21,23], we considered testing triplets consisting of a SNP, a pcGene, and a set of ncRNAs, such that: the SNP and protein-coding gene form a distal-eQTL at nominal $P < 10^{-6}$, and the SNP forms local-eQTLs with all ncRNAs in the set at nominal $P < 10^{-6}$. As in previous distal-eQTL studies,

we use a liberal P-value threshold to increase the number of testing triplets subjected to rigorous permutation-based mediation analysis[18,21,23]. We then compute the total mediation effect (TME) of the SNP on the pcGene through the set of ncRNAs and test its magnitude using a permutation test at Benjamini-Hochberg false discovery rate adjusted $P < 0.05$ (**Methods**) [21,30,31]. The number of distal-eQTLs of pcGenes that are significantly mediated by ncRNAs are reported in **Table 1** (full list in **Supplemental Table S1-2**), across non-cancerous and tumor breast and prostate tissue. Distributions of TME and mediation proportion (the ratio of the TME to the sum of the TME and the direct distal-eQTL effect) in tumor tissue show a larger range than in non-cancerous tissue for both breast and prostate, and the median TME is higher in tumor tissue and median mediation proportion (**Supplemental Figure S2**).

In non-cancerous breast tissue, *PUM1* and *PCBP1* are among pcGenes with cross-chromosomal distal-eQTLs with large mediated effects through ncRNAs (**Figure 2A**), both of which have profound impacts on tumorigenesis[32–34]. In breast tumors, we find that *MT4* and *GPRC6A* have large mediated distal effects; *MT4* is a part of the metallothionein family with roles in breast cancer carcinogenesis[35,36], and *GPRC6A* is involved in the androgen receptor signaling pathway[37,38]. In non-cancerous prostate tissue, we detect large distal mediation effects for genes such as *RAF1*, a proto-oncogene[39], *LGR5*, a gene associated with prostatic regeneration and overexpressed in prostate tumors[40], and *SAXO2*, a microtubule stabilizer that has been targeted for cancer therapies[41,42]. We also detect a number of chemical stimulus and olfactory receptors (*OR2T1*, *OR10G8*, *OR10S1*) with large mediated effects in prostate tumor; olfactory receptors have shown some involvement with prostate cancer progression, though they generally have low expression in prostate tumors[43,44].

Probably due to a larger set of nominally-significant distal-eQTLs in the tumor tissues and larger sample sizes in TCGA than GTEx, we find nearly four times as many significantly mediated

distal-eQTLs in tumor tissue compared to non-cancerous tissue. In these tumor-enriched mediated distal-eQTLs, we observed that multiple ncRNAs mediated distal-eQTLs with many different pcGenes, leading to vertical bands in the location plots in **Figure 2A**. One example of such an ncRNA in breast tumors is *LINC00301*, showing significant mediation of 1,103 distal-eQTLs, comprising SNP-pcGene pairs across of 53 unique SNPs and 53 unique pcGenes). Many of these pcGenes belong to the GAGE protein family, which promotes breast cancer cell invasion and has shown evidence of distal genetic regulation[45,46]. *LINC00301* itself has been implicated in facilitating tumor progression and immune suppression, albeit in lung cancer[47]. Another example of such an ncRNA in breast tumors is *miR-548f-4*, a commonly mutated microRNA in multiple cancers[48], which mediated more than 300 distal-eQTLs for 16 unique pcGenes. These pcGenes include *GUCA2B*, a gene upregulated in breast cancer metastases[49], and *CYP2C9*, a gene suppressed by the breast cancer drug tamoxifen[50]. In contrast, only two ncRNAs in non-cancerous breast tissue showed significant mediation of more than 50 distal pcGenes. One of these genes, *FAM106A*, mediates distal-eQTLs of 13 pcGenes, the majority with *CCDC144A* and *BTN3A2*, the latter of which is a prognostic marker for breast cancer[51]. Across non-cancerous and tumor breast tissue, three pcGenes (*PRCC1*, *CYP2C9*, and *ATG14*) showed significant mediation through ncRNAs, though the sets of ncRNAs for these three genes are distinct.

In prostate tumors, *LINC02903*, known to harbor SNPs that interact with known prostate cancer risk SNPs[52], showed significant mediation with the highest number of distal-eQTLs (177 eQTLs across 8 pcGenes). These pcGenes include *FABP9*, an upregulated gene in prostate carcinomas with prognostic value[53], and *MTNR1B*, a gene harboring nominal risk variants for prostate cancer[54]. Another ncRNA with significant TME for nearly 100 distal-eQTLs across 20 pcGenes is *SDHAP2*. Many of these pcGenes have been implicated in prostate cancer and metastasis pathways, including *TMEM207*, *FADS6*, *MTNR1B, SLC26A8,* and *FGF23*[55–57].

Similar to breast tissue, only two genes showed significant mediation of more than 20 distal eQTLs in prostate tissue: *FBXO30-DT*, a divergent transcript, and *SNHG2*, which has been implicated in tumorigenesis and proliferation in multiple cancer indications[58–60]. The overwhelming majority (22/26) of these distal-eQTLs are for *OVCH2*, which has been implicated in prostate risk through GWAS[61,62]. The majority of distal-eQTLs mediated by *SNHG2* are for *RAF1*, a therapeutic target for multiple cancers[39]. We did not detect any shared ncRNAs or pcGenes across non-cancerous and tumor prostate tissue.

Gene-based distal-eQTL mapping

Next, we conducted gene-level distal eQTL mapping using GBAT[24] in non-cancerous and tumor breast and prostate tissue to identify ncRNAs that are regulated by multiple weak local genetic effects and may have distal effects on pcGenes; these ncRNAs are likely to be missed by the mediation framework outlined previously. We then use SuSiE fine-mapping to identify a set of SNPs in the locus that explain the distal genetic association[29]. Comparing to mediated pcGenes from mediation analysis (**Table 1**), we find a similar order of magnitude of pcGenes with distal associations with ncRNAs using GBAT (**Table 2**). Similar to results from mediation analysis, we find far more distal ncRNA-pcGene directional associations in tumor compared to non-cancerous tissue, as shown in **Table 2** (full list in **Supplemental Table S1-2**). In addition, we find that the distribution of ncRNA-pcGene effect sizes is shifted downwards in tumor tissue, compared to non-cancerous tissue, though the number of effect sizes in these distributions are far less for non-cancerous tissue distal associations (**Supplemental Figure S2**). In breast tumors, we find large distal genetic associations with pcGenes like *OXNAD1*, an RNA-binding protein that is associated with pan-cancer survival rates and involved with tumor invasion and metastasis[62,63], and *LCN9*, belonging to the lipocalin family which promotes breast cancer metastasis[64]. In prostate tumors, we detected multiple large distal genetic associations with

genes in the pro-proliferative keratin-associated protein (KRTAP) family (*KRTAP13-3*, *KRTAP13-4*, *KRTAP10-8*)[65] (**Figure 2**).

Though we did not find pervasive distal genetic effects of ncRNAs in non-cancerous breast or prostate tissue, we found three ncRNAs with genetic associations with pcGenes across both breast and prostate tissue: *LINC01678*, *FAM106A*, and *AP001056.1*. These ncRNAs also target the same pcGenes (*LOC102724159*, *CCDC144A*, *GATD3B*, and a paralog to *TRAPPC10*), though these pcGenes do not have catalogued functions in cancer tumorigenesis or progression.

Again, in tumor-specific gene-gene associations, we find that many ncRNAs have associations with multiple pcGenes, leading to vertical bands in the location plots in **Figure 2B**. For example, in breast tumors, *LINC000906*, with predicted activity as an miRNA sponge in breast tumors and correlated with overall survival in breast cancer patients, was associated with 115 pcGenes, the largest association with *MS4A5*, a gene whose hypomethylation is shown to be prognostic for multiple cancer types[66,67]. *LINC00115*, a known promoter of breast and ovarian cancer metastasis and progression[68–70], is another example of an ncRNA associated with multiple different pcGenes, many of which are interferon-related genes (*IFNA17* and *IFNW1*), related to immune system cytotoxicity (*RAC2* and *DDB2*), or coding for secretory proteins (*PRH1* and *PRH2*).

In prostate tumors, *FAM138F* showed distal genetic associations with 57 distinct pcGenes, many of which are involved in amino acid activation in prostate cancer (*HDAC2*, *NT5DC1*, *NUS1*, and *PREP*) and protein stability (*PDCD2*, *TCP1*). Another ncRNA with multiple distal genetic associations with pcGenes is *TDRG1*, shown to be associated with progression and metastases in multiple cancers[71,72]. The associated pcGene targets of *TDRG1* include the

cancer-initiating pluripotency factor *PRDM14*[73] and multiple genes related to olfactory stimulus (*OR10H3*, *OR2T1*, *OR51V1*, and *UGT2A1*), which are associated with prostate cancer progression[44]. Taken together, these distal eQTL mappings suggest the ncRNAs have strong influences on gene expression of multiple pcGenes in both non-cancerous and tumor tissue.

Overlap of miRNA-pcGene pairs with target prediction databases

To provide a level of in-silico validation, we queried TargetScan[74,75], a database that curates computationally predicted RNA targets of miRNAs, for any miRNAs that our analysis detected to mediate distal-eQTLs of pcGenes. We only queried found miRNA-mediated distal-eQTLs of pcGenes in non-cancerous and tumor breast tissue. Out of 522 pairs of miRNAs and pcGenes across 72 unique miRNAs identified through our analysis, we found that 184 pairs were included in the TargetScan database (**Supplemental Table S3**). miRNA-pcGene pairs identified in our eQTL analysis are found in TargetScan at an enrichment ratio of 8.2 (95% CI: [6.68, 10.09]), compared to the universe of all miRNA-target pairs in TargetScan (approximately 3.56 million) and roughly 159,000 miRNA-target pairs in TargetScan for the 72 miRNA families identified. A large majority of these miRNAs (82% of miRNAs detected in eQTLs) are only conserved in humans and mice but well-annotated. Though this intersection with TargetScan does not implicate the miRNA in distal regulation of the proposed pcGene, it provides some computational validation with different methodology (sequence similarity vs. eQTL mapping).

Distal-eGenes are enriched for gene pathways involved in tumorigenesis and cancer progression

To assess enriched biological processes or pathways by sets of prioritized pcGenes, we conducted gene ontology enrichments[76]. Overall, compared to the universe of expressed pcGenes in the transcriptome, pcGenes with ncRNA-mediated distal eQTLs in non-cancerous tissue (combining breast and prostate) were significantly enriched (FDR-adjusted $P < 0.05$) for

many relevant ontologies: immune processes, genes targeted by epigenetic regulation, microRNA targets in cancer, and oxidoreductase activity. In comparison, pcGenes detected in ncRNA-mediated distal-eQTLs in tumor tissue showed enrichments mainly for detection and perception to chemical and sensory perception and genes involved in and localized in intermediate filament cytoskeleton (**Supplemental Figure S3-4**). Comparing tissue-prioritized pcGenes (breast-specific or prostate-specific pcGenes to the protein-coding transcriptome), we found pcGene enrichments detected in prostate tissue for immune pathway ontologies, including the multiple activations of immune cells, and known tumorigenic pathways, like the JAK-STAT cascade and PI3K-Akt signaling[77,78]. We did not detect any significant enrichments for the breast-specific pcGenes (**Supplemental Figure S5**).

We also conducted comparisons of pcGenes with ncRNA-mediated distal eQTLs between non-cancerous and tumor state in both breast and prostate tissue (**Figure 3**). We find that, compared to pcGenes prioritized in breast tumors, non-cancerous breast pcGenes were enriched for cytokine and leukocyte production, response, and function, as well as membrane transport and binding. We observed similar enrichments when we compared non-cancerous prostate pcGenes to those from prostate tumors, with additional cell death and morphogenesis ontologies enrichments. In contrast, across both breast and prostate tissue, tumor-specific pcGenes, compared to non-cancerous pcGenes, mainly showed enrichments for olfactory and chemical stimulus response, intermediate filament cytoskeleton localization, and epidermis development. These ontologies are consistent with cancer progression, as olfactory receptors have been validated as prognostic biomarkers in prostate cancers and are overexpressed in more aggressive breast tumors[43,44,79,80], more aggressive breast cancers are enriched for genes that influence epidermal growth[81] and cytoskeletal dysregulation is key to cancer cell invasion, progression, and metastasis[82–84].

Mediated distal-eQTL signal in breast and prostate overlaps with genetic signal with cancer risk

Lastly, we integrated these eQTL results with GWAS summary statistics for overall prostate and overall and molecular subtype-specific breast cancer risk[1,2]. We studied five intrinsic breast cancer molecular subtypes, defined by combinations of estrogen (ER)-, progesterone (PR)-, and human epidermal growth factor receptor (HER2) and tumor grade (REF): Luminal A-like (ER + and/or PR + , HER2-, grade 1 or 2); (2) luminal B-like/HER2-negative (ER + and/or PR + , HER2-, grade 3); (3) luminal B-like/HER2-positive (ER + and/or PR + , HER2 +); (4) HER2-positive/non-luminal (ER- and PR-, HER2 +), and (5) TNBC (ER-, PR-, HER2-). Of the ncRNAs that mediate at least one distal-eQTL of a pcGene, we find 84 ncRNAs are within 0.5 Mb of a GWAS SNP at $P < 5 \times 10^{-8}$, the majority for overall prostate and breast cancer risk, which have the largest GWAS sample sizes (**Figure 4A, Supplemental Table S1-2**). For these ncRNAs and any pcGenes we found in mediated distal-eQTLs, we conducted Bayesian colocalization analysis using *coloc* to estimate the probability that the eQTL signal and GWAS signal can be explained by a shared SNP, denoted PP.H4[27]. We conducted independent colocalization analyses for the local-eQTL signal and the distal-eQTL signal.

