

1 **A splicing-based multi-tissue joint transcriptome-wide association study**
2 **identifies susceptibility genes for breast cancer**

3 Guimin Gao^{1,#,*}, Julian McClellan^{1,#}, Alvaro N. Barbeira^{2,#}, Peter N. Fiorica¹, James L. Li¹,
4 Zepeng Mu², Olufunmilayo I. Olopade³, Dezheng Huo^{1,2,*}, Hae Kyung Im^{2,*}

5

6 ¹Department of Public Health Sciences, University of Chicago, IL, 60637, USA

7 ²Section of Genetic Medicine, Department of Medicine, University of Chicago, IL, 60637, USA

8 ³Section of Hematology & Oncology, Department of Medicine, University of Chicago, IL,
9 60637, USA

10 [#]These authors made equal contribution.

11 ^{*}Corresponding authors:

12 Guimin Gao, ggao5@bsd.uchicago.edu

13 Dezheng Huo, dhuo@bsd.uchicago.edu

14 Hae Kyung Im, haky@uchicago.edu

15

1 **Abstract**

2 Splicing-based transcriptome-wide association studies (splicing-TWASs) of breast cancer have
3 the potential to identify new susceptibility genes. However, existing splicing-TWASs test
4 association of individual excised introns in breast tissue only and have thus limited power to
5 detect susceptibility genes. In this study, we performed a multi-tissue joint splicing-TWAS that
6 integrated splicing-TWAS signals of multiple excised introns in each gene across 11 tissues that
7 are potentially relevant to breast cancer risk. We utilized summary statistics from a meta-analysis
8 that combined genome-wide association study (GWAS) results of 424,650 European ancestry
9 women. Splicing level prediction models were trained in GTEx (v8) data. We identified 240
10 genes by the multi-tissue joint splicing-TWAS at the Bonferroni corrected significance level; in
11 the tissue-specific splicing-TWAS that combined TWAS signals of excised introns in genes in
12 breast tissue only, we identified 9 additional significant genes. Of these 249 genes, 88 genes in
13 62 loci have not been reported by previous TWASs and 17 genes in 7 loci are at least 1 Mb away
14 from published GWAS index variants. By comparing the results of our splicing-TWASs with
15 previous gene expression-based TWASs that used the same summary statistics and expression
16 prediction models trained in the same reference panel, we found that 110 genes in 70 loci
17 identified by our splicing-TWASs were not reported in the expression-based TWASs. Our results
18 showed that for many genes, expression quantitative trait loci (eQTL) did not show significant
19 impact on breast cancer risk, while splicing quantitative trait loci (sQTL) showed strong impact
20 through intron excision events.

21 **Keywords:** alternative splicing; transcriptome-wide association studies; breast cancer; intron
22 excision; multi-tissue; joint analysis; susceptibility genes.

1 **Introduction**

2 Breast cancer is a complex genetic disorder caused by high-penetrance genes, multiple common
3 variants, and non-genetic factors (i.e. environmental and lifestyle/reproductive factors). To date,
4 genome-wide association studies (GWASs) have identified over 200 loci significantly associated
5 with breast cancer. However, the common susceptibility variants identified by GWASs account
6 for a relatively small proportion of the familial relative risk (1) and specific causal genes in most
7 of these loci have not been identified. To further explore the role of genetic variants on breast
8 cancer, expression-based transcriptome-wide association studies (expression-TWASs) have
9 identified hundreds of genes whose genetically regulated gene expression is significantly
10 associated with breast cancer and its various subtypes (2-5). While expression-TWAS have
11 contributed to our understanding of genetic risk in breast cancer, research has recently shown
12 that RNA splicing is major contributor to complex traits. In fact, splicing quantitative trait loci
13 (sQTLs) and expression quantitative loci (eQTLs) may both have significant effects on
14 phenotypes (6). Splicing-based TWASs (splicing-TWASs) can use splicing information of
15 individual intron excision events (i.e., the read proportions for individual introns within an
16 alternatively excised intron cluster in RNA-seq data), which often cannot be captured by
17 traditional expression-based TWASs that use the total gene expression levels in a gene. Thus,
18 investigating alternative splicing in the context of breast cancer may help identify new
19 susceptibility genes. Indeed, recent studies suggest that alternative splicing plays a critical role in
20 the genetic regulation of breast cancer (7) and a splicing-TWAS has identified 85 genes
21 associated with breast cancer by applying a susceptible transcription factor (sTF)-TWAS method
22 to alternative splicing levels measured at individual excised introns in breast tissue (8).
23 However, existing splicing-TWASs suffer from limited power from the multiple testing

1 correction required to account for the large number of tests used to determine the association
2 individual excised introns with disease phenotypes. We hypothesize the power of splicing-
3 TWAS could be increased by modeling multiple intron excision splicing events jointly—a
4 reasonable approach since the multiple (intron excision) splicing events in a gene may jointly
5 impact the phenotype. Existing splicing-TWASs for breast cancer are also limited by their use
6 breast tissue only. Other tissues potentially relevant to breast cancer development may provide
7 useful information in the identification of genes susceptible to breast cancer (9) and including
8 multiple tissues in splicing-TWASs could prove fruitful.

9 In this study, we performed a joint splicing-TWAS that combine information from
10 multiple excised introns in each gene across multiple tissues that are potentially relevant to breast
11 cancer. We used intron splicing level prediction models for 11 tissues relevant to breast cancer
12 trained in GTEx v8 data (10, 11). We used summary statistics from a meta-analysis of the
13 GWAS results from the Breast Cancer Association Consortium (BCAC) (including 122,977
14 breast cancer cases and 105,974 controls) (1) and GWAS results of 10,534 breast cancer cases
15 and 185,116 controls in UK Biobank (UKB) (12). Finally, we compared results from our current
16 joint splicing-TWAS with those of the previous joint expression-TWAS, which uses the same
17 GWAS summary statistics as the splicing-TWAS and for which the gene expression prediction
18 models are trained in the same 11 tissues in GTEx v8 data.

19 **Results**

20 **Joint splicing-TWAS combines information from multiple intron excision events in a gene**
21 **across multiple tissues.** To create our multi-tissue joint splicing-TWAS, we used summary
22 statistics from a meta-analysis (5) that combined GWAS results on women of European ancestry
23 from the Breast Cancer Association Consortium (BCAC) (1) and GWAS results on European

1 women from UK Biobank (UKB) (5). We used splicing level prediction models for 11 selected
2 tissues potentially relevant to breast cancer (5) (see also Methods). The models were trained on
3 samples with European ancestry from GTEx v8 data (sample sizes ranging from 129 to 670 with
4 a median of 227) using a multivariate adaptive shrinkage (MASH) method (10, 13, 14). In total,
5 we tested 14,528 genes across the 11 tissues with splicing prediction models, including 10,931
6 genes in breast tissue in our splicing TWAS analysis. In the prediction models, splicing levels
7 were quantified at clustered excised introns using short-read RNA-seq data by LeafCutter (15).
8 Specifically, read proportions for the introns in each cluster were estimated and quantile
9 normalized.

