

---

# AI-based histopathology phenotyping reveals germline loci shaping breast cancer morphology

---

Shubham Chaudhary<sup>1,2,3,✉</sup>, Almut Voigts<sup>1,4</sup>, Sergey Vilov<sup>5</sup>,  
Matthias Heinig<sup>3,5</sup>, Francesco Paolo Casale<sup>1,2,3,✉</sup>

<sup>1</sup> Institute of AI for Health, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

<sup>2</sup> Helmholtz Pioneer Campus, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

<sup>3</sup> School of Computation, Information and Technology, Technical University of Munich, Garching, Germany

<sup>4</sup> TUM School of Medicine and Health, TU Munich and Klinikum Rechts der Isar, Munich, Germany

<sup>5</sup> Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

✉shubham.chaudhary@helmholtz-munich.de, francescopaolo.casale@helmholtz-munich.de

## Abstract

AI foundation models have transformed cancer histopathology by enabling rich, data-driven feature extraction from H&E-stained whole-slide images. However, their application to studying how germline variation shapes tumor morphology remains limited. Here, we perform the first genome-wide association study of breast cancer morphology, independently analyzing AI-derived features from histology images and diagnostic pathology reports. Analyzing H&E slides from 753 patients with matched germline data, we identified six genome-wide significant loci associated with either imaging or textual features, two of which replicated across modalities. We then linked these two loci to histological features described in pathology reports, visual histological features through generative modelling, gene expression modules and patient survival. We found that *rs819976* in *ATAD3B* is associated with disorganized, necrotic tumor morphology, poor-prognosis expression programs, and clinical features including invasive lobular carcinoma and ER positivity. These findings demonstrate the power of AI-based histology to uncover and characterize germline variants that shape tumor morphology, and assess their clinical significance.

## 1 Introduction

Understanding how germline genetic variation influences cancer progression remains a fundamental challenge in oncology. While most studies have focused on variants that predispose individuals to cancer onset [Menden et al., 2018, Chatrath et al., 2020], accumulating evidence suggests that the genetic architecture underlying tumor subtype, progression and prognosis is often distinct from that of susceptibility [Escala-Garcia et al., 2021, Zhu et al., 2022, Escala-Garcia et al., 2019, Morra et al., 2021].

Tissue morphological features visible in histopathology are prognostically informative and strongly associated with clinical outcomes [Morra et al., 2021, Zhu et al., 2022]. Despite their clinical relevance, these features remain underutilized in germline association studies, limiting discovery of regulatory variants that influence tumor morphology and clinical course.

To address this gap, we present the first germline genome-wide association study (GWAS) of breast cancer histology that jointly analyzes phenotypes derived from both histology images and diagnostic pathology reports. Image-based phenotypes are extracted from H&E-stained slides using a foundation model trained on histopathology [Chen et al., 2024], while text-based phenotypes are derived from diagnostic reports using semantic embeddings from a large language model [OpenAI, 2023].

Using matched histology and germline genotypes from 753 breast cancer patients in The Cancer Genome Atlas (TCGA), we identify genetic loci associated with morphological variation captured independently in image and text modalities. Loci that replicate across both provide modality-independent support for germline contributions to tumor architecture. We further integrate generative modeling, gene expression, structured clinical traits, and survival data to interpret these associations in morphological and clinical terms. Together, our results position AI-derived histological phenotypes as a scalable and interpretable axis for germline discovery in oncology.

## 2 Related Work

**Genetic analysis of cancer subtypes.** Although germline variants can impact various aspects of tumor biology, including metastatic potential [Escala-Garcia et al., 2021], drug response [Menden et al., 2018], and survival in specific patient subgroups [Morra et al., 2021], genetic studies of cancer histology have traditionally focused on somatic alterations associated with prognosis, such as driver mutations and chromosomal changes [Wagner et al., 2023, Fu et al., 2020, Lindeman et al., 2013, Woodman et al., 2012, Russnes et al., 2017, Kather et al., 2019, Cao et al., 2020]. In contrast, our study presents the first GWAS of breast cancer histology.