Among pairs of ncRNAs and pcGenes where the ncRNA is 0.5 Mb from a GWAS SNP for either breast or prostate cancer risk, we identified 30 pairs (15 for overall breast cancer and 10 for LumA breast cancer, five for prostate cancer) where at least one of the ncRNA local-eQTL or pcGene distal-eQTL colocalized with the GWAS signal at the locus with PP.H4 ≥ 0.75 (**Supplemental Figure S6**, **Supplemental Table S4**). In sum, we provide computational evidence that 10 independent overall or LumA-specific breast cancer risk GWAS signals and one independent prostate cancer risk GWAS signal may be, in part, explained by distal genetics effects on pcGenes mediated by local-ncRNAs.

We found that local-eQTLs of *RFPL1S* (though only reaching associations with P $\approx 10^{-4}$) in prostate tumor colocalized with prostate cancer risk at PP.H4 = 0.865; this ncRNA is also implicated in survival of other cancers and is hypothesized to regulate other genes in the RFLP gene family[85–87]. We find strong colocalization of this locus with distal-eQTLs of *CSH1* with PP.H4 = 0.913 (**Figure 4B**). *CSH1* codes for a somatotropin hormone, belonging to a family of hormones with paracrine signaling functions promoting cell division and growth in glands[88]. Using non-cancerous breast tissue eQTLs, we find strong colocalization with overall breast cancer risk with local-eQTLs of *RUSC1-AS1* and distal-eQTLs of *SH2B1* (**Figure 4C**). Two different SNPs carry the largest posterior probability of colocalization for the local- (rs2297480) and distal-eQTLs (rs2075571), these two SNPs are in moderate, yet statistically significant, linkage disequilibrium (D' = 0.51, R^2 = 0.131, P < 0.001). *RUSC1-AS1* has shown evidenced silencing of genes on different chromosomes through epigenetic signaling and is correlated with breast cancer progression[89,90]. Additionally, *SH2B1* is involved with cytokine and growth factor receptor signaling in cell proliferation and migration[91]. We also find strong colocalization with LumA-specific breast cancer risk with local-eQTLs of *THBS3-AS1* and distal-eQTLs of *SLC39A13* in non-cancerous breast tissue. Interestingly, the same SNP shows the maximum posterior probability for colocalization with both the local- and distal-eQTL signal. Though the ncRNA has not been implicated in cancer risk or progression, the pcGene belongs to a family of solute carrier proteins with catalogued associations in breast cancer[92]. In fact, *SLC39A13* facilitates metastasis in ovarian cancer by activating the Src/FAK signaling pathway, a cascade that affects epithelial cell migration[93].

For overall breast cancer risk, we also detect strong colocalization between eQTLs and GWAS signals at the 17q21.31 locus. This locus houses a large and common inversion polymorphism with associations with breast and ovarian cancer prognosis[94–96], as well as widespread associations with multiple phenotypes, including neuropsychiatric traits, Parkinson's disease,

and metabolic traits[97–100]. In particular, we found that the ncRNA *KANSL1-AS1* potentially mediates distal-eQTLs of multiple pcGenes. All but one of these pcGenes are on Chromosome 17, near the end of the 17q21.31 region; the last pcGene we detected is *TXNRD3* at 3q21.3. We were interested in disentangling effects of the H2 inversion on gene expression and breast cancer risk, with the proposed causal diagram presented in **Figure 5A**. First, we estimated haplotypes of the H2 inversion in GTEx[101]. We found that *KANSL1-AS1* and the associated pcGenes near 17q21.31 all have significant associations with the H2 inversion (**Figure 5B**). Next, we reran mediation analysis for SNPs local to *KANSL-AS1* and these detected pcGenes. After accounting for H2, the distal mediation signal of *KANSL-AS1* is attenuated for all pcGenes except *TXNRD3*, suggesting that the inversion may be driving a significant portion of this signal. These analyses support our proposed causal model, that the genetics of the H2 inversion affect expression of the local ncRNA KANSL-AS1 and the distal pcGenes at the end of the inversion, which induces the eQTL associations (**Figure 5A**).

H2 inversion-adjusted colocalization analyses also support this model. A salient example is shown in **Figure 5D**, with Manhattan plots from colocalization analysis of overall breast cancer risk, local-eQTLs of *KANSL-AS1*, and distal-eQTLs of *CRHR1*. Colocalization analysis unadjusted for H2, showed nearly-perfect colocalization for both eQTL signals with GWAS (PP.H4 > 0.98 for both eQTLs), but after adjustments, the eQTL signals are completely removed. In fact, the standardized effect sizes for local-eQTLs of *KANSL1-AS1* showed strong correlation with the standardized effect sizes for local-QTLs of the H2 inversion (**Figure 5E**), and the local-QTL signal for the H2 inversion strongly colocalized (PP.H4 = 0.99) with the GWAS signal in the locus (**Figure 5F**). These results illustrate that the structural inversion in the locus likely accounts for both *cis*-genetic control of *KANSL1-AS1* and the associated pcGenes further downstream on Chromosome 17. Furthermore, we emphasize further examination of the H2 inversion and other structural variants in the locus for its impact on local and distal gene

expression, as well as cancer susceptibility, especially to elucidate if SNPs with widespread local or distal associations with gene expression are affected by confounding due to structural variants or other aberrations.

We also studied the cancer risk associations of ncRNAs with predicted large transcriptomic effects using a genetically-regulated expression (GReX) approach. For ncRNAs with significant mediation of distal-eQTLs of multiple pcGenes, we train predictive models of expression in GTEx and TCGA and conduct ncRNA-level disease mapping using the weighted burden test and permutation test from FUSION[28]; this permutation test is critical, as it quantifies the significance of the expression-trait associations conditional on the SNP-trait effects at the locus. At $P < 2.5 \times 10^{-6}$ and permutation $P < 0.05$, we identify 33 ncRNA-level associations (**Figure 6**, **Supplemental Table S5**), predominantly for overall and subtype-specific breast cancer risk. Only one ncRNA, *SDHAP2*, showed an association with overall prostate cancer risk (**Figure 6**); as we restrict to ncRNAs mediating distal-eQTLs and not necessarily in known prostate cancer GWAS loci, we do not recover any ncRNAs detected by Guo et al's analysis of long-ncRNAs in prostate cancer[102]. A couple ncRNAs prioritized in associations of breast cancer risk using ncRNA GReX in non-cancerous breast tissues have been previously noted. Common non-synonymous SNPs in *HCG9* have previously been implicated in breast cancer GWAS[103]. In addition, *CEROX1*, a cataloged post-transcriptional regulator of mitochondrial catalytic activity, has been implicated in distal alterations of metabolic pathways in breast cancers[14,104]. ncRNAs prioritized in breast tumor GReX-associations with breast cancer risk mainly included micro- and snoRNAs. Of these, *miR-519d* has been shown to suppress breast cancer cell growth by targeting distal molecular features[105,106]. GReX associations prioritize these ncRNAs for further investigation of their functional effects in breast and prostate tissue.

DISCUSSION

In this work, we systematically map distal-eQTLs of protein-coding genes (pcGenes) that are mediated by non-coding RNAs (ncRNAs) in prostate and breast tissue. We identify multiple ncRNAs such that local genetic architecture associated with ncRNA expression potentially mediates pcGenes far away from the ncRNA. We then show that many of these ncRNA-mediated distal-eQTLs of pcGenes overlap with GWAS signals for breast and prostate cancer. Through colocalization analysis, we show evidence for strong overlap with ncRNA expression with nine breast cancer GWAS loco (excluding 17q21.31) and one prostate cancer GWAS locus. In addition, our GReX analysis prioritizes multiple novel genetic loci associated with prostate and breast cancer risk, with the same mediated distal regulation of pcGenes. Taken together, our results suggest that distal genetic effects on pcGenes mediated by ncRNAs may be a common mechanism underlying GWAS signals and ncRNAs have a widespread role as distal transcriptional regulators in prostate and breast tissue. We observed more distal ncRNA-pcGene directional associations in tumor than non-cancerous tissue, suggesting that tumor tissues have multiple activated gene regulatory networks with potential effects on disease pathogenesis or progression. Our results implicating distal interactions between ncRNAs and pcGenes are attractive for further in silico and experimental study.

We find many colocalized eQTL signals for ncRNAs or pcGenes in the 17q21.31 region, many of which are for genes with categorized functions in cancer pathways or transcripts prioritized by previous integrative genetic association studies. For example, a previous transcriptome-wide association study (TWAS) of breast cancer stratified by ER status, identified ncRNA *KANSL-AS1*[107]; which has also been associated with ovarian cancer[108]. In our analysis, *KANSL-AS1* mediated multiple pcGenes further downstream on chromosome 17 and colocalized almost perfectly with the breast cancer GWAS association. Some of the pcGenes that according to our analysis are targets of *KANSL-AS1* are also implicated in breast cancer progression, including

WNT3, *ARL17A*, and *TXNRD3*[107,109–113]. These results suggest widespread distal effects of *KANSL-AS1* on chromosome 17. However, after accounting for the large and common H2 inversion in the 17q21.31 locus, the distal effects of KANSL1-AS1 were largely attenuated, pointing to correlated effects of large structural aberrations on gene expression and disease etiology[114–117] in this region. Further, these results also suggest that structural aberrations, like the 17q21.31 H2 inversion, may have large effects on gene expression, not only locally, but also distally. In particular, more comprehensive analyses of the transcriptomic effects of structural inversions and other aberrations in tumor tissue is needed. Our results also serve as a cautionary tale: future gene expression analyses must delineate the eQTL signal in a locus from any chromosomal aberrations when following up on GWAS signals.

We conclude with limitations of our study. First, we rely on Ensembl annotations of gene biotypes to define our sets of ncRNAs and pcGenes. These annotations may be incomplete, and accordingly, we may have ignored multiple transcripts that may be non-coding[118]. Next, as our local- and distal-eQTL signals are derived from entirely shared samples, we were unable to use multiple trait colocalization methods, like *moloc* or *Primo*[119,120]. A more flexible framework that allows for shared molecular QTL signal could be developed to fully interrogate the mediated-distal QTL signal, in the context of complex trait etiology. As mentioned before, we do not systematically account for copy number variation or structural variation to disentangle these potentially disparate signals. Future studies along this same vein should consider corrections for inversions, translocations, or genomic imbalances. Lastly, due to limited sample sizes in TCGA, we could not assess the effects of molecular subtype heterogeneity on eQTL mapping in tumor tissue. Previously, subtype-specific genetic architecture of gene expression regulation has been suggested by previous studies[5,121,122], with some distal genetic associations detected for genes that are highly predictive of molecular subtypes[123]. However, more robust subtype-specific analyses in breast cancer can be informative for both subtype-specific risk and outcomes.

We find evidence that ncRNAs play an important role in mediating multiple distal genetic regulatory mechanisms for pcGenes in breast and prostate tissue. In addition, this genetic signal for ncRNA-mediated distal-eQTLs of protein-coding genes overlaps with genetic signal for prostate and breast cancer risk. From a methodological perspective, our framework provides a systematic approach to use functional genomic features to characterize genetic variants that contribute to transcription through trans-acting pathways.

METHODS

A graphical representation of our methods is provided in **Supplemental Figure S1**.

Data acquisition and processing

We used pre-processed genotype, gene expression, and demographic and clinical covariate data for non-cancerous breast mammary and prostate tissue and breast and prostate tumor tissue from the Genotype-Tissue Expression Project (GTEx) v8[25] and The Cancer Genome Atlas (TCGA)[26], respectively. We included only individuals of European ancestry due to the small sample sizes available for other demographics ($N = 337$ for GTEx breast, $N = 186$ for GTEx prostate, $N = 437$ for TCGA breast, $N = 349$ for TCGA prostate). For both GTEx and TCGA, we only consider SNPs and genes on autosomes, restricted to SNPs with minor allele frequency (MAF) greater than or equal to 1%, and excluded SNPs that deviated from Hardy-Weinberg equilibrium at $P < 10^{-6}$.