10 In our joint splicing-based TWAS analysis, we analyzed genes with splicing prediction
11 models for at least one intron (excision) event. First, we performed a traditional individual
12 intron-based TWAS analysis for all intron excision events in all genes across the genome in each
13 tissue by the S-PrediXcan method (16). Second, for each gene with prediction models for
14 multiple intron excision events, we combined the S-PrediXcan p-values for the multiple introns
15 in each tissue by the aggregated Cauchy association test (ACAT) method (17) to obtain a tissue-
16 specific gene-level p-value. Finally, for the multi-tissue TWAS, we combined the tissue-specific
17 p-values across the 11 selected tissues to generate an overall p-value for each gene using ACAT;
18 we ignored tissues that had no prediction models and, therefore, no p-values.

19 Of the 14,528 genes tested in our multi-tissue joint splicing-TWAS analysis, we
20 identified 240 genes (located in 94 loci) at the Bonferroni corrected significance level ($p <$
21 3.44×10^{-6}) to be associated with breast cancer (Supplementary Table 1). We also created a
22 breast-tissue-specific joint splicing-TWAS using the ACAT method to combine multiple excised
23 introns in a gene in breast tissue only, we identified 158 genes significant at the Bonferroni

1 corrected significance level ($p < 4.57 \times 10^{-6}$), of which 149 genes were also identified by the
2 multi-tissue joint splicing-TWAS that combined information across 11 tissues (Supplementary
3 Table 2). A total of 249 significant genes that were identified by either the multi-tissue or breast-
4 tissue-specific joint splicing-TWAS. Of these 249 genes, 88 have not been reported by previous
5 TWASs (Supplementary Table 1). We found 17 genes in 7 loci that are at least 1 Mb away from
6 previously published GWAS index variants, including 11 genes in 7 loci not reported by
7 previous TWASs (Table 1). Of the 17 genes in Table 1, 12 were identified by both the multi-
8 tissue and breast-tissue-specific joint splicing TWASs and two genes (*AFF1* and *SRP54*) were
9 identified only by the breast-tissue-specific joint splicing-TWAS. We further performed
10 conditional splicing-TWAS tests for the significant genes adjusting for nearby GWAS index risk
11 variants (within ± 2 Mb of the transcription start or stop sites of a gene). The 17 genes remained
12 significant in the conditional splicing-TWASs (see “Conditional ACAT P- value” in
13 Supplementary Table 1 and more details in Methods); this suggests the TWAS signals at the 17
14 genes are independent of nearby GWAS index variants. Our results also suggest that the multi-
15 tissue, joint splicing-TWAS, especially when used in tandem with breast-tissue-specific joint
16 splicing-TWAS, could provide additional information regarding disease susceptibility genes than
17 GWAS or expression-TWAS could provide alone.

18 Of the 17 genes in Table 1 that were identified by our current splicing-TWASs and are at
19 least 1 Mb away from previous GWAS hits, 11 genes have not been identified by any previous
20 published TWASs (Table 2). For each of these genes, we further explored which tissues and
21 which specific excised introns in the tissues had the strongest signals. Table 2 lists the tissue-
22 specific joint splicing-TWAS p-values for breast tissue and/or the tissues with strongest tissue-
23 specific signals; in addition, Table 2 lists the S-PrediXcan z-scores and p-values for excised

1 introns in these tissues. For example, at the gene *FCGR1CP*, there were four intron events with
2 prediction models in whole blood tissue, and the *FCGR1CP* isoforms with introns excised from
3 143,874,823 to 143,875,219 base pairs in chromosome 1 (intron_1_143874823_143875219) had
4 the smallest p-value (4.75×10^{-10}); however, in breast tissue, no excised intron events had
5 prediction models (i.e., intron phenotypes could not be predicted by sQTL variants).
6 Furthermore, In Table 2, 9 of the 11 genes showed significant associations with breast cancer in
7 breast tissue (i. e. with tissue specific splicing-TWAS p-values $< 4.57 \times 10^{-6}$), indicating these
8 genes potentially impact breast cancer risk by splicing in breast tissue.

9

Table 1. The 17 genes identified by splicing-TWASs located at 7 loci at least 1 Mb away from previous GWAS hits.

Cytoband	Gene symbol	Gene position (hg38)	Gene type (v40)	Multi-Tissue ACAT P ¹	Breast-tissue ACAT P ²	Max PIP ³	MAX RCP ⁴	Reported in previous TWASs
1q21.1	FCGR1CP	1:143874793-143883575	pseudogene	1.90E-09	NA	0.994	0	No
4q21.3	AFF1*	4:86935002-87141039	protein_coding	1.33E-05	3.58E-06	0.659	0.001	No
6q24.1	TXLNB	6:139240061-139291998	protein_coding	1.87E-06	NA	0.56	0.364	Yes
	ENSG00000226571	6:139271362-139667284	lncRNA	2.24E-06	NA	0.76	0.452	No
7q22.1	TRIM4	7:99876958-99919531	protein_coding	6.51E-07	4.55E-07	0.0917	0.065	No
	GJC3	7:99923266-99929620	protein_coding	1.06E-06	8.23E-07	0.0577	0.002	No
	AZGP1	7:99966720-99976042	protein_coding	1.06E-06	8.23E-07	0.0577	0.002	No
	PMS2P1	7:100300024-100336307	pseudogene	1.90E-07	2.02E-07	0.0604	0.077	No
	STAG3L5P	7:100336079-100351900	pseudogene	2.52E-07	7.94E-07	0.08	0.235	No
	PILRB	7:100352176-100367831	protein_coding	2.38E-07	2.36E-07	0.144	0.258	Yes
	PILRA	7:100367530-100400096	protein_coding	1.15E-06	7.50E-07	0.0277	0.087	Yes
	ZCWPW1	7:100400826-100428992	protein_coding	3.43E-07	3.57E-07	0.0566	0.104	Yes
	TSC22D4	7:100463359-100479232	protein_coding	4.23E-07	3.42E-07	0.0412	0.075	Yes
	NYAP1	7:100483927-100494802	protein_coding	5.73E-07	NA	0.017	0.016	Yes
14q13.2	SRP54*	14:34981957-35029686	protein_coding	6.71E-06	1.63E-06	0.684	0.379	No
	PRORP	14:35121846-35277622	protein_coding	3.53E-06	3.75E-06	0.774	0.134	No
17p12	ZNF18	17:11977439-11997475	protein_coding	2.98E-06	1.24E-06	0.905	0.456	No

* These two genes in bold are significant in the ACAT test in breast tissue but marginally significant in ACAT test across 11 tissues.

¹ P value of multi-tissue ACAT test that combined multiple intron-based TWAS p-values in a gene across 11 tissues.

² P value of ACAT test that combined multiple intron-based TWAS p-values in a gene in breast tissue only.

³ Maximum PIP for all introns across 11 tissues.

⁴ Maximum RCP for all introns across 11 tissues.

Abbreviation: GWAS, genome-wide association study; TWAS, transcriptome-wide association study; ACAT, aggregated Cauchy association test; PIP, posterior inclusion probability; RCP, regional colocalization probability.

Table 2. The 11 genes identified by splicing-TWASs and were at least 1 Mb away from previous GWAS hits but not identified by previous TWASs.