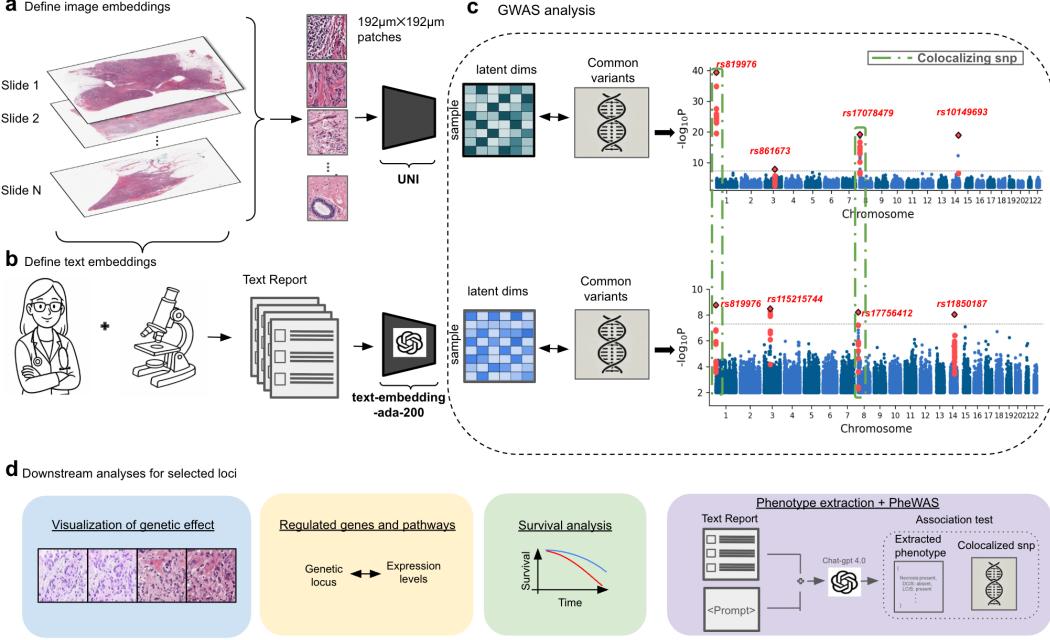
**Pathology foundation models.** We use a pre-trained pathology foundation model to automatically quantify histological features from breast cancer slides. These models generate compact representations of tissue morphology and support diverse downstream tasks [Wang et al., 2022a, Azizi et al., 2023, Saldanha et al., 2023, Lu et al., 2023, Mokhtari et al., 2023]. They fall into two main categories: self-supervised models that learn from unlabeled tissue [Chen and Krishnan, 2022, Wang et al., 2022a, Azizi et al., 2023, Kang et al., 2023, Li et al., 2021, Lazard et al., 2023], and multimodal models that align images and text to support diagnostic interpretation [Huang et al., 2023, Lu et al., 2024]. We adopt the self-supervised UNI model [Chen et al., 2024] to extend GWAS discovery to histological traits beyond those explicitly labeled in clinical records. This work presents the first genetic analysis of cancer histology features derived from foundation model representations.

**Multivariate GWAS.** We assess associations between genetic variants and histological embeddings using a multivariate GWAS framework designed for multi-trait analysis. Among available methods [Wang et al., 2016, Casale et al., 2015, Lippert et al., 2014, Turley et al., 2018, Furlotte and Eskin, 2015], we employ a recent approach developed specifically for histology embeddings [Chaudhary et al., 2024], which performs GWAS on top latent factors to retain statistical power [Kirchler et al., 2022, Xie et al., 2024, Yun et al., 2024, Chaudhary et al., 2024].

**Generative models.** We use conditional generative models to visualize allele-specific changes in tissue morphology through the synthesis of high-resolution histology images conditioned on embeddings. Generative models have been widely used in computational biology to explore latent representations and simulate system perturbations [Goodfellow et al., 2014, Mirza, 2014, Lamiable et al., 2023, Lotfollahi et al., 2023, Palma et al., 2023]. Building on recent work [Chaudhary et al., 2024], our approach directly inverts the UNI pathology foundation model, enabling interpretable visualizations of how genetic variants influence tumor architecture.

## 3 Methods: HistoGWAS for Cancer Histology

Our analysis follows three main stages: (i) defining and validating image- and text-based embeddings (**Figure 1a–b**); (ii) identifying genetic variants associated with these embeddings through multivariate GWAS (**Figure 1c**); and (iii) interpreting the biological relevance of significant loci via downstream analyses (**Figure 1d**).