We used the BioConductor package *biomaRt* for ENSEMBL gene biotype annotations[124]. Using these annotations, we defined protein-coding genes (pcGenes) as those labeled “protein-coding” and non-coding RNAs (ncRNAs) as those labeled “tRNA”, “rRNA”, “scRNA”, “snRNA”, “snoRNA”, “miRNA”, “miscRNA”, “lincRNA”, “sRNA” and “scaRNA”. These annotations included

16,582 pcGenes and 5,650 ncRNAs in GTEx breast, 16,827 pcGenes and 5,862 ncRNAs in GTEx prostate, 21,648 pcGenes and 1,261 ncRNAs in TCGA breast, and 15,773 pcGenes and 548 ncRNAs in TCGA prostate. In analyses, we considered all provided GTEx covariates: five genotype-based principal components (PCs), up to 60 probabilistically-estimated expression residuals (PEER) factors, age, sex, and sequencing platform and protocol[25]. For TCGA, we calculated genotype PCs using PLINK v1.93[125], calculated up to 50 hidden components of expression (HCP) using *Rhcpp*[126,127], and included the following demographic and clinical covariates: age, estrogen receptor subtype, menopausal status, and disease pathological stage. For prostate tumors, we include the following demographic and clinical covariates: age, sequencing platform, and protocol.

We also integrated eQTL results with genome-wide association summary statistics for overall prostate cancer risk, overall breast cancer risk, and subtype-specific breast cancer risk. We obtained European-ancestry specific GWAS summary statistics for prostate cancer risk from the PRACTICAL Consortium, described in Schumacher, Al Olama, and Berndt et al 2018[1]. We obtain European-ancestry specific overall and subtype-specific GWAS summary statistics for breast cancer risk from the BCAC Consortium, described in Zhang et al 2020[2].

eQTL mapping

We used a multiple linear regression model implemented in the *MatrixEQTL* R package to detect both local- and distal-eQTLs[128], with the following model:

$$E_g = X_s\beta_s + X_C\beta_C + \epsilon,$$

where E_g is a vector of expression values corresponding to a gene g, X_s is a vector of genotypes for SNP s, X_C is a matrix of covariates, and ϵ is random error independently and identically distributed as Gaussian random variables with mean 0 and variance σ^2. To maximize

the number of detected distal-eQTLs associated with pcGenes and local-eQTLs associated with ncRNAs, we iterated on this model using SNPs on Chromosome 22. For breast tumor tissue from TCGA, we found that including all of the clinical covariates (age, estrogen receptor subtype, menopause status, and disease pathological stage), the first 3 PCs, and the first 8 HCPs maximized the number of distal-eQTLs associated with pcGenes and including all of the clinical covariates, the first 4 PCs, and the first 10 HCPs maximized the number of local-eQTLs associated with ncRNAs. As GTEx eQTLs are well-catalogued, for non-cancerous breast tissue, we used the entire set of covariates provided by the consortium (5 genotype PCs, 60 PEER factors, and all previously described technical, clinical, and demographic variables). For prostate tumor tissue, we included age, sequencing platform, protocol as covariates and used 5 genotype PCs and 10 HCPs as the optimized set of covariates for both local- and distal-eQTL mapping[126]. We used the full set of provided covariates from the GTEx consortium for non-cancerous prostate tissue eQTL mapping.

Mediation analysis for distal-eQTL mapping

To identify distal-eQTLs, we fit two regressions to perform mediation analysis[21,23]. We define a local window of 1 Mb and a distal association if two features are more than 1 Mb apart or on separate chromosomes. First, we identify a testing triplet, consisting of (1) a SNP s, (2) a distal pcGene G associated with SNP s with nominal $P < 10^{-6}$, and (3) a set of local ncRNAs $m_1, \ldots, m_M$, all associated univariately with SNP s with nominal $P < 10^{-6}$. Next, we fit the following two linear regressions for mediation analysis:

$$Y_G = X_s\beta_s + M\beta_M + X_C\beta_C + \epsilon_G, \qquad \epsilon_G \sim N(0, \sigma^2 I_n)$$

where Y_G is a vector of expression for gene G, X_s is a vector of dosages for SNP s, β_s is the effect size of SNP s on gene G, M is a matrix of expression of the of the set of m ncRNAs local

to s, β_M is the effects of the M ncRNAs on Y_G, X_C is a matrix of covariates, and ϵ_G is a random error term. The jth ncRNA is modeled as

$$M_j = X_s\alpha_{M_j} + X_C\alpha_{C,j} + \epsilon_{M_j}, 1 \leq j \leq m, \qquad \epsilon_{M_j} \sim N(0, \sigma^2_{M_j} I_n)$$

where M_j is the vector of expression for the jth ncRNA, α_{M_j} is a vector of effects of SNP s on mediator M_j, X_C is a matrix of covariates, $\alpha_{C,j}$is a vector of covariate effects on the mediator, and ϵ_{M_j} represents a random error term.

We define the total mediation effect (TME) as $TME = \alpha_M \cdot \beta_M$ and the mediation proportion, or MP, such that $MP = \frac{\alpha_M \cdot \beta_M}{\beta_s + \alpha_M \cdot \beta_M}$. We test $H_0: TME = 0$ vs. $H_1: TME \neq 0$ via permutation testing with 10,000 draws. We bound the mediation proportion from above at 1.

Gene-based association testing (GBAT)

We also applied GBAT[24] with slight modifications to identfy ncRNAs with genetically-regulated expression (GReX) associated distally (farther than 1 Mb away) with pcGenes. First, we removed multi-mapped reads using previously provided annotations[129]. Then, we estimated the heritability of ncRNA expression using GCTA v1.93[130] excluding all genes with no evidence of heritability at P > 0.05. Next, using leave-one-out cross-validation, we constructed the GReX of ncRNA expression using SNPs local to the ncRNA using elastic net, lasso[131] and SuSiE[29]. We kept ncRNAs with cross-validation adjusted $R^2 > 0.01$. We then employed *MatrixEQTL* to estimate the association between the GReX of ncRNAs and expression of pcGenes that are farther than 1 Mb away using the optimized set of covariates for distal-QTL mapping[128]. Lastly, to identify a credible set of SNPs local to the ncRNA that best explains the gene-level association with the distal pcGene, we used *SuSiE* fine-mapping with default parameters to define a 90% credible set[29].

Colocalization with cancer risk

To identify any potentially overlapping signals between local-eQTLs of ncRNAs, ncRNA-mediated distal-eQTLs of pcGenes, and cancer risk, we employed the Bayesian colocalization method, *coloc*[27]. *coloc* estimates the posterior probability that the same SNP explains both the eQTL and the GWAS signal at a given locus. We used standard parameters with default priors ($p_1 = 10^{-4}$, $p_2 = 10^{-4}$, and $p_{12} = 10^{-5}$) to estimate the colocalization posterior probabilities for five hypotheses. We considered an eQTL signal to colocalize with a GWAS signal if the posterior probability of colocalization through one SNP (PP.H4 in Giambartolomei et al[27]) was greater than 0.75.

Genetically-regulated expression analysis of ncRNAs

We also identified any cancer associations for the genetically-regulated expression (GReX) of any ncRNAs that showed significant mediation of multiple distal-eQTLs of distant pcGenes. First, using elastic net regression (mixing parameters α of 0, 0.25, 0.5, 0.75, and 1), linear mixed modeling[132,133], and SuSiE[29,131], we built predictive models of ncRNAs in both GTEx and TCGA data across breast and prostate tissue and select only models with 10-fold cross-validation McNemar's adjusted R^2. We then employed the weighted burden test and permutation test from the FUSION TWAS framework to detect a trait association with the GReX of an ncRNA[28]. We define transcriptome-wide significance as $P < 2.5 \times 10^{-6}$ and permutation test $P < 0.05$.

AVAILABILITY OF DATA AND MATERIALS

GTEx v8 genotype, expression, and covariate data were obtained through dbGAP Study Accession phs000424.v8.p2. TCGA genotype were obtained through dbGAP Study Accession phs000178.v11.p8 and expression and covariate data was obtained from the Broad GDAC Firehose repository (https://gdac.broadinstitute.org). Prostate cancer risk summary statistics

were obtained from the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) Consortium: http://practical.icr.ac.uk/blog/wp-content/uploads/uploadedfiles/oncoarray/MetaSummaryData/meta_v3_onco_euro_overall_ChrAll_1_release.zip. Breast cancer risk summary statistics were obtained from the Breast Cancer Association Consortium (BCAC): https://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-associations-breast-cancer-risk-2020/. Sample code for this analysis are available at https://github.com/ColetheStatistician/ncRNAInBreastCancer/.

REFERENCES

1. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics* **50**, 928–936 (2018).
2. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics* **52**, 572–581 (2020).
3. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
4. Mancuso, N. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature Communications* **9**, (2018).
5. Li, Q. *et al.* A novel eQTL-based analysis reveals the biology of breast cancer risk loci. *Cell* **152**, 633–633 (2013).
6. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
7. Kopp, F. & Mendell, J. T. Functional classification and experimental dissection of long noncoding RNAs. *Cell* **172**, 393–393 (2018).

8. Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology 2020 22:2* **22**, 96–118 (2020).

9. Vance, K. W. & Ponting, C. P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in Genetics* **30**, 348–348 (2014).

10. Anastasiadou, E., Jacob, L. S. & Slack, F. J. Non-coding RNA networks in cancer. *Nature Reviews Cancer 2017 18:1* **18**, 5–18 (2017).

11. Slack, F. J. & Chinnaiyan, A. M. The Role of Non-coding RNAs in Oncology. *Cell* **179**, 1033–1055 (2019).

12. Chen, D. L. *et al.* Long noncoding RNA XIST expedites metastasis and modulates epithelial–mesenchymal transition in colorectal cancer. *Cell Death & Disease 2017 8:8* **8**, e3011–e3011 (2017).

13. Arun, G., Diermeier, S. D. & Spector, D. L. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends in molecular medicine* **24**, 257–257 (2018).

14. Kamada, S., Takeiwa, T., Ikeda, K., Horie-Inoue, K. & Inoue, S. Long Non-coding RNAs Involved in Metabolic Alterations in Breast and Prostate Cancers. *Frontiers in Oncology* **10**, 2119–2119 (2020).

15. Peterson, S. M. *et al.* Common features of microRNA target prediction tools. *Frontiers in Genetics* **5**, 23–23 (2014).

16. Porcu, E. *et al.* Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications 2021 12:1* **12**, 1–9 (2021).

17. Pierce, B. L. *et al.* Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. *PLoS Genetics* **10**, (2014).

18. Pierce, B. L. *et al.* Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nature Communications* **9**, 1–12 (2018).
19. Yang, F. *et al.* CCmed: cross-condition mediation analysis for identifying replicable trans-associations mediated by cis-gene expression. *Bioinformatics* **37**, 2513–2520 (2021).
20. Yang, F., Wang, J., Pierce, B. L. & Chen, L. S. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Research* **27**, 1859–1871 (2017).
21. Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLoS Genetics* **17**, e1009398–e1009398 (2021).
22. Bhattacharya, A. *et al.* Placental genomics mediates genetic associations with complex health traits and disease. *Nat Commun* **13**, 706 (2022).
23. Shan, N., Wang, Z. & Hou, L. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics* **20**, (2019).
24. Liu, X. *et al.* GBAT: a gene-based association test for robust detection of trans-gene regulation. *Genome Biology* **21**, 211–211 (2020).
25. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
26. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45**, 1113–1120 (2013).
27. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics* **10**, e1004383–e1004383 (2014).
28. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

29. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B* **82**, 1273–1300 (2020).
30. Mackinnon, D. P., Lockwood, C. M. & Williams, J. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research* **39**, 99–128 (2004).
31. Benjamini, Y. & Hochberg, Y. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. *Source: Journal of the Royal Statistical Society. Series B (Methodological)* vol. 57 289–300 (1995).
32. Carrasco, A. M. *et al.* PUM1 and RNase P genes as potential cell-free DNA markers in breast cancer. *Journal of Clinical Laboratory Analysis* **35**, e23720–e23720 (2021).
33. Shi, P. *et al.* Long non-coding RNA NORAD inhibition upregulates microRNA-323a-3p to suppress tumorigenesis and development of breast cancer through the PUM1/eIF2 axis. *https://doi.org/10.1080/15384101.2021.1934627* **20**, 1295–1307 (2021).
34. Shi, H. *et al.* PCBP1 depletion promotes tumorigenesis through attenuation of p27Kip1 mRNA stability and translation. *Journal of Experimental & Clinical Cancer Research 2018 37:1* **37**, 1–18 (2018).
35. Goulding, H. *et al.* Metallothionein expression in human breast cancer. *British Journal of Cancer 1995 72:4* **72**, 968–972 (1995).
36. Si, M. & Lang, J. The roles of metallothioneins in carcinogenesis. *Journal of Hematology & Oncology 2018 11:1* **11**, 1–20 (2018).
37. Michmerhuizen, A. R., Spratt, D. E., Pierce, L. J. & Speers, C. W. ARe we there yet? Understanding androgen receptor signaling in breast cancer. *npj Breast Cancer 2020 6:1* **6**, 1–19 (2020).
38. Masi, M. *et al.* OXER1 and RACK1-associated pathway: a promising drug target for breast cancer progression. *Oncogenesis 2020 9:12* **9**, 1–15 (2020).