Gene symbol	Multi-tissue expression-TWAS P ¹	Tissue-specific Splicing-TWAS		Intron-based Splicing-TWAS ⁴		
		Most significant tissue and/or breast tissue ²	P ³	Intron ID (chr_start_end)	Z	P
FCGR1CP	NA	Whole_Blood	1.90E-09	intron_1_143882396_143883295	1.05	2.94E-01
				intron_1_143874823_143875219	6.23	4.75E-10
				intron_1_143880691_143882113	-1.10	2.70E-01
				intron_1_143876306_143880437	-0.22	8.24E-01
AFF1	2.64E-01	Breast_Mammary_Tissue	3.58E-06	intron_4_87047594_87084120	-1.16	2.45E-01
				intron_4_87105845_87108159	-4.79	1.64E-06
				intron_4_87106888_87108159	4.59	4.36E-06
ENSG00000226571	NA	Cells_EBV-transformed_lymphocytes	2.24E-06	intron_6_139283166_139286845	-4.73	2.24E-06
TRIM4	1.71E-05	Cells_Cultured_fibroblasts	2.12E-07	intron_7_99892746_99903218	-5.26	1.47E-07
				intron_7_99902174_99903218	5.42	6.05E-08
				intron_7_99908812_99909565	1.60	1.11E-01
		Breast_Mammary_Tissue	4.55E-07	intron_7_99909660_99919009	-4.45	8.66E-06
				intron_7_99917885_99919009	4.45	8.66E-06
				intron_7_99902174_99903218	5.12	3.08E-07
intron_7_99892746_99903218	-4.92	8.69E-07				
GJC3	NA	Ovary	8.23E-07	intron_7_99968430_99971746	-4.93	8.23E-07
		Breast_Mammary_Tissue	8.23E-07	intron_7_99968430_99971746	-4.93	8.23E-07
AZGP1	4.75E-05	Ovary	8.23E-07	intron_7_99968430_99971746	-4.93	8.23E-07
	4.75E-05	Breast_Mammary_Tissue	8.23E-07	intron_7_99968430_99971746	-4.93	8.23E-07
PMS2P1	NA	Spleen	1.58E-07	intron_7_100332581_100335838	-5.24	1.58E-07
		Breast_Mammary_Tissue	2.02E-07	intron_7_100332581_100335838	-5.24	1.58E-07
				intron_7_100329019_100330235	-5.14	2.82E-07
				intron_7_100332581_100335838	-5.24	1.58E-07
		Ovary	1.58E-07	intron_7_100332581_100335838	-5.24	1.58E-07
		Adipose_Visceral_Omentum	1.58E-07	intron_7_100332581_100335838	-5.24	1.58E-07
Cells_EBV-transformed_lymphocytes	1.58E-07	intron_7_100332581_100335838	-5.24	1.58E-07		
STAG3L5P	NA	Breast_Mammary_Tissue	7.94E-07	intron_7_100345968_100349717	4.76	1.97E-06
				intron_7_100345968_100349731	2.35	1.88E-02

				intron_7_100345968_100349799	-5.12	3.06E-07
		Cells_EBV-transformed_lymphocytes	1.80E-07	intron_7_100345968_100349717	5.25	1.56E-07
				intron_7_100345968_100349731	5.09	3.50E-07
				intron_7_100345968_100349799	-5.27	1.35E-07
SRP54	4.43E-02	Adipose_Visceral_Omentum	9.76E-07	intron_14_35001020_35007283	-5.11	3.25E-07
				intron_14_35019074_35022910	-1.69	9.01E-02
				intron_14_35019074_35028088	1.68	9.38E-02
		Breast_Mammary_Tissue	1.63E-06	intron_14_34999649_35000936	2.73	6.36E-03
				intron_14_35001020_35007283	-5.11	3.25E-07
				intron_14_35019074_35022910	-0.48	6.29E-01
				intron_14_35019074_35028088	1.68	9.38E-02
				intron_14_35023080_35028088	1.72	8.55E-02
PRORP	1.31E-04	Liver	2.03E-06	intron_14_35124231_35126735	4.75	2.03E-06
		Breast_Mammary_Tissue	3.75E-06	intron_14_35122767_35122952	1.44	1.49E-01
				intron_14_35124231_35126735	4.75	2.03E-06
				intron_14_35124231_35127479	-4.75	2.03E-06
				intron_14_35126782_35127479	4.38	1.18E-05
ZNF18	2.52E-03	Breast_Mammary_Tissue	1.24E-06	intron_17_11983407_11984113	-2.33	1.96E-02
				intron_17_11992911_11997431	4.98	6.21E-07

¹ P values from our previous multi-tissue expression-TWAS that used the same summary statistics as in the splicing-TWASs and that used ACAT to combine expression TWAS p-values calculated in 11 tissues (5).

² the most significant tissues denote those with smallest p-values in tissue-based splicing-TWAS, which used ACAT method to combine S-PrediXcan p-values of introns in a gene; breast tissue is also listed if breast tissue splicing-TWAS p-values were available (i.e., intron PrediXction models were available in breast tissue).

³ P values of tissue-specific splicing-TWAS that used ACAT test to combine S-PrediXcan p values of introns in a gene in a specific tissue.

⁴ S-PrediXcan p-values for single introns in single tissues.

1 **Joint splicing-TWASs identify 110 genes not found by expression-TWASs that used the**
2 **same GWAS summary statistics.** We previously performed both a multi-tissue and a breast-
3 tissue-specific expression-TWASs that used the same meta-analysis summary statistics as
4 described above and expression prediction models trained in the same 11 tissues (or breast tissue
5 only) from GTEx v8 (5). We compared the combined results of our current multi-tissue and
6 breast-tissue-specific joint splicing-TWASs with the combined results from our previous multi-
7 tissue and breast-tissue-specific expression-TWASs. The expression-TWASs identified 309
8 genes (5) compared to 249 identified by splicing-TWASs. The splicing-TWASs and expression-
9 TWASs mutually identified 139 genes; the remaining 110 and 170 genes were unique to
10 splicing-TWASs and expression-TWASs, respectively. Supplementary Table 3 lists the
11 comparison of p-values and/or z-scores of the two types of (expression- and splicing-based)
12 TWASs for the 110 genes unique to joint splicing-TWASs. For 83 genes, the expression-TWASs
13 had weak signals but did not reach the Bonferroni corrected significance level ($p \leq 2.59 \times 10^{-6}$).
14 For the remaining 27 genes, no expression prediction models were available in the 11 tissues
15 considered (Supplementary Table 3, genes with “NA” in the column “Expression-TWAS ACAT
16 joint P-value”); because eQTL signals in the reference panel (GTEx V8) were insufficient in cis-
17 gene regions to predict the gene expression levels. While the eQTLs in these cis-gene regions did
18 not show significant impact on breast cancer risk through gene expression in our data analyses,
19 prediction models for excised introns of these genes were created in at least one of the 11 tissues
20 in the reference panel and the splicing-TWASs showed significance in these genes. This
21 evidence suggests sQTL in these cis regions have strong impact on breast cancer risk through the
22 excised introns. As a special case, in Table 2 we also included p-values of our previous multi-
23 tissue expression-TWAS for the 12 genes that were not identified by our expression-TWASs but

1 identified by our splicing-TWASs and are at least 1Mb away from GWAS index SNPs. For five
2 genes (*FCGR1CP*, *ENSG00000226571*, *GJC3*, *PMS2P1*, *STAG3L5P*), we found no available
3 expression prediction models in the 11 tissues but having intron splicing prediction models in at
4 least one of the 11 tissues.