39. Zhang, L. *et al.* A Multifunctional Therapy Approach for Cancer: Targeting Raf1- Mediated Inhibition of Cell Motility, Growth, and Interaction with the Microenvironment. *Molecular Cancer Therapeutics* **19**, 39–51 (2020).
40. Wang, B. *et al.* Castration-Resistant Lgr5+ Cells Are Long-Lived Stem Cells Required for Prostatic Regeneration. *Stem Cell Reports* **4**, 768–768 (2015).
41. Shivakumar, M. *et al.* Genetic analysis of functional rare germline variants across 9 cancer types from the DiscovEHR study. *medRxiv* 2019.12.09.19013334-2019.12.09.19013334 (2019) doi:10.1101/2019.12.09.19013334.
42. Mukhtar, E., Adhami, V. M. & Mukhtar, H. Targeting Microtubules by Natural Agents for Cancer Therapy. *Molecular cancer therapeutics* **13**, 275–275 (2014).
43. Abaffy, T. *et al.* A Testosterone Metabolite 19-Hydroxyandrostenedione Induces Neuroendocrine Trans-Differentiation of Prostate Cancer Cells via an Ectopic Olfactory Receptor. *Frontiers in Oncology* **0**, 162–162 (2018).
44. Neuhaus, E. M. *et al.* Activation of an Olfactory Receptor Inhibits Proliferation of Prostate Cancer Cells. *The Journal of Biological Chemistry* **284**, 16218–16218 (2009).
45. Maine, E. A. *et al.* The cancer-testis antigens SPANX-A/C/D and CTAG2 promote breast cancer invasion. *Oncotarget* **7**, 14708–14708 (2016).
46. Quiroz-Zárate, A. *et al.* Expression Quantitative Trait loci (QTL) in tumor adjacent normal breast tissue and breast tumor tissue. *PLOS ONE* **12**, e0170181–e0170181 (2017).
47. Sun, C.-C. *et al.* FOXC1-mediated LINC00301 facilitates tumor progression and triggers an immune-suppressing microenvironment in non-small cell lung cancer by regulating the HIF1α pathway. *Genome Medicine 2020 12:1* **12**, 1–27 (2020).
48. Tuna, M., Machado, A. S. & Calin, G. A. Genetic and epigenetic alterations of microRNAs and implications for human cancers and other diseases. *Genes Chromosomes and Cancer* **55**, 193–214 (2016).

49. Xiao, L. *et al.* RNA Sequence Profiling Reveals Unique Immune and Metabolic Features of Breast Cancer Brain Metastases. *Frontiers in Oncology* **0**, 3213–3213 (2021).
50. MC, B., U, Y., MO, B., O, S. & A, B. Tamoxifen inhibits cytochrome P450 2C9 activity in breast cancer patients. *Journal of chemotherapy (Florence, Italy)* **18**, 421–424 (2006).
51. Cai, P. *et al.* BTN3A2 serves as a prognostic marker and favors immune infiltration in triple-negative breast cancer. *Journal of Cellular Biochemistry* **121**, 2643–2654 (2020).
52. Tao, S. *et al.* A genome-wide search for loci interacting with known prostate cancer risk-associated genetic variants. *Carcinogenesis* **33**, 598–598 (2012).
53. MS, A. F. *et al.* The increased expression of fatty acid-binding protein 9 in prostate cancer and its prognostic significance. *Oncotarget* **7**, 82783–82797 (2016).
54. MJ, M. *et al.* Association of type 2 diabetes susceptibility variants with advanced prostate cancer risk in the Breast and Prostate Cancer Cohort Consortium. *American journal of epidemiology* **176**, 1121–1129 (2012).
55. Marín-Aguilera, M. *et al.* Molecular profiling of peripheral blood is associated with circulating tumor cells content and poor survival in metastatic castration-resistant prostate cancer. *Oncotarget* **6**, 10604–10604 (2015).
56. Saigo, C., Kito, Y. & Takeuchi, T. Cancerous Protein Network That Inhibits the Tumor Suppressor Function of WW Domain-Containing Oxidoreductase (WWOX) by Aberrantly Expressed Molecules. *Frontiers in Oncology* **0**, 350–350 (2018).
57. S, F., J, W., Y, Z., CJ, C. & M, I. FGF23 promotes prostate cancer progression. *Oncotarget* **6**, 17291–17301 (2015).
58. Li, Y.-L., Tian, H., Jiang, J., Zhang, Y. & Qi, X.-W. Multifaceted regulation and functions of fatty acid desaturase 2 in human cancers. *American Journal of Cancer Research* **10**, 4098–4098 (2020).
59. Chi, J.-R. *et al.* SNHG5 Promotes Breast Cancer Proliferation by Sponging the miR-154-5p/PCNA Axis. *Molecular Therapy - Nucleic Acids* **17**, 138–149 (2019).

60. Damas, N. D. *et al.* SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. *Nature Communications 2016 7:1* **7**, 1–14 (2016).

61. Wang, M. *et al.* Large-scale association analysis in Asians identifies new susceptibility loci for prostate cancer. *Nature Communications 2015 6:1* **6**, 1–7 (2015).

62. Sinnott-Armstrong, N. *et al.* Convergent mutations in tissue-specific regulatory regions reveal novel cancer drivers. *bioRxiv* 2020.08.21.239954-2020.08.21.239954 (2020) doi:10.1101/2020.08.21.239954.

63. Kang, D., Lee, Y. & Lee, J.-S. RNA-Binding Proteins in Cancer: Functional and Therapeutic Perspectives. *Cancers* **12**, 1–33 (2020).

64. Hu, C. *et al.* Lipocalin 2: a potential therapeutic target for breast cancer metastasis. *OncoTargets and therapy* **11**, 8099–8099 (2018).

65. Sack, L. M. *et al.* Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns. *Cell* **173**, 499–499 (2018).

66. Smith, J. C. & Sheltzer, J. M. Genome-wide identification and analysis of prognostic features in human cancers. *bioRxiv* 2021.06.01.446243-2021.06.01.446243 (2021) doi:10.1101/2021.06.01.446243.

67. Sun, L., Zhang, Y. & Zhang, C. Distinct Expression and Prognostic Value of MS4A in Gastric Cancer. *Open Medicine* **13**, 178–178 (2018).

68. Hou, R. & Jiang, L. LINC00115 promotes stemness and inhibits apoptosis of ovarian cancer stem cells by upregulating SOX9 and inhibiting the Wnt/β-catenin pathway through competitively binding to microRNA-30a. *Cancer Cell International 2021 21:1* **21**, 1–15 (2021).

69. Li, Q. *et al.* A Positive Feedback Loop of Long Noncoding RNA LINC00152 and KLF5 Facilitates Breast Cancer Growth. *Frontiers in Oncology* **0**, 899–899 (2021).

70. Yuan, C., Luo, X., Duan, S. & Guo, L. Long noncoding RNA LINC00115 promotes breast cancer metastasis by inhibiting miR-7. *FEBS Open Bio* **10**, 1230–1237 (2020).

71. Ma, Y., Xu, X. L., Huang, H. G., Li, Y. F. & Li, Z. G. LncRNA TDRG1 promotes the aggressiveness of gastric carcinoma through regulating miR-873-5p/HDGF axis. *Biomedicine & Pharmacotherapy* **121**, 109425–109425 (2020).

72. Hu, X., Mu, Y., Wang, J. & Zhao, Y. LncRNA TDRG1 promotes the metastasis of NSCLC cell through regulating miR-873-5p/ZEB1 axis. *Journal of Cellular Biochemistry* **122**, 969–982 (2021).

73. Tracey, L. J. & Justice, M. J. Off to a bad start: Cancer initiation by pluripotency regulator PRDM14. *Trends in genetics : TIG* **35**, 489–489 (2019).

74. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).

75. McGeary, S. E. *et al.* The biochemical basis of microRNA targeting efficacy. *Science* **366**, eaav1741 (2019).

76. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research* **47**, 199–205 (2019).

77. Thomas, S. J., Snowden, J. A., Zeidler, M. P. & Danson, S. J. The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *British Journal of Cancer 2015 113:3* **113**, 365–371 (2015).

78. Hoxhaj, G. & Manning, B. D. The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism. *Nature Reviews Cancer 2019 20:2* **20**, 74–88 (2019).

79. Masjedi, S., Zwiebel, L. J. & Giorgio, T. D. Olfactory receptor gene abundance in invasive breast carcinoma. *Scientific Reports 2019 9:1* **9**, 1–12 (2019).

80. Kalra, S. *et al.* Analysis of single-cell transcriptomes links enrichment of olfactory receptors with cancer cell differentiation status and prognosis. *Communications Biology 2020 3:1* **3**, 1–10 (2020).

81. Feng, Y. *et al.* Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & Diseases* **5**, 77–77 (2018).

82. Holle, A. W. *et al.* Intermediate filament reorganization dynamically influences cancer cell alignment and migration. *Scientific Reports 2017 7:1* **7**, 1–14 (2017).

83. Ong, M. S. *et al.* Cytoskeletal Proteins in Cancer and Intracellular Stress: A Therapeutic Perspective. *Cancers* **12**, (2020).

84. Tang, Y. *et al.* LncRNAs regulate the cytoskeleton and related Rho/ROCK signaling in cancer metastasis. *Molecular Cancer 2018 17:1* **17**, 1–10 (2018).

85. Seroussi, E. *et al.* Duplications on Human Chromosome 22 Reveal a Novel Ret Finger Protein-Like Gene Family with Sense and Endogenous Antisense Transcripts. *Genome Research* **9**, 803–814 (1999).

86. Li, X. & Meng, Y. Survival analysis of immune-related lncRNA in low-grade glioma. *BMC Cancer* **19**, (2019).

87. Mao, Y., Dong, L., Zheng, Y., Dong, J. & Li, X. Prediction of recurrence in cervical cancer using a nine-lncRNA signature. *Frontiers in Genetics* **10**, 284–284 (2019).

88. Strous, G. J. *et al.* Growth Hormone Receptor Regulation in Cancer and Chronic Diseases. *Frontiers in Endocrinology* **11**, 867–867 (2020).

89. Hu, C. C. *et al.* LncRNA RUSC1-AS1 promotes the proliferation of breast cancer cells by epigenetic silence of KLF2 and CDKN1A. *European review for medical and pharmacological sciences* **23**, 6602–6611 (2019).

90. Zhu, L. *et al.* A panel of 8-lncRNA predicts prognosis of breast cancer patients and migration of breast cancer cells. *PLOS ONE* **16**, e0249174–e0249174 (2021).

91. Yousaf, N., Deng, Y., Kang, Y. & Riedel, H. Four PSM/SH2-B Alternative Splice Variants and Their Differential Roles in Mitogenesis *. *Journal of Biological Chemistry* **276**, 40940–40948 (2001).

92. Liu, L., Yang, J. & Wang, C. Analysis of the prognostic significance of solute carrier (SLC) family 39 genes in breast cancer. *Bioscience reports* **40**, (2020).

93. Cheng, X. *et al.* Zinc transporter SLC39A13/ZIP13 facilitates the metastasis of human ovarian cancer cells via activating Src/FAK signaling pathway. *Journal of Experimental and Clinical Cancer Research* **40**, 1–15 (2021).

94. Permuth-Wey, J. *et al.* Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. *Nature communications* **4**, 1627–1627 (2013).

95. Ruiz-Arenas, C., Cáceres, A., Moreno, V. & González, J. R. Common polymorphic inversions at 17q21.31 and 8p23.1 associate with cancer prognosis. *Human Genomics* **13**, 1–9 (2019).

96. Orsetti, B., Courjal, F., Cuny, M., Rodriguez, C. & Theillet, C. 17q21-q25 aberrations in breast cancer: combined allelotyping and CGH analysis reveals 5 regions of allelic imbalance among which two correspond to DNA amplification. *Oncogene 1999 18:46* **18**, 6262–6270 (1999).

97. O'Brien, H. E. *et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biology* **19**, (2018).

98. Trampush, J. W. *et al.* GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. *Molecular Psychiatry 2017 22:3* **22**, 336–345 (2017).

99. Wei, X. T. *et al.* Pleiotropic genomic variants at 17q21.31 associated with bone mineral density and body fat mass: a bivariate genome-wide association analysis. *European Journal of Human Genetics 2020 29:4* **29**, 553–563 (2020).

100. Bowles, K. R. *et al.* 17q21.31 sub-haplotypes underlying H1-associated risk for Parkinson's disease are associated with LRRC37A/2 expression in astrocytes. *bioRxiv* 860668–860668 (2021) doi:10.1101/860668.

101. Boettger, L. M., Handsaker, R. E., Zody, M. C. & Mccarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature genetics* **44**, 881–881 (2012).

102. Guo, H. *et al.* Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. *Nat Genet* **48**, 1142–1150 (2016).

103. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics* **45**, 353–361 (2013).

104. Sirey, T. M. *et al.* The long non-coding rna cerox1 is a post transcriptional regulator of mitochondrial complex i catalytic activity. *eLife* **8**, (2019).

105. Deng, X., Zhao, Y. & Wang, B. miR-519d-mediated downregulation of STAT3 suppresses breast cancer progression. *Oncology reports* **34**, 2188–2194 (2015).