5 **Intron-based fine mapping identifies 114 genes with high posterior inclusion probability.**

6 Our splicing-TWASs were based on S-PrediXcan (16) that were applied to test intron–trait
7 association for individual excised introns in single tissues. The intron–trait association statistics
8 for introns in a linkage disequilibrium (LD) region can be correlated as a function of LD among
9 genetic variants and sQTL weights. As a result, when a causal intron is associated with breast
10 cancer, S-PrediXcan may identify significant intron-trait associations at a set of introns in the LD
11 region, including non-causal introns. To narrow down the list of potential causal genes from the
12 249 genes identified by our splicing-TWASs, we performed intron-based fine mapping on a set
13 of excised introns in an LD region in each tissue using the package FOCUS (18). We calculated
14 marginal posterior inclusion probability (PIP) for each intron. For a gene, we calculated
15 maximum PIP (Max-PIP) for introns in the gene across 11 tissues. We considered genes with a
16 $\text{Max-PIP} > 0.80$ to have a high likelihood of being causal. In our fine-mapping analysis, 114
17 genes had Max-PIP greater than 0.80 (Supplementary Table 4).

18 **Colocalization combined with fine mapping refines list of likely causal genes.**

19 We performed colocalization using the package ENLOC (19) to identify evidence of colocalization between
20 GWAS and sQTL signals by calculating regional colocalization probabilities (RCP). Since
21 ENLOC can only be applied to an intron region (LD block including the intron) in single tissues,
22 for each gene, we calculated the maximum RCP (Max-RCP) across introns in the gene across the
23 11 tissues. We found single introns often had lower RCP compared to gene expression-based

1 colocalization, therefore, we set a threshold of 0.10 for Max-RCPs. Genes with Max-RCP greater
2 than the threshold are more likely to be causal. In our analysis, 88 of 249 genes had Max-RCP
3 values greater than 0.1 (Supplementary Table 1). Overall, 56 genes exceeded both the fine-
4 mapping and colocalization thresholds (Max-PIP greater than 0.80; Max-RCP greater than 0.10),
5 exhibiting strong evidence of being causal genes (Supplementary Table 4, top 56 rows). Still,
6 these genes need to be investigated in future functional experiments.

7 **Gene set enrichment and functional annotation corroborate splicing-TWAS results.** Of the
8 249 genes identified by our joint splicing-TWASs, 216 are protein-coding genes, 19 are long
9 non-coding RNA (lncRNA) genes, and 14 are pseudogenes. We tested the enrichment of the set
10 of 235 protein coding and lncRNA genes against background gene sets from multiple databases
11 using the FUMA software package (20). We limited the enrichment analysis to 33,527
12 background genes in FUMA. Three genes (*ENSG00000281357*, *ENSG00000284237*,
13 *ENSG00000280670*) were not recognized in FUMA. We found the set of 235 genes identified by
14 splicing-TWASs were significantly enriched in 42 background gene-sets at the threshold 0.05 for
15 Bonferroni-adjusted p-values (Supplementary Table 5). These gene sets include a
16 mammographic density set, an alcohol use disorder set, two body fat distribution sets (trunk fat
17 ratio and leg fat ratio), and five breast related sets (breast cancer, estrogen-receptor negative
18 breast cancer, breast size, NIKOLSKY BREAST CANCER 1Q21 AMPLICON, and
19 NIKOLSKY BREAST CANCER 7Q21 Q22 AMPLICON). The enrichment in these gene sets
20 suggests that the genes identified by splicing-TWASs may contribute to breast cancer etiology
21 directly or through their impacts on known lifestyle/environmental risk factors. FUMA also
22 identified differentially expressed gene (DEG) sets (genes which are significantly more or less
23 expressed in a given tissue compared to others) for each of the 30 general tissues that FUMA

1 selected from GTEx v8 data. The genes identified by splicing TWASs showed strong tissue
2 specificity; for example, these genes are significantly enriched in the DEG sets in heart,
3 pancreas, liver, blood, muscle, ovary, cervix uteri, and uterus tissues (Supplementary Figure 2).

4 **Discussion**

5 In this study, we performed a multi-tissue and a breast-tissue-specific joint splicing-TWAS for
6 overall breast cancer risk that combine information from multiple (excised) introns in a gene
7 across multiple tissues (or in breast tissue only). We identified 249 significant genes. Among
8 them, 88 genes in 62 loci have not been reported by previous TWASs; 17 genes in seven loci
9 were at least 1 Mb away from previously published GWAS index variants and the remaining 232
10 genes are located known GWAS susceptibility loci. Of the 17 genes, 11 genes in 7 loci were not
11 reported by previous TWASs.

12 As another focus of this study, we compared the results of two types of TWASs: splicing-
13 and expression-TWASs. Our findings illustrated that multi-tissue and breast-tissue-specific joint
14 splicing-TWASs identified genes that were not identified by the multi-tissue and breast-tissue-
15 specific expression-TWAS when the two types of TWASs used the same summary statistics and
16 prediction models trained in the same reference panel (GTEx v8). These findings suggested that
17 sQTL-based splicing-TWASs may provide different information from eQTL-based expression-
18 TWASs for breast cancer risk and may reveal new insights into genetic etiology of breast cancer.

19 We have checked in the literature the functional importance of the 11 genes (see Table 2)
20 that are at least 1Mb away from published GWAS index SNPs and are not reported by previous
21 TWASs. Here we briefly describe the importance of six genes, *TRIM4*, *GJC3*, *AZGP1*, *AFF1*,
22 *SRP54*, and *ZNF* in cancer biology. Han et al (21) reported that *TRIM4* is downregulated in

1 tamoxifen (TAM)□resistant breast cancer cells, while the loss of TRIM4 is associated with an
2 unfavorable prognosis; In vitro and in vivo experiments confirm that TRIM4 increased estrogen
3 receptor alpha (ER□ α) expression and the sensitivity of breast cancer cells to TAM.

4 *GJC3* and *AZGP1* are two genes also located at the same locus as *TRIM4* at 7q22.1, and
5 the in-frame fusion of these genes (*AZGP1-GJC3*) has previously been reported in both triple-
6 negative breast cancer and prostate cancer cells (22-24). This fusion event is a well-documented
7 transcription-induced chimera (TIC). TICs occur when consecutive genes on a chromosome are
8 spliced together, rendering their fusion product a functional protein. The intron-level significant
9 association ($P=8.23 \times 10^{-7}$) in breast tissue of the same intron (intron_7_99968430_99971746) in
10 both genes suggests that aberrant splicing in breast tissue could play a role in the development of
11 this fusion in breast cancer.

12 *AFF1* is a proto-oncogene and member of the family of ALF transcription elongation
13 factors located on chromosome 4 (25, 26). *AFF1* is translocated to chromosome 11 to fuse with
14 *KMT2A* in nearly 50% of infant acute lymphoblastic leukemias (ALL). In these fusions, the
15 transactivation domain of *AFF1* remains functional. Similarly, *AFF1*'s homolog, *AFF3* retains
16 its transactivation domain when translocated in the minority of ALL *t(4;11)* translocations.
17 Additionally, increased *AFF3* expression has been associated with tamoxifen resistance and
18 breast cancer development in breast ductal acini cells(27, 28). More in-depth splicing
19 quantification of RNA-seq in normal and malignant breast tissue is needed to elucidate the *AFF1*
20 association with breast cancer.

21 *SRP* is a ribonucleoprotein with six subunits that targets proteins to the endoplasmic
22 reticulum as they are translated (29) , and in particular, *SRP54* has been shown to interact and

1 decrease circulating copies of *TP53* in cervical cancer (30). *SRP54* has 23 documented splice
2 variants. In our analysis, the splicing events intron_14_35001020_35007283 in this gene showed
3 strong association with breast cancer in both the breast and adipose visceral omentum tissues
4 (Table 2).

5 Another gene *ZNF18* at the 17p13.3 locus has been previously implicated in multiple
6 cancer sites including breast cancer (31), diffuse large B cell lymphoma (32), clear cell
7 endometrial carcinoma (33), and lung cancer(34). Interestingly, in lung cancer cell lines,
8 overexpression of the tumor suppressor *MEN1* was shown to decrease the isoform abundance of
9 *ZNF18* (34), suggesting that decreased expression of specific isoforms of *ZNF18* may play a role
10 in carcinogenesis. Furthermore, the PIPs of intron excision events of *ZNF18* and *SRP54* in breast
11 and several other tissues were high (PIP > 0.50, Supplementary Table 1) suggesting these genes
12 may contain candidate causal isoforms that affect breast cancer risk.

13 Our splicing-TWASs can identify significant associations of non-causal introns and
14 genes; this is similar to a GWAS, which can identify a susceptibility locus with a set of
15 significant genetic variants, but cannot identify which variants in the locus are causal. The PIPs
16 for individual introns in single tissues can provide useful information about how likely the
17 corresponding genes are causal. However, compared to the PIP for a gene that was calculated
18 based on an expression-TWAS, the PIPs for individual excised introns in a gene seemed
19 relatively smaller on average. It is possible that a joint PIP combining information from multiple
20 excised introns in a gene can be more useful for determining causal genes from the splicing-
21 TWASs identified. We also noticed that the intron-based sQTL colocalization signal is weaker
22 compared to the gene-based eQTL colocalization. We suggest a relatively small sQTL

1 colocalization threshold 0.1 for the Max RCP to indicate association between corresponding genes
2 and breast cancer.

3 We are not the first to attempt splicing-TWASs in the breast cancer context. He et al. (8)
4 proposed an approach by integrating prior knowledge of susceptible transcription factor-
5 occupied cis-regulatory elements (STFCREs) with TWAS (sTF-TWAS) in an effort to improve
6 susceptible gene discovery. By applying their method to individual excised introns in the breast
7 tissue in GTEx v8 and using the summary statistics of BCAC, He et al. performed a splicing-
8 TWAS and identified 85 putative susceptibility genes for breast cancer at a threshold of 0.05 for
9 Bonferroni adjusted p-values. In contrast, by using the same threshold for adjusted p-values, our
10 multi-tissue splicing-TWAS and breast-tissue-specific joint splicing-TWAS identified 240 and
11 158 susceptible genes, respectively. Both of our multi-tissue and breast-tissue-specific joint
12 splicing-TWAS analyses identified substantially more significant genes than He et al., possibly
13 because of several notable differences in methodologies. First, we used GWAS data from a large
14 number of breast cancer cases (N=133,511) and controls (N=291,090) combined from BCAC
15 and UKB, while He et al. used the GWAS summary statistics of BCAC with a total of 122,977
16 cases and 105,974 controls. Second, for each gene, both our multi-tissue and breast-tissue-
17 specific joint TWASs combined splicing-TWAS signals for multiple excised introns in the gene
18 into one test, while He et al. performed multiple tests for the multiple excised introns, which
19 increased the number of tests in the multiple testing correction and may have resulted in lower
20 power. Third, our joint splicing-TWAS combined information across 11 tissues while He et al.
21 only used the breast tissue from GTEx v8. Our results show that the multi-tissue approach
22 identifies more genes compared to splicing-TWAS using breast tissue alone. This suggests that
23 while breast tissue is an important tissue to utilize when conducting breast cancer splicing-

1 TWASs, other tissues can contribute additional information for gene discovery. Fourth, we used
2 splicing prediction models trained in GTEx v8 with the MASH method based on fine mapping to
3 select possible causal sQTLs as predictors for each excised intron. Selecting possibly causal
4 sQTL through fine mapping can reduce the probability that non-causal sQTLs were used in the
5 prediction models (14). In addition, MASH can more accurately estimate the true sQTL effects
6 (i.e., beta coefficients) on intron excision levels by jointly analyzing the sQTL summary statistics
7 estimated in single tissues and accounting for correlation of non-zero sQTL effect sizes across
8 the tissues; the estimates of beta coefficients of sQTLs by MASH were used as final weights in
9 the splicing prediction models.

10 The current study has several limitations. First, although the multi-tissue joint splicing-
11 TWAS identified more genes than breast-tissue-specific splicing-TWAS, it may have generated
12 more false positive hits because: 1) it utilized other tissues that may not be truly causal to breast
13 cancer (35), and 2) it used splicing prediction models trained in 11 tissues; the splicing prediction
14 biases in any tissues may cause false positive findings. This last concern is mediated by the fact
15 that the ACAT method used in our multi-tissue joint splicing-TWAS analysis calculates a
16 weighted average of p-values from multiple tissues and is relatively conservative in identifying
17 significant genes.

18 Second, the current study focused on overall breast cancer risk in women of European
19 ancestry. We are currently working on splicing-TWASs that focus on ER-positive and ER-
20 negative subtypes as well as intrinsic subtypes. In addition, future studies in other racial/ethnic
21 populations are highly desirable. To date, RNA-seq data in the GTEx v8 have a small number of
22 samples from non-European populations, creating a barrier to building accurate prediction
23 models in these populations.

1 **Methods**

2 **GWAS summary statistics and study population:** We used results from a meta-analysis of
3 GWAS summary statistics from BCAC GWAS and GWAS of breast cancer cases extracted from
4 UK Biobank (UKB). BCAC GWAS is composed of 122,977 breast cancer cases and 105,974
5 controls. UKB GWAS includes 10,853 breast cancer cases and 262,614 controls. The details of
6 UKB GWAS and meta-analysis are described in the methods of Gao et al.(5). We performed
7 summary statistic imputation to optimize the accuracy of our GTEx splicing prediction models.

8 **Summary statistic-based imputation.** For variants included in the GTEx prediction models but
9 not in the GWAS summary statistics, we imputed z-scores with the method ImpG-Summary
10 (36). The ImpG-Summary method estimates posterior mean of z-scores at unobserved SNPs
11 based on the assumption that under the null hypothesis of no association, the vector \mathbf{Z} of z-scores
12 at all SNPs in a locus is approximately distributed as a Gaussian distribution, $\mathbf{Z} \sim N(\mathbf{0}, \Sigma)$, where
13 Σ is the correlation matrix among all pairs of SNPs induced by LD. We used the GWAS
14 summary statistics and correlation matrix estimated by using the genotype data in the GTEx
15 samples as input of the ImpG-Summary method.