106. Li, D. *et al.* MiR-519d-3p suppresses breast cancer cell growth and motility via targeting LIM domain kinase 1. *Molecular and cellular biochemistry* **444**, 169–178 (2018).

107. Feng, H. *et al.* Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genetic Epidemiology* **44**, 442–468 (2020).

108. Gusev, A. *et al.* A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nature Genetics* **51**, 815–823 (2019).

109. Pashirzad, M. *et al.* Role of Wnt3a in the pathogenesis of cancer, current status and prospective. *Molecular Biology Reports* **46**, 5609–5616 (2019).

110. Wu, Y. *et al.* Expression of Wnt3 activates Wnt/β-catenin pathway and promotes EMT-like phenotype in trastuzumab-resistant HER2-overexpressing breast cancer cells. *Molecular cancer research : MCR* **10**, 1597–1606 (2012).

111. Zhou, J. X. *et al.* Identification of KANSARL as the first cancer predisposition fusion gene specific to the population of European ancestry origin. *Oncotarget* **8**, 50594–50594 (2017).

112. Wu, W. *et al.* A pan-cancer study of selenoprotein genes as promising targets for cancer therapy. *BMC Medical Genomics* **14**, 1–14 (2021).

113. Yu, Y., Hu, H., Doust, A. N. & Kellogg, E. A. Divergent gene expression networks underlie morphological diversity of abscission zones in grasses. *New Phytologist* nph.16087-nph.16087 (2019) doi:10.1111/nph.16087.

114. de Jong, S. *et al.* Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics* **13**, 1–6 (2012).

115. Lu, Y. *et al.* A transcriptome-wide association study among 97,898 women to identify candidate susceptibility genes for epithelial ovarian cancer risk. *Cancer Research* **78**, 5419–5430 (2018).

116. Su, X. *et al.* Transcriptome-Wide Association Study Provides Insights Into the Genetic Component of Gene Expression in Anxiety. *Frontiers in Genetics* **12**, 1901–1901 (2021).

117. Lee, C. T. *et al.* Functional Consequences of 17q21.31/WNT3-WNT9B Amplification in hPSCs with Respect to Neural Differentiation. *Cell Reports* **10**, 616–632 (2015).

118. Zhang, D. *et al.* Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances* **6**, 8299–8299 (2020).

119. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

120. Gleason, K. J., Yang, F., Pierce, B. L., He, X. & Chen, L. S. Primo: Integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome Biology* **21**, 236–236 (2020).

121. Bhattacharya, A. *et al.* A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biology* **21**, 42–42 (2020).

122. Geeleher, P. *et al.* Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biology* **19**, 130–130 (2018).

123. Patel, A. *et al.* Gene level germline contributions to clinical risk of recurrence scores in Black and White breast cancer patients. *Cancer Research* canres.1207.2021-canres.1207.2021 (2021) doi:10.1158/0008-5472.CAN-21-1207.

124. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols 2009 4:8* **4**, 1184–1191 (2009).

125. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet* **81**, 559–575 (2007).

126. Mostafavi, S. *et al.* Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge. *PLoS ONE* **8**, e68141–e68141 (2013).

127. van Iterson, M. *mvaniterson/Rhcpp: A fast R implementation using Rcpp based on the original matlab HCP method for Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge*. (2020).

128. Shabalin, A. A. Gene expression Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)* **28**, 1353–1358 (2012).

129. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors [version 2; peer review: 3 approved]. *F1000Research* **7**, 1860–1860 (2019).

130. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76–82 (2011).

131. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22 (2010).

132. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* **4**, 250–255 (2011).

133. Mefford, J. *et al.* Efficient Estimation and Applications of Cross-Validated Genetic Predictions to Polygenic Risk Scores and Linear Mixed Models. *Journal of Computational Biology* **27**, 599–612 (2020).

ACKNOWLEDGEMENTS

We thank Nicholas Mancuso, Harold Pimentel, Mike Love, Achal Patel, and Kangcheng Hou for engaging conversation during the research process. We also thank the UCLA Bruins-in-Genomics Summer Program for the research opportunities.

The Prostate cancer genome-wide association analyses are supported by the Canadian Institutes of Health Research, European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), Cancer Research UK Grants C5047/A7357, C1287/A10118, C1287/A16563, C5047/A3354, C5047/A10692, C16913/A6135, and The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148537-01 (the GAME-ON initiative).

Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I] and by Cancer Research UK grant A8197/A16565. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher).

We would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now PCUK), The Orchid Cancer Appeal, Rosetrees Trust, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust.

The breast cancer genome-wide association analyses for BCAC and CIMBA were supported by Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710, C12292/A20861, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), 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 Quebec 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).

FUNDING

TS was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number T32NS048004. This research was supported by the National Institute of Mental Health of the National Institutes of Health under Award number 5R01MH115676-04. AG was partially supported by R01 CA227237. SL was partially supported by NIH award R01 CA194393. BP were partially supported by NIH awards R01 HG009120, R01 MH115676, R01 CA251555, R01 AI153827, R01 HG006399, R01 CA244670, U01 HG011715. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR INFORMATION

Contributions

AB conceived the study. AB, TS, and PL developed the statistical approaches, performed the analysis, and drafted the paper. AB, TS, and BP provided insight in methodological approaches and analysis. BP provided data resources. AB and BP supervised the study. All authors approved and edited the final manuscript.

Corresponding author

Correspondence to Arjun Bhattacharya (abtbhatt@ucla.edu)

ETHICS DECLARATION

Ethics approval and consent to participate

This study was approved by the Office of Human Research Ethics at the University of California, Los Angeles, and written informed consent was obtained from each participant. All experimental methods abided by the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

FIGURES

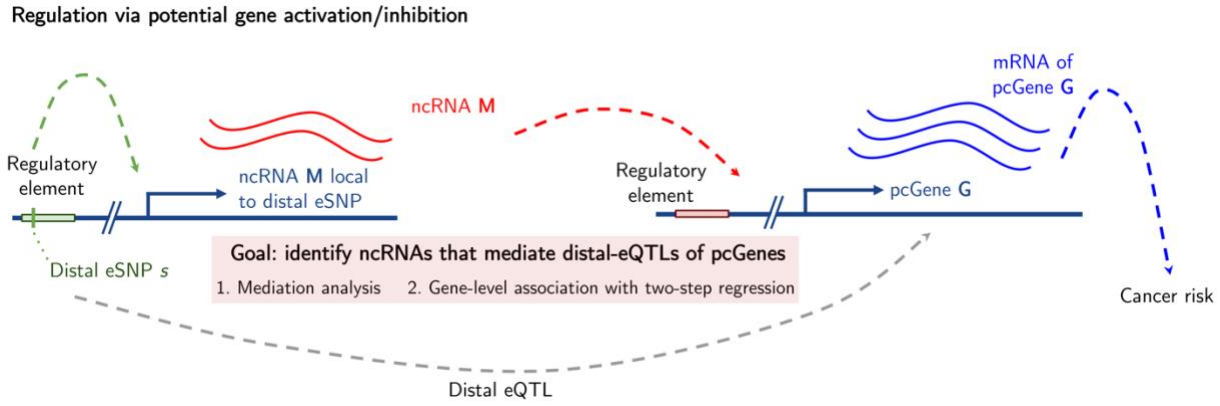


Figure 1: *Schematic of mediation analysis to identify ncRNA-mediated distal-eQTLs of protein-coding genes.* A SNP s is distal to protein-coding gene G and local to ncRNA M, where M has some distal regulatory effect on G. We use expression quantitative trait locus mapping to identify the local-eQTL between s and M (green dotted line) and the distal-eQTL between s and G (grey dotted line). Using either mediation analysis or gene-level association testing, we estimate the indirect mediation effect of s on G through effects from M (red line). Lastly, we use colocalization and genetically-regulated expression analysis to find any intersecting genetic signal between distal-eQTLs of G and genetic associations with breast and prostate cancer risk.

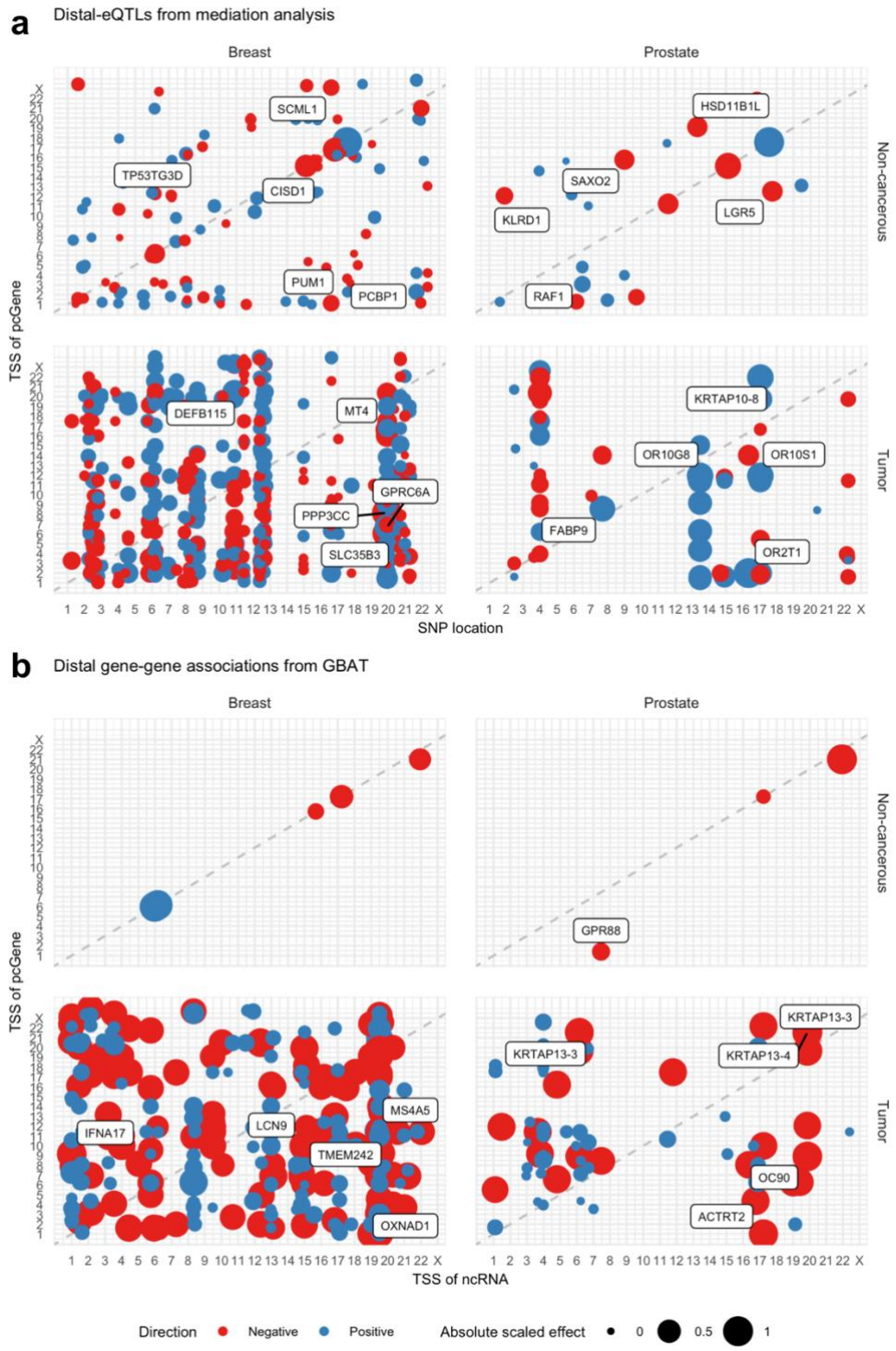


Figure 2: *Location plot of distal-eQTL or gene-gene associations across healthy and tumor samples of breast and prostate tissue.* **(a)** eSNP (X-axis) position vs. transcription start site (TSS) of pcGene (Y-axis) at FDR-adjusted P < 0.01, sized by absolute scaled TME and colored by direction of effect. **(b)** ncRNA TSS (X-axis) vs. pcGene TSS (Y-axis) at FDR-adjusted P < 0.01, sized by absolute scaled gene-gene effect and colored by direction of effect. Top cross-chromosomal distal-eGenes with largest effects are labeled.