16 **Quantification of RNA splicing with LeafCutter.** Li et al. (15) proposed an approach
17 LeafCutter for the quantification of alternative splicing events by focusing on intron excisions
18 (rather than whole isoform quantification). Leafcutter quantifies RNA splicing variation using
19 short-read RNA-seq data. The core idea is to leverage spliced reads (reads that span an intron) to
20 quantify (differential) intron usage across samples. Specifically, to identify alternatively excised
21 introns, LeafCutter pools all mapped reads from a study and finds overlapping introns
22 demarcated by split reads. LeafCutter then constructs a graph that connects all overlapping
23 introns that share a donor or acceptor splice site. The connected components of the graph form

1 clusters, which represent alternative intron excision events. Then LeafCutter estimate read
2 proportions for all introns within alternatively excised intron clusters; the read proportions can be
3 further standardized across individuals for each intron and quantile normalized across introns and
4 then used as intron phenotype matrix for sQTL analysis or prediction model construction.

5 **Selection of tissues:** For our multi-tissue joint splicing TWAS, we selected 11 tissues from the
6 GTEx v8 data that are potentially relevant to breast cancer development or carcinogen
7 metabolism (5), including female tissues (breast, ovary, uterus, and vagina), tissues that resemble
8 connective and fat tissues in the breast (subcutaneous adipose, visceral adipose, and cultured
9 fibroblasts), tissues related to immune cells (spleen, EBV-transformed lymphocytes, and whole
10 blood), and liver.

11 **Intron splicing prediction models:** Splicing prediction models were originally built in 49
12 tissues in GTEx (v8) samples of European ancestry that have the genotype and RNA-seq data;
13 each of these 49 tissues has sample size in each tissue greater than 70. Sample size less than 70
14 may result in inaccuracy in prediction (10, 11, 14). Specifically, the prediction models were built
15 with the following steps: 1) cis-sQTL analysis was performed by using fastQTL (37) in each
16 tissue with the intron excision phenotypes (i.e., proportions standardized and then normalized by
17 LeafCutter) (see previous section). For each intron excision event, all variants within the cis-
18 window ($\pm 1\text{Mb}$) with $\text{MAF} > 0.01$ were considered and the following covariates were corrected in
19 the linear regression models: sex, WGS platform, WGS library preparation protocol, top 5
20 genetic principal components, and PEER factors (10, 11). 2) Fine mapping was performed for
21 each intron and its cis-region in each tissue by the dap-g method (19, 38) to select variants with
22 minor allele frequency > 0.01 and posterior inclusion probabilities (PIPs) > 0.01 and to select
23 excised introns with at least one credible set that had $\text{PIP} > 0.1$ (where the credible set PIP is

1 sum of PIPs of variants in the set). Then in each credible set, only the variant with the highest
2 PIP was kept. For the 49 tissues, a union of selected variants across 49 tissues was obtained and
3 LD pruning was applied to the union of variants to remove redundant variants. 3) The
4 multivariate adaptive shrinkage method (13) was used to estimate the true effects at the selected
5 sQTL variants by jointly analyzing the marginal effect sizes and standard errors (SEs) of the
6 sQTLs across the 49 tissues accounting for correlation among nonzero effects in different tissues
7 (Barbeira, 2021 Genome Biology). 4) The predicted intron splicing level in each tissue was
8 calculated as the linear combination of genotypes multiplying by their estimated effect sizes at
9 the selected variants. In this study, we used the prediction models for 11 tissues potentially
10 relevant to breast cancer. It is possible no prediction models could be constructed for some intron
11 splicing events in some tissues because there are no strong sQTL signals for the intron
12 phenotypes.

13 **Joint Splicing-TWAS test for multiple excised introns in a gene across multiple tissues.**

14 Suppose there are J excised introns in a gene with prediction models. The joint TWAS analysis
15 generate a p-value for the gene by three steps: 1) performing traditional TWAS test for each
16 intron in each of the 11 tissues by the software S-PrediXcan to obtain the p-values p_{jk} ($j =$
17 $1, \dots, J; k = 1, \dots, 11$), where j denotes j -th intron and k denotes k -th tissue. 2) generating a tissue-
18 specific p-value $p_{ACAT,k}$ for k -th tissue by constructing a tissue specific test statistic $T_{ACAT,k}$ with
19 the ACAT method that combines p-values p_{jk} ($j = 1, \dots, J$) of all J introns with prediction
20 models in the gene. Specifically, the ACAT test statistic is $T_{ACAT,k} = \sum_{j=1}^J w_{jk} \tan((0.5 -$
21 $p_{jk})\pi)$, where w_{jk} are nonnegative weights. We used $w_{jk} = 1/J$. The tissue specific p-value
22 $p_{ACAT,k}$ of the ACAT test statistic is approximated by $p_{ACAT,k} = \frac{1}{2} - (\arctan T_{ACAT,k})/\pi$. 3)

1 suppose the gene has p-values in $K \leq 11$ tissues, we generate a joint p-value for the gene by the
2 ACAT method again to combine the K tissue specific p-values $p_{ACAT,k}$ by a similar way as in
3 step 2 except using weight $=1/K$.

4 **Conditional joint TWAS.** To test if the signals at the 249 genes identified by our multi-tissue
5 and breast-tissue-specific splicing- TWASs are independent of previously published GWAS
6 index SNPs that were genome-wide significant ($p < 5 \times 10^{-8}$), we performed splicing-TWASs that
7 were conditional on these index SNPs. For each intron excision event, we defined two sets of
8 SNPs: the target set of SNPs used for predicting the intron phenotype and the conditioning set of
9 significant index SNPs from published GWASs within ± 2 Mb of the transcription start or stop
10 sites of the gene. By using the conditional and joint multiple-SNP (COJO) analysis method of
11 Yang et al (39), for the target set of SNPs, we calculated adjusted effects (beta) on breast cancer
12 risk and standard deviation of adjusted beta conditioning on the conditioning set of index SNPs.
13 After performing COJO, we applied S-PrediXcan to these conditional summary statistics in
14 single tissues and performed joint splicing TWAS to combine p-values from single introns in a
15 gene and across individual tissues with the ACAT method.

16 **Intron-based colocalization analysis.** For intron splicing events in the 249 genes that were
17 identified by our splicing-TWASs, we calculated RCPs by the method ENLOC in each of the 11
18 tissues. ENLOC divides the genome into roughly independent LD blocks using the approach
19 described in Berisa & Pickrell (40). For an intron located in a specific LD block, we calculated
20 the colocalization probability of causal GWAS hits and causal sQTLs in the LD block by
21 ENLOC. We used the GTEx (v8) sQTLs for the intron and the meta-analysis GWAS summary
22 statistics in the LD block. For a gene with multiple introns, we assigned the maximum RCP
23 across the 11 tissues as the gene-level RCP.

1 **Intron-based fine mapping.** We performed intron-based statistical fine-mapping over the
2 intron-trait association signals from S-PrediXcan using the software package FOCUS. For a LD
3 block, we estimated a number of intron sets, each contained the causal introns at a predefined
4 confidence level ρ (that is, ρ -credible gene sets; for example, $\rho = 90\%$). We also computed the
5 marginal PIP for each intron in the region to be causal given the observed TWAS statistics
6 calculated from S-PrediXcan. FOCUS accounts for the correlation structure induced by LD and
7 prediction weights used in the TWAS and controls for certain pleiotropic effects. FOCUS takes
8 as input GWAS summary data, intron prediction weights, and LD among all SNPs in the LD
9 region. We applied FOCUS to each of the 11 tissues and related splicing prediction weights from
10 the GTEx v8. We assigned the maximum PIP of all introns across all tissues to a gene as gene-
11 level PIP.