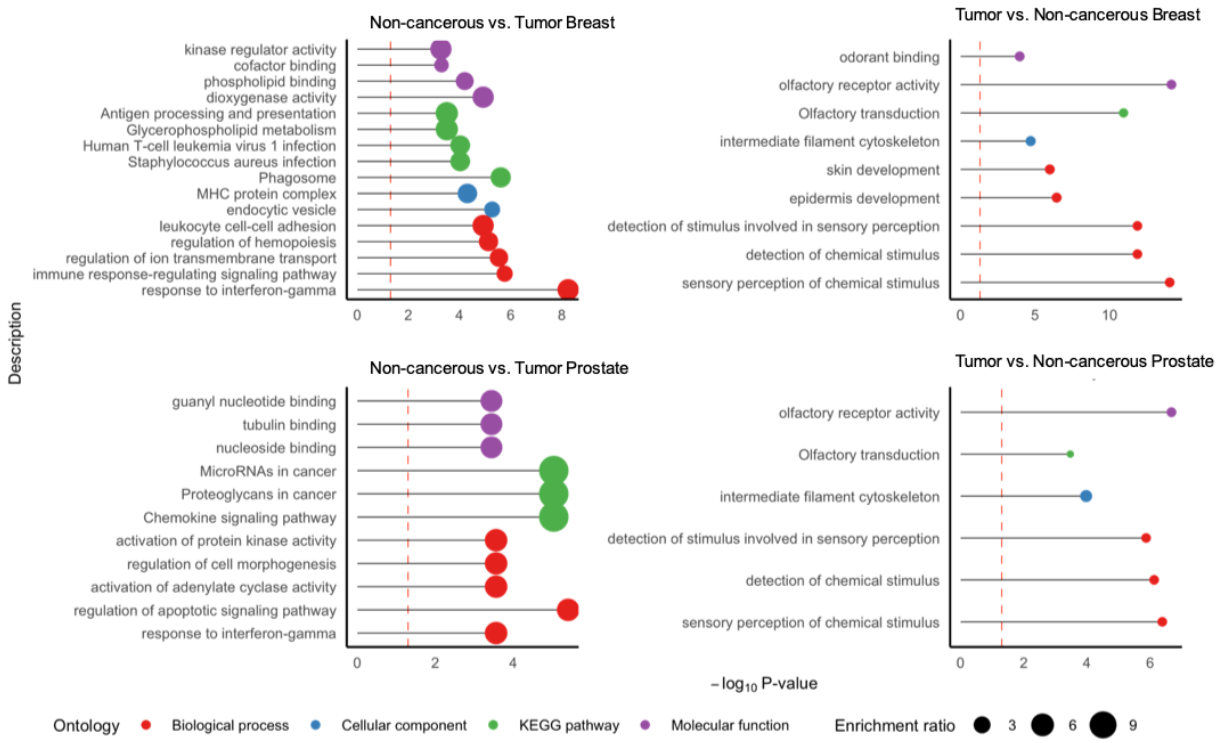


Figure 3: *Over-represented ontologies for tissue-specific eGenes*. -$\log_{10}$ P-value of enrichment (X-axis) of over-represented gene sets (Y-axis), with point sized by enrichment ratio and colored by ontology category. Here, for a tissue, we compare the set of pcGenes from healthy or tumor state to the universe of all pcGenes for tissue across both healthy and tumor states.

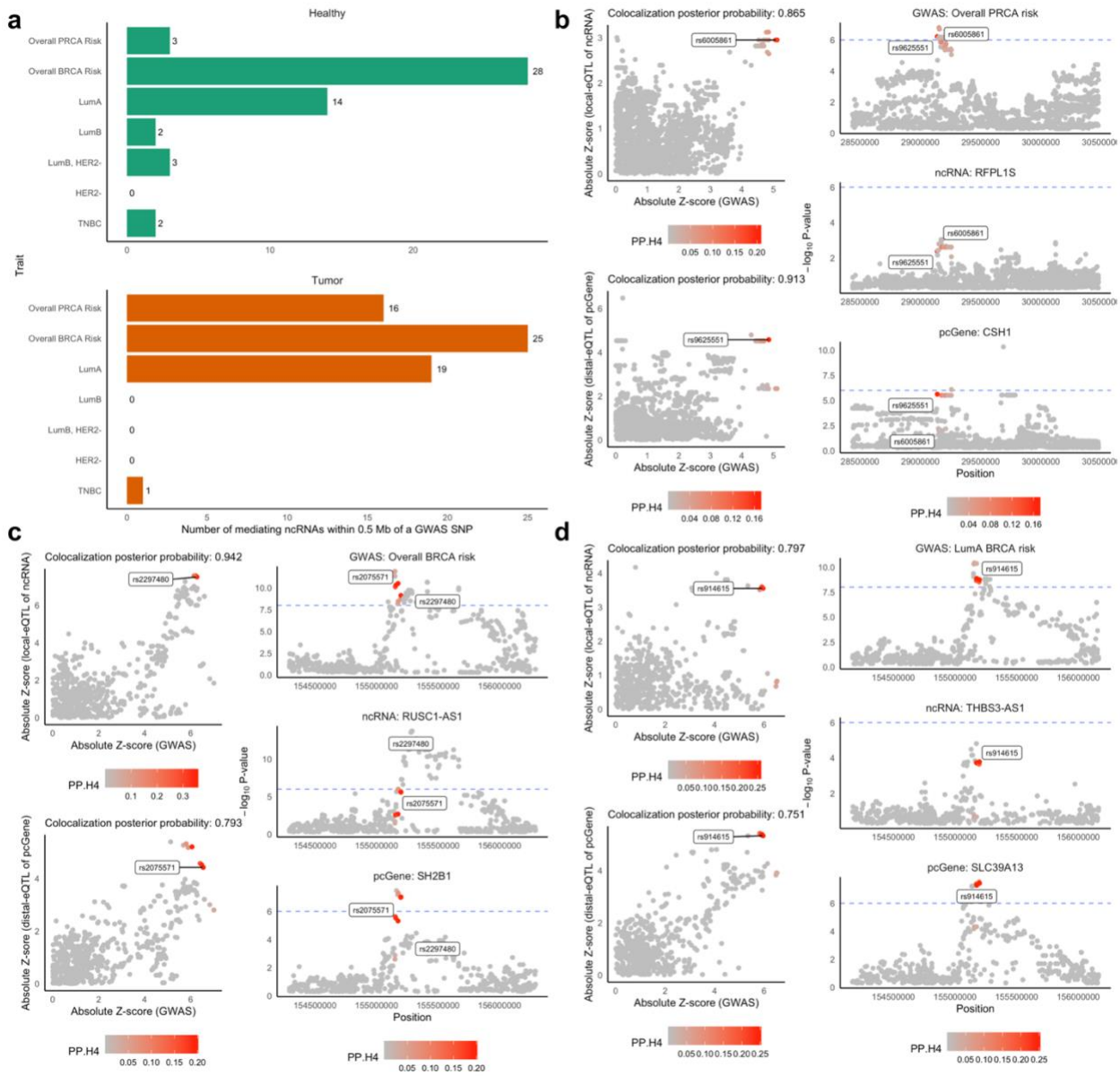


Figure 4: *Colocalization of local-eQTLs of ncRNA and distal-eQTLs of pcGene with GWAS*. **(a)** Barplot of numbers of mediating ncRNAs within 1 Megabase of a GWAS SNP (X-axis) from cancers (Y-axis). **(b-d)** Colocalization results for example ncRNAs and pcGenes, with phenotype in the GWAS, ncRNA, and pcGene provided. Left panel shows scatterplot of absolute Z-scores of GWAS (X-axis) and eQTL associations (Y-axis) with points colored by posterior probability of colocalization (PP.H4). Right panel shows a Manhattan plot of GWAS (top), ncRNA local-eQTL (middle), and pcGene distal-eQTL (bottom) signal, colored by PP.H4.

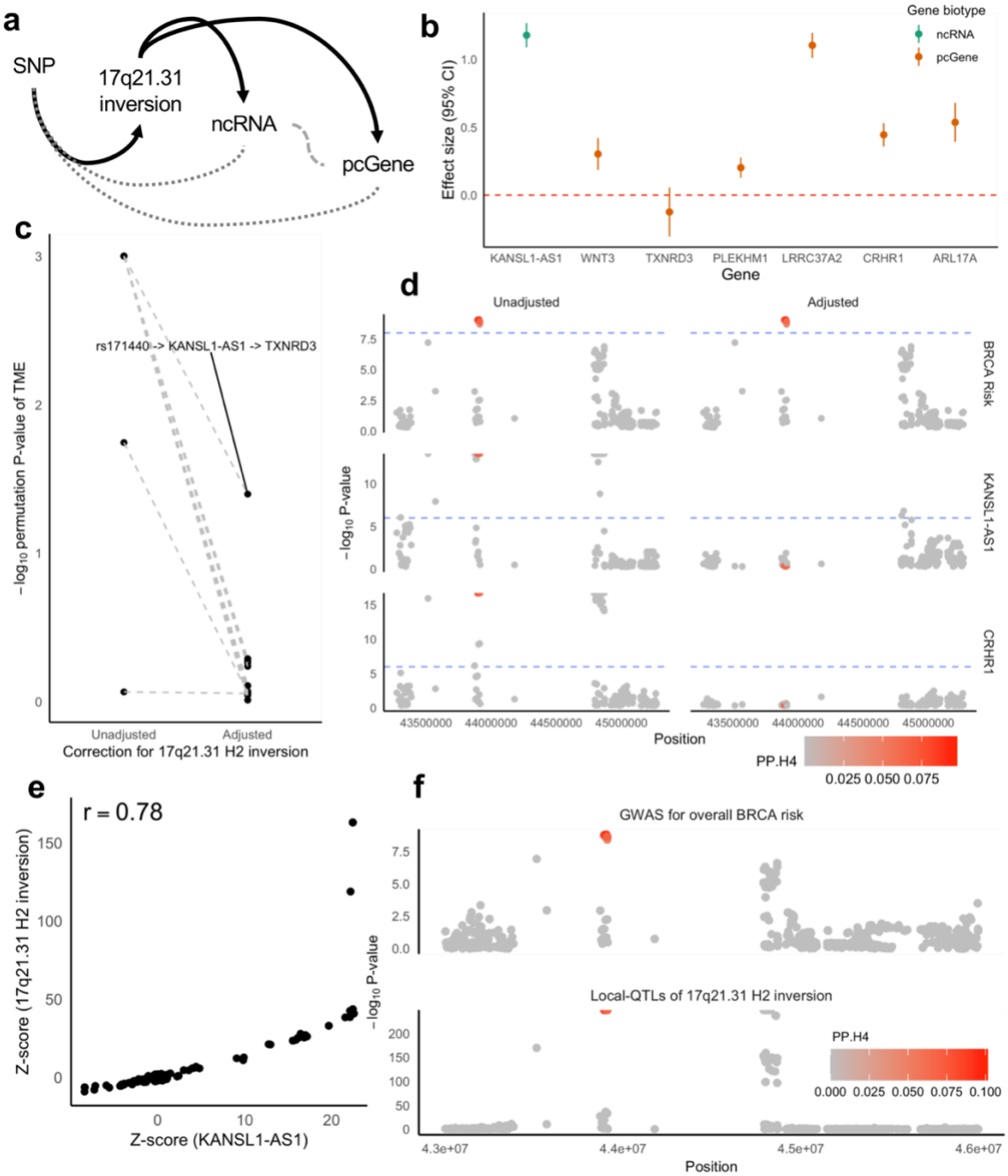


Figure 5: *Impact of H2 inversion on eQTLs in the 17q21.31 locus.* **(a)** Causal diagram of genetic effects in 17q21.31, where strong effects of genetically-determined H2 inversion on ncRNA and pcGene (in black) induces the observed SNP-ncRNA and -pcGene associations (in grey). **(b)** Forest plot of effect sizes and 95% confidence interval (Y-axis) on H2 inversion on pcGenes or ncRNAs (X-axis). **(c)** Difference in -log10 permutation P-value (Y-axis) of total mediation effect of ncRNA on pcGene, with or without adjustment for H2 inversion (X-axis). **(d)** Manhattan plots of GWAS, local-eQTLs of KANSL-AS1, and distal-eQTLs of CRHR1, unadjusted (left) and adjusted (right) for H2 inversion, colored by per-SNP PP.H4. **(e)** Scatterplot of Z-score of local-eQTLs on KANSL1-AS1 (X-axis) against Z-score of local-QTLs of H2 inversion. **(f)** Manhattan plots of GWAS and local-QTLs of H2 inversion, colored by per-SNP PP.H4.

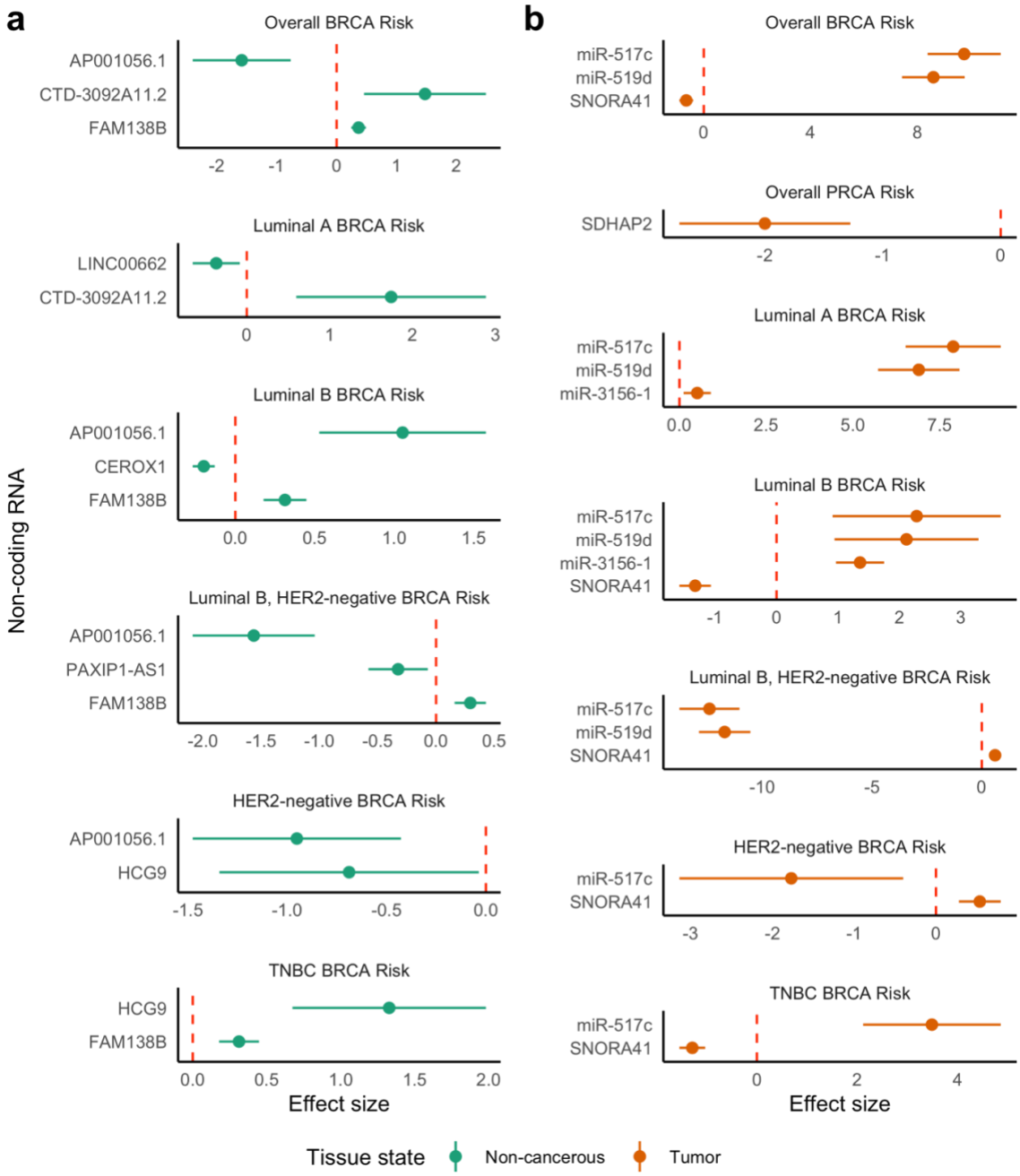


Figure 6: *GReX associations with cancer risk for ncRNAs mediating multiple distal-eQTLs of pcGenes*. Forest plot of effect size and confidence intervals at significance level of P = 2.5 x 10^{-6} (Y-axis) of GReX-associations with overall and subtype-specific cancer risk across ncRNAs that showed significant mediation of multiple distal-eQTLs of distinct pcGenes (X-axis) in non-cancerous **(a)** and tumor **(b)** tissue states.

TABLES

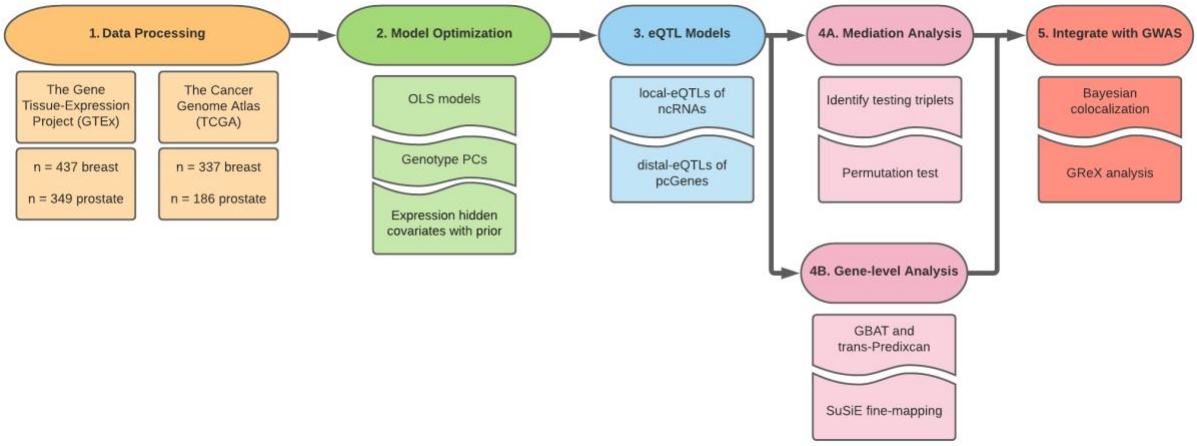
Table 1: *Summary of local- and distal-eQTL mapping results across breast and prostate non-cancerous and tumor tissue through mediation analysis.*

	Breast		*Prostate*	
	Non-cancerous	Tumor	Non-cancerous	Tumor
Total eQTLs				
Local	22,832	1,298	12,511	6,368
Distal	29,512	316,097	31,782	363,587
Total eGenes				
local-ncRNAs	1,113	87	773	60
distal-pcGenes	8,849	19,376	8,711	15,560
SNPs associated in both local and distal eQTLs				
Local	1,569	601	281	6,368
Distal	1,580	5,089	264	9,259
Mediated distal-eQTLs				
eSNPs	703	3,017	103	425
e-ncRNAs	157	45	22	18
e-pcGenes	173	562	24	107

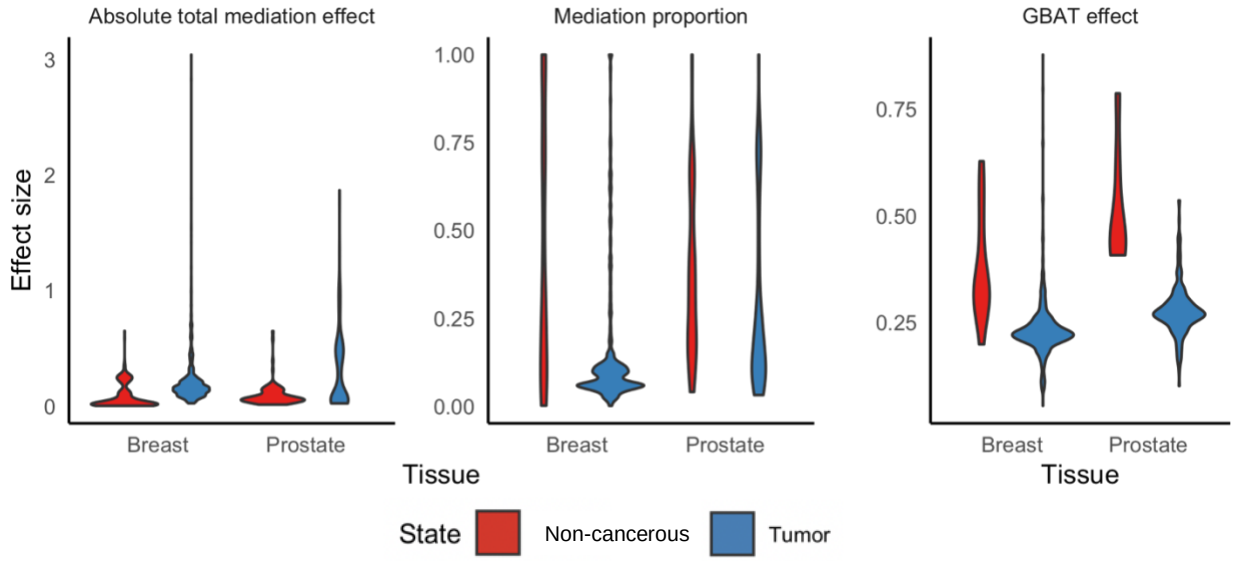
Table 2: *Summary of distal-eQTL mapping results across breast and prostate normal and tumor tissue using GBAT.*

	Breast		*Prostate*	
	Normal	Tumor	Normal	Tumor
Total gene-gene associations	13	1,375	7	297
Unique ncRNAs	9	209	4	84
Unique pcGenes	10	1,127	5	268