12 **Gene Set Enrichment and Functional Annotation.** For the set of 249 significant genes
13 identified by our splicing-TWASs, we conducted enrichment of 235 protein-coding and lncRNA
14 genes against gene sets from multiple biological pathways, functional categories, and databases
15 by the FUMA package. Specifically, we used the GENE2FUNC module of FUMA and specified
16 33,527 protein-coding and lncRNA genes as the background genes for enrichment testing.
17 Multiple testing correction was performed per data source of tested gene sets (e.g., canonical
18 pathways, GWAScatalog categories) using Bonferroni adjustment. We reported
19 pathways/categories with adjusted p-value ≤ 0.05 and at least 2 genes that overlapped with the
20 gene set of interest.

21 **Multi-tissue expression-TWAS:** Our previous multi-tissue expression-TWAS (5) includes
22 two steps: 1) performing a traditional TWAS analysis in each of the 11 tissues by the software S-
23 PrediXcan to obtain the p-values p_k ($k = 1, \dots, 11$), and 2) constructing test statistic by the

1 ACAT method that combined p-values for each gene from the single tissue TWAS analyses
2 across the 11 tissues. Gene expression prediction models were built with the genotype and RNA-
3 seq data in 49 tissues of European ancestry from the GTEx project (v8) by a similar approach as
4 described in the Section of Intron splicing prediction models and the prediction models for 11
5 tissues were used for the multi-tissue expression-TWASs (5). We used the summary statistics
6 from meta-analysis of the BCAC GWAS and UKB GWAS results. Of the 19,274 genes tested in
7 our joint expression-TWAS analysis, we identified 299 genes whose predicted expression was
8 associated with breast cancer risk at the Bonferroni-corrected significance level ($p < 2.59 \times 10^{-6}$).
9 Only 141 genes were identified when TWAS analysis used only breast tissue, i.e. conventional
10 single-tissue TWAS approach. Of these 141 genes, 131 genes were also identified in the multi-
11 tissue TWAS. The remaining 10 genes identified only in the breast-tissue TWAS analysis were
12 also marginally significant in the multi-tissue TWAS ($p < 0.05$), so we considered the 309 genes
13 from either expression-TWASs in this study for comparison with the results of splicing-TWASs.

14 **Figure and Table Legends**

15 **Supplementary Figure 1.** Comparison of joint splicing-TWAS and joint expression-TWAS
16 Manhattan plots.

17 **Supplementary Figure 2.** Differential analysis of expression of the splicing-TWAS identified
18 genes in GTEx v8 shows tissue specificity. Significantly enriched differentially expressed gene
19 sets (Bonferoni adjusted $p < 0.05$) are highlighted in red. The P values were from
20 hypergeometric test.

21 **Table 1.** The 17 genes identified by splicing-TWASs located at 7 loci at least 1 Mb away from
22 previous GWAS hits.

1 **Table 2.** The 11 genes that were identified by splicing-TWASs and were at least 1 Mb away
2 from previous GWAS hits but not identified by previous TWASs.

3 **Supplementary Table 1.** The 249 genes identified by multi-tissue or breast-tissue-specific joint
4 splicing-TWAS.

5 **Supplementary Table 2.** The 158 genes identified by breast-tissue-specific splicing-TWAS.

6 **Supplementary Table 3.** The 110 genes identified by our multi-tissue or breast-tissue-specific
7 splicing-TWAS but not by our previous multi-tissue or breast-tissue-specific expression-TWAS
8 using the same GWAS summary statistics.

9 **Supplementary Table 4.** The 114 candidate causal genes identified by fine-mapping analysis
10 (sorted by Max RCP).

11 **Supplementary Table 5.** Significant gene sets in the enrichment analysis using FUMA.

12 **Declaration of interests**

13 Dr. Olopade reported receiving grants from Tempus (scientific advisory board) during the
14 conduct of the study; being cofounder of CancerIQ, serving as a board of director member for
15 54gene, and receiving grants from Color Genomics (research support) and grants from Roche
16 (clinical trial support for IIT) outside the submitted work. No other disclosures were reported.

17 **Acknowledgements**

18 This work was supported by the National Cancer Institute (R01 CA242929, R01 CA228198, P20
19 CA233307), Breast Cancer Research Foundation (BCRF-22-071), National Institutes of Health
20 (R01 MD013452), and the NIDDK (P30 DK20595). For BCAC data, the breast cancer genome-
21 wide association analyses were supported by the Government of Canada through Genome

1 Canada and the Canadian Institutes of Health Research, the ‘Ministère de l’Économie, de la
2 Science et de l’Innovation du Québec’ through Genome Québec and grant PSR-SIIRI-701, The
3 National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research UK
4 (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-
5 2009-223175 and H2020 633784 and 634935). All studies and funders are listed in Michailidou
6 et al (Nature, 2017). We thank Yang Li for consulting on quantification of intron excisions and
7 Sarah Sumner for help editing the paper.

8 **Author contributions**

9 G.G., D.H., H.K.I. conceived the study and contributed to the study design. J.M., G.G., A.N.B.,
10 P.N.F., Z.M., D.H., and H.K.I. performed data analyses. G.G., and P.N.F. wrote the first version
11 of the manuscript. G.G., J.M., A.N.B., P.N.F., J.L.L., O.I.O., D.H., and H.K.I revised the
12 manuscript.

13 **Web resources**

14 PrediXcan GTEx v8 MASHR models, <https://predictdb.org/>; Summary statistics of meta-analysis
15 of BCAC and UKB, <https://zenodo.org/record/7814694#.ZDaspXbMK5d>; FUMA
16 software, <http://fuma.ctglab.nl>; COJO (GCTA), <https://yanglab.westlake.edu.cn/software/gcta/>;
17 Enloc, <https://github.com/xqwen/integrative>; S-PrediXcan,
18 <https://github.com/hakyimlab/MetaXcan> and [https://github.com/hakyimlab/summary-gwas-](https://github.com/hakyimlab/summary-gwas-imputation)
19 [imputation; UK Biobank, http://ukbiobank.ac.uk](http://ukbiobank.ac.uk)

20 **Data and code availability**

21 In this study, we only used existing datasets that are publicly available (see the section of Web
22 resources). The code pipeline and results for our multi-tissue joint splicing-TWAS analysis will

1 be available at <https://zenodo.org/>. For specific method code, we made minor modifications to S-
2 PrediXcan to combine results with ACAT
3 (https://github.com/shugamoe/MetaXcan/tree/catch_up). We also made minor modifications to
4 FOCUS to accommodate PrediXcan GTEx v8 MASHR models
5 (<https://github.com/shugamoe/focus>).