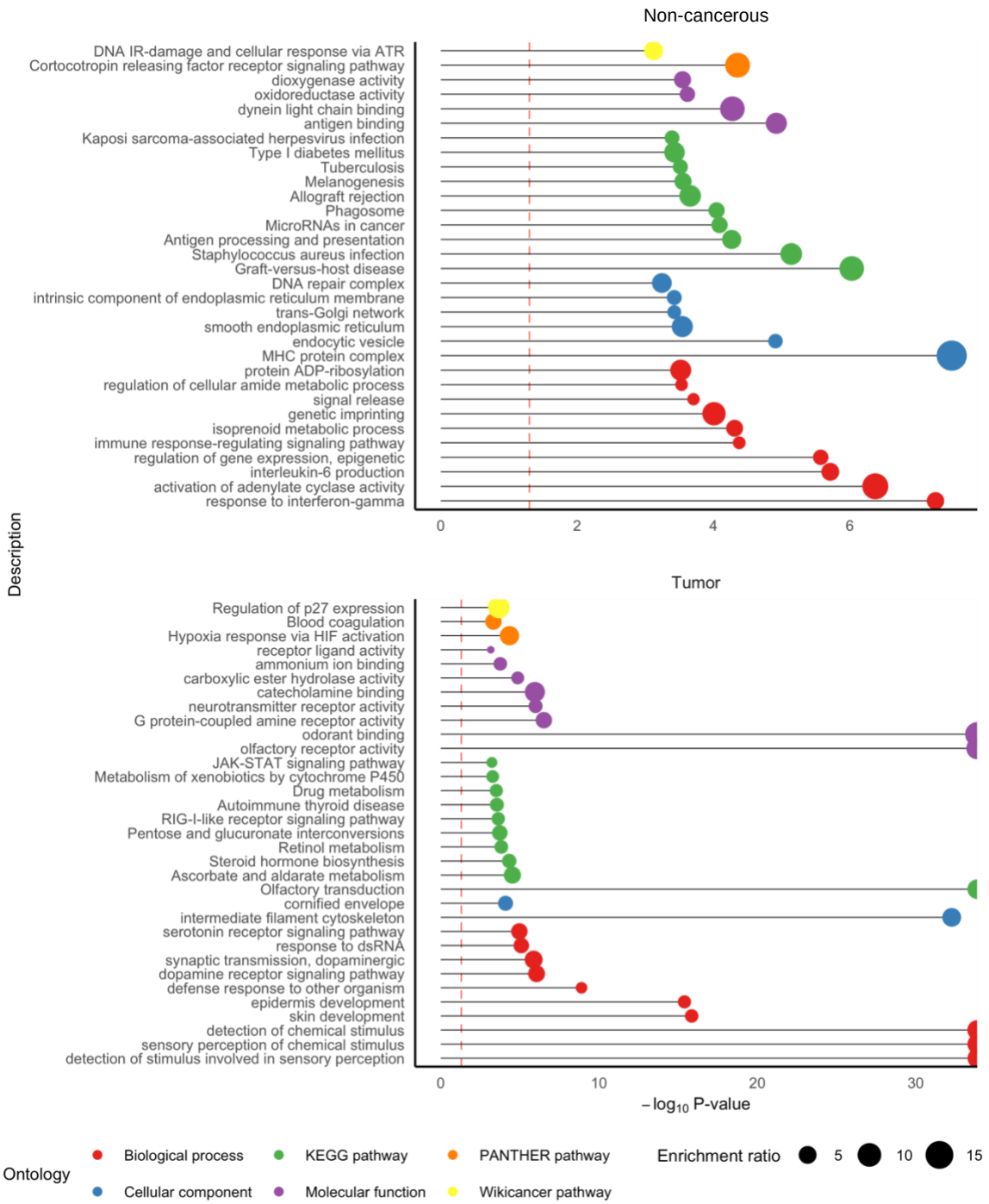
SUPPLEMENTAL FIGURES



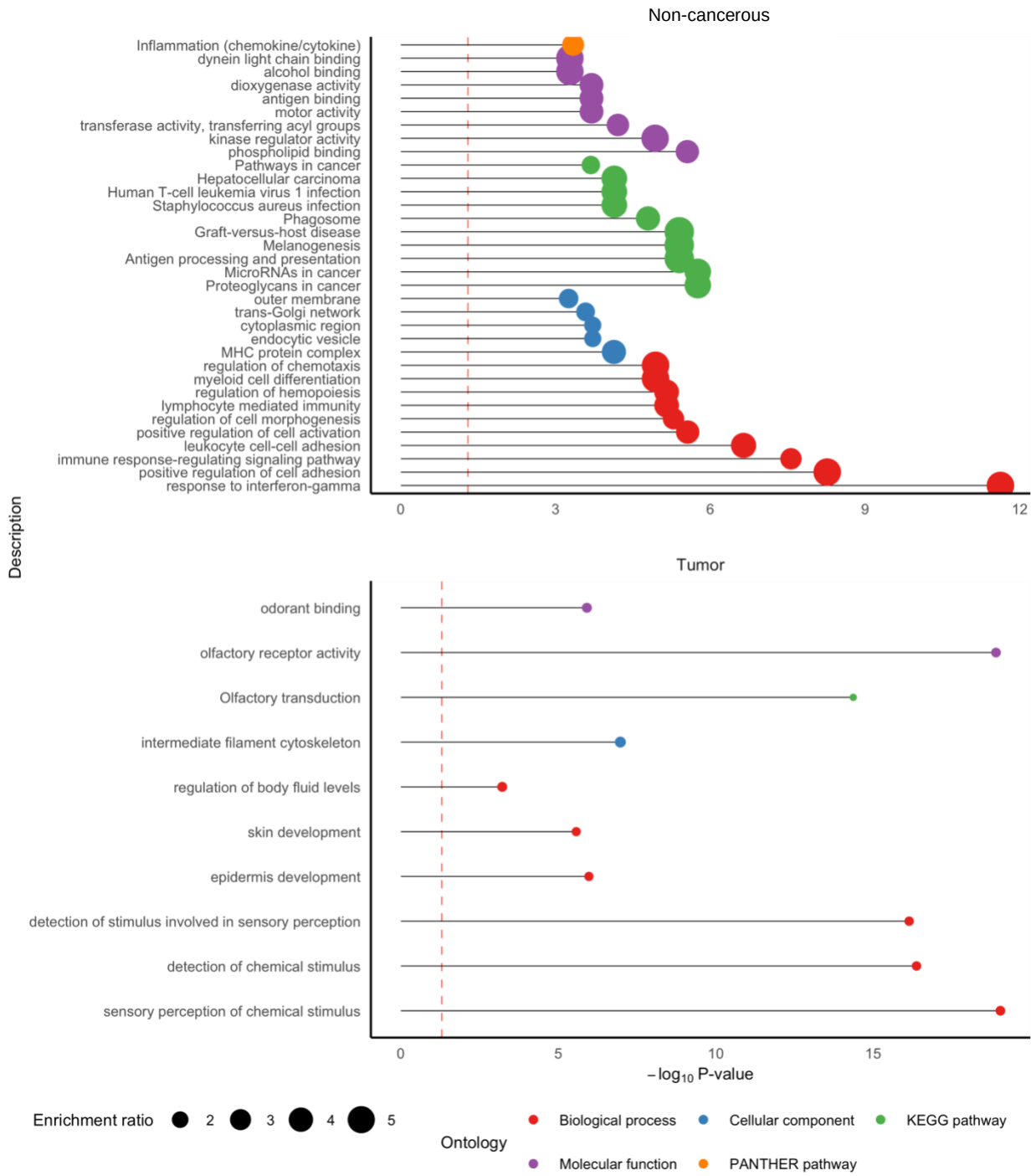
Supplemental Figure S1: *Analysis scheme*. We analyze genetic and transcriptomic data from non-cancerous breast and prostate tissue from GTEx and breast and prostate tumors from TCGA. We optimize eQTL mapping using ordinary least squares regression for numbers of genotype principal components and expression hidden covariates with prior to optimize eQTL discovery. We then conduct a genome-wide local and distal eQTL analysis using the optimized set of covariates. Next, we conduct mediation analysis or gene-based association testing to identify distal-eQTLs of pcGenes that are mediated by local-eQTLs of ncRNAs. Lastly, we integrate eQTLs results with GWAS using colocalization and analysis of genetically-regulated expression.



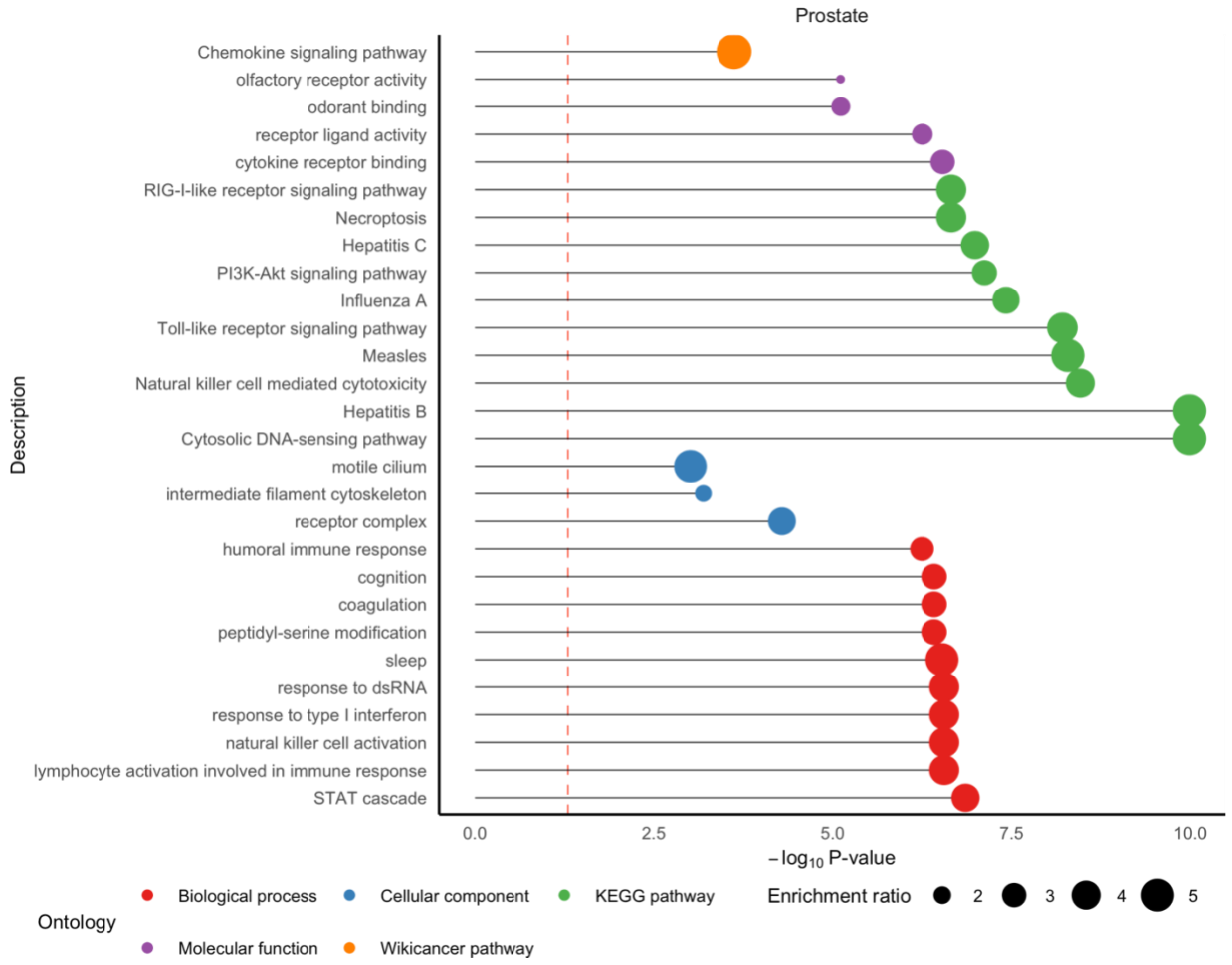
Supplemental Figure S2: *Distribution of total mediation effects, mediation proportions, and gene-level distal effect sizes in healthy and tumor breast and prostate tissue.*



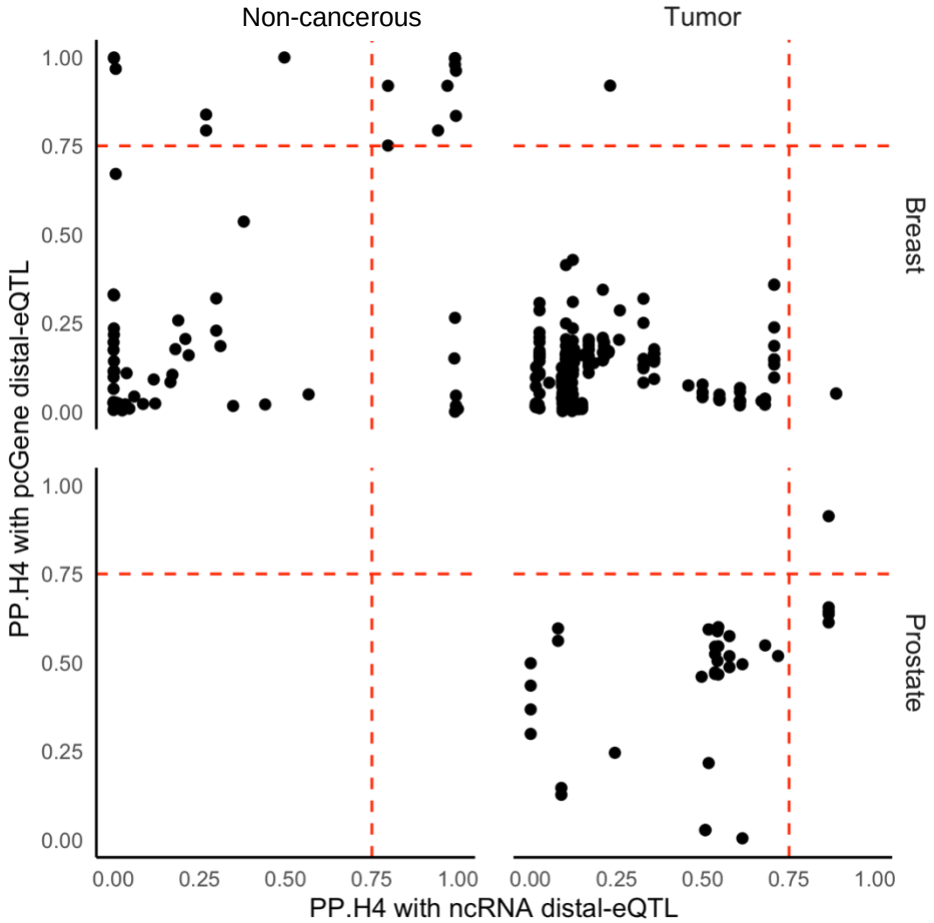
Supplemental Figure S3: *Over-represented ontologies for healthy- or tumor-specific eGenes, compared to all protein-coding genes in the transcriptome*. -$\log_{10}$ P-value of enrichment (X-axis) of over-represented gene sets (Y-axis), with point sized by enrichment ratio and colored by ontology category. Here, combining pcGenes with a distal genetic association across breast and prostate tissue, we compare the set of pcGenes from healthy or tumor state to the universe of all pcGenes in the transcriptome.



Supplemental Figure S4: *Over-represented ontologies for healthy or tumor state-specific eGenes, compared to all protein-coding genes detected in distal-QTL mapping*. -$\log_{10}$ P-value of enrichment (X-axis) of over-represented gene sets (Y-axis), with point sized by enrichment ratio and colored by ontology category. Here, combining pcGenes with a distal genetic association across breast and prostate tissue, we compare the set of pcGenes detected for breast or prostate and compare to the universe of all pcGenes detected across breast and prostate. No enrichments at $P < 0.05$ were detected for breast-specific pcGenes.



Supplemental Figure S5: *Over-represented ontologies for breast or prostate-specific eGenes, compared to all protein-coding genes detected in distal-QTL mapping*. -$\log_{10}$ P-value of enrichment (X-axis) of over-represented gene sets (Y-axis), with point sized by enrichment ratio and colored by ontology category. Here, combining pcGenes with a distal genetic association across healthy and tumor state, we compare the set of pcGenes detected for breast or prostate and compare to the universe of all pcGenes detected across breast and prostate. No enrichments at $P < 0.05$ were detected for breast-specific pcGenes.



Supplemental Figure S6: *Scatterplot of posterior probability of colocalization for local- and distal-eQTLs with cancer risk GWAS*. Red lines show PP.H4 = .75.

SUPPLEMENTAL TABLE LEGENDS

Table S1: *Prioritized ncRNA-mediated distal-eQTLs of pcGenes in healthy and tumor prostate tissue.* We provide the tissue state (healthy or tumor), SNP, ncRNA, pcGene, effect size, P-value, method of detection (mediation analysis or GBAT), and closest GWAS risk SNP and P-value.

Table S2: *Prioritized ncRNA-mediated distal-eQTLs of pcGenes in healthy and tumor breast tissue.* We provide the tissue state (healthy or tumor), SNP, ncRNA, pcGene, effect size, P-value, method of detection (mediation analysis or GBAT), and closest GWAS risk SNP and P-value.

Table S3: *In-silico validation of miRNA-pcGene pairs using TargetScan.* For miRNAs detected to mediated distal-eQTLs of pcGenes, we list miRNAs shown to target the pcGene using TargetScan.

Table S4: *Colocalization results for ncRNA-mediated distal-eQTLs with GWAS signal of cancer risk.* We provide the trait, tissue (breast or prostate), tissue state (healthy or tumor), ncRNA and its location, pcGene and its location, posterior probabilities of colocalization between the ncRNA and pcGene with the GWAS signal, and the colocalized SNP for each signal.

Table S5: *GReX analysis results for ncRNAs and cancer risk.* We provide the trait, tissue (breast or prostate), tissue state (healthy or tumor), ncRNA and its location, effect size, standard error, and Z-score of association.