6 **References:**

- 7 1. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association
8 analysis identifies 65 new breast cancer risk loci. *Nature*. 2017;551(7678):92-4.
- 9 2. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide
10 association study of 229,000 women identifies new candidate susceptibility genes for breast
11 cancer. *Nat Genet*. 2018;50(7):968-78.
- 12 3. Feng H, Gusev A, Pasaniuc B, Wu L, Long J, Abu-Full Z, et al. Transcriptome-wide
13 association study of breast cancer risk by estrogen-receptor status. *Genet Epidemiol*.
14 2020;44(5):442-68.
- 15 4. Jia G, Ping J, Shu X, Yang Y, Cai Q, Kweon SS, et al. Genome- and transcriptome-wide
16 association studies of 386,000 Asian and European-ancestry women provide new insights into
17 breast cancer genetics. *Am J Hum Genet*. 2022;109(12):2185-95.
- 18 5. Gao G, Fiorica PN, McClellan J, Barbeira AN, Li JL, Olopade OI, et al. A joint
19 transcriptome-wide association study across multiple tissues identifies candidate breast cancer
20 susceptibility genes. *Am J Hum Genet*. 2023;110(6):950-62.
- 21 6. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a
22 primary link between genetic variation and disease. *Science*. 2016;352(6285):600-4.
- 23 7. Koedoot E, Wolters L, van de Water B, Devedec SEL. Splicing regulatory factors in
24 breast cancer hallmarks and disease progression. *Oncotarget*. 2019;10(57):6021-37.
- 25 8. He J, Wen W, Beeghly A, Chen Z, Cao C, Shu XO, et al. Integrating transcription factor
26 occupancy with transcriptome-wide association analysis identifies susceptibility genes in human
27 cancers. *Nat Commun*. 2022;13(1):7118.
- 28 9. Ferreira MA, Gamazon ER, Al-Ejeh F, Aittomaki K, Andrulis IL, Anton-Culver H, et al.
29 Genome-wide association and transcriptome studies identify target genes and risk loci for breast
30 cancer. *Nat Commun*. 2019;10(1):1741.
- 31 10. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al.
32 Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome biology*.
33 2021;22(1):49.
- 34 11. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human
35 tissues. *Science*. 2020;369(6509):1318-30.
- 36 12. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank
37 resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-9.
- 38 13. Uebachs SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for
39 estimating and testing effects in genomic studies with multiple conditions. *Nat Genet*.
40 2019;51(1):187-95.

- 1 14. Barbeira AN, Melia OJ, Liang Y, Bonazzola R, Wang G, Wheeler HE, et al. Fine-
2 mapping and QTL tissue-sharing information improves the reliability of causal gene
3 identification. *Genet Epidemiol.* 2020;44(8):854-67.
- 4 15. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-
5 free quantification of RNA splicing using LeafCutter. *Nat Genet.* 2018;50(1):151-8.
- 6 16. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al.
7 Exploring the phenotypic consequences of tissue specific gene expression variation inferred from
8 GWAS summary statistics. *Nat Commun.* 2018;9(1):1825.
- 9 17. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation
10 under arbitrary dependency structures. *J Am Stat Assoc.* 2020;115(529):393-402.
- 11 18. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic
12 fine-mapping of transcriptome-wide association studies. *Nat Genet.* 2019;51(4):675-82.
- 13 19. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic
14 association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.*
15 2017;13(3):e1006646.
- 16 20. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and
17 annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.
- 18 21. Han D, Wang L, Long L, Su P, Luo D, Zhang H, et al. The E3 Ligase TRIM4 Facilitates
19 SET Ubiquitin-Mediated Degradation to Enhance ER-alpha Action in Breast Cancer. *Adv Sci*
20 *(Weinh).* 2022;9(25):e2201701.
- 21 22. Jung J, Jang K, Ju JM, Lee E, Lee JW, Kim HJ, et al. Novel cancer gene variants and
22 gene fusions of triple-negative breast cancers (TNBCs) reveal their molecular diversity
23 conserved in the patient-derived xenograft (PDX) model. *Cancer Lett.* 2018;428:127-38.
- 24 23. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion
25 detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics (Oxford, England).*
26 2011;27(8):1068-75.
- 27 24. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, et al. Deep RNA sequencing
28 analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples.
29 *BMC Med Genomics.* 2011;4:11.
- 30 25. Heerema NA, Sather HN, Ge J, Arthur DC, Hilden JM, Trigg ME, et al. Cytogenetic
31 studies of infant acute lymphoblastic leukemia: poor prognosis of infants with t(4;11) - a report
32 of the Children's Cancer Group. *Leukemia.* 1999;13(5):679-86.
- 33 26. Domer PH, Fakharzadeh SS, Chen CS, Jockel J, Johansen L, Silverman GA, et al. Acute
34 mixed-lineage leukemia t(4;11)(q21;q23) generates an MLL-AF4 fusion product. *Proc Natl Acad*
35 *Sci U S A.* 1993;90(16):7884-8.
- 36 27. Shi Y, Zhao Y, Zhang Y, AiErken N, Shao N, Ye R, et al. AFF3 upregulation mediates
37 tamoxifen resistance in breast cancers. *J Exp Clin Cancer Res.* 2018;37(1):254.
- 38 28. To MD, Faseruk SA, Gokgoz N, Pinnaduwege D, Done SJ, Andrulis IL. LAF-4 is
39 aberrantly expressed in human breast cancer. *Int J Cancer.* 2005;115(4):568-74.
- 40 29. Kellogg MK, Miller SC, Tikhonova EB, Karamyshev AL. SRPassing Co-translational
41 Targeting: The Role of the Signal Recognition Particle in Protein Targeting and mRNA
42 Protection. *Int J Mol Sci.* 2021;22(12).
- 43 30. Abdelmohsen K, Panda AC, Kang MJ, Guo R, Kim J, Grammatikakis I, et al. 7SL RNA
44 represses p53 translation by competing with HuR. *Nucleic acids research.* 2014;42(15):10099-
45 111.

- 1 31. Sun X, Luo Z, Gong L, Tan X, Chen J, Liang X, et al. Identification of significant genes
2 and therapeutic agents for breast cancer by integrated genomics. *Bioengineered*.
3 2021;12(1):2140-54.
- 4 32. Zhu QY. Bioinformatics analysis of the pathogenic link between Epstein-Barr virus
5 infection, systemic lupus erythematosus and diffuse large B cell lymphoma. *Sci Rep*.
6 2023;13(1):6310.
- 7 33. O'Hara AJ, Le Gallo M, Rudd ML, Bell DW. High-resolution copy number analysis of
8 clear cell endometrial carcinoma. *Cancer Genet*. 2020;240:5-14.
- 9 34. Jin B, Zhu J, Pan T, Yang Y, Liang L, Zhou Y, et al. MEN1 is a regulator of alternative
10 splicing and prevents R-loop-induced genome instability through suppression of RNA
11 polymerase II elongation. *Nucleic acids research*. 2023;51(15):7951-71.
- 12 35. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et
13 al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*.
14 2019;51(4):592-9.
- 15 36. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate
16 imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*
17 (Oxford, England). 2014;30(20):2906-14.
- 18 37. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL
19 mapper for thousands of molecular phenotypes. *Bioinformatics* (Oxford, England).
20 2016;32(10):1479-85.
- 21 38. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient Integrative Multi-SNP Association
22 Analysis via Deterministic Approximation of Posteriors. *Am J Hum Genet*. 2016;98(6):1114-29.
- 23 39. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional
24 and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants
25 influencing complex traits. *Nat Genet*. 2012;44(4):369-75, s1-3.
- 26 40. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in
27 human populations. *Bioinformatics* (Oxford, England). 2016;32(2):283-5.