**Figure 1: Overview of analysis workflow for genetic discovery from breast cancer histology.** (a) Image embeddings: H&E-stained whole-slide images are divided into 192 $\mu$ m × 192 $\mu$ m patches and encoded using the pre-trained UNI model [Chen et al., 2024]. Patch embeddings are aggregated via average pooling to generate sample-level image embeddings. (b) Text embeddings: Diagnostic pathology reports are embedded using openAI text-embedding-ada-200. (c) Genetic association: Multivariate GWAS is performed using histological (top) and textual (bottom) embeddings as phenotypes. The two Manhattan plots summarize genome-wide results for each modality, with colocalizing loci across image and text highlighted in green. (d) Downstream analysis for selected variants: Visualization of allele-specific histological changes using a generative model; gene expression-based association and pathway analysis; survival analysis; and association with structured traits extracted from diagnostic reports via GPT prompting.

### 3.1 Defining image and text embeddings

**Histology data preprocessing.** We curated a dataset from the TCGA breast cancer cohort [Ciriello et al., 2015, 13 et al., 2012], selecting samples with both histological and germline genetic data, yielding 1,054 breast tissue slides from unique individuals. Patches were defined on each slide at the highest resolution (192  $\mu$ m × 192  $\mu$ m) using a regular grid spanning the entire tissue section. Slides were converted to grayscale, and binary tissue masks were generated using `cv2.threshold` from the OpenCV library [Bradski, 2000]. Patches with at least 50% tissue content were retained, resulting in 5,278,793 high-resolution (256 × 256 pixels, 0.75  $\mu$ m/pixel) images for downstream analysis.

**Definition of image embeddings.** To extract high-dimensional features from histology patches, we applied the pre-trained UNI self-supervised foundation model, which has demonstrated strong performance on histopathology tasks [Chen et al., 2024]. Importantly, this model was trained across diverse staining protocols and acquisition conditions, yielding representations that are robust to scanner and staining variability [Chen et al., 2024]. This produced 5,278,793 embeddings (1,024 dimensions each) from all retained tissue patches. To exclude clusters dominated by imaging artifacts, contaminants, or rare histological patterns [Chaudhary et al., 2024], we used the `scipy` Python module [Wolf et al., 2018] to construct a nearest-neighbor graph (30 principal components, 10 nearest neighbors), applied Leiden clustering (resolution 0.5), and retained only clusters represented by at least 50 patches in a minimum of 50 slides. This filtering resulted in 14 clusters comprising 4,205,039 patches, capturing the most prevalent histological phenotypes across the analyzed breast cancer slides. For population-level analyses across individuals, we defined slide-level image embeddings by averaging the leading principal components (PCs) of patch embeddings per slide. We performed principal component analysis on patch embeddings to reduce dimensionality—an essential step for

maintaining well-calibrated P-values in multivariate GWAS [Chaudhary et al., 2024, Kirchler et al., 2022]. Simulation studies demonstrated that using 64 principal components best balanced power and calibration, avoiding P-value deflation observed at higher dimensionalities (**Supplementary Figure A1**).

**Definition of text embeddings.** To capture high-dimensional semantic features from diagnostic pathology reports, we used preprocessed TCGA-BRCA diagnostic text, as described in Kefeli and Tatonetti [2024]. Each report was passed to the OpenAI `text-embedding-ada-002` model via the embedding API to generate dense vector representations, following the approach outlined in Chen and Zou [2025]. This yielded embeddings for all 1,054 tissue samples, each represented as a 1,536-dimensional latent vector. These embeddings encode unstructured clinical information from pathology reports—such as histological diagnoses, architectural descriptions, and pathological observations—into a format suitable for downstream multivariate GWAS and phenotypic association analysis. Proceeding analogously to the image embeddings, we applied principal component analysis (PCA) across all text embeddings and retained the top 64 components. These reduced-dimensional representations were used as input for all population-level analyses, including gene expression prediction and multivariate GWAS.

**Biological relevance of image and text embeddings.** To evaluate the biological relevance of image and text embeddings, we assessed their correlation with gene expression by testing, for each gene, whether expression levels could be predicted out-of-sample from the embeddings. Focusing on highly variable genes, identified using the `highly_variable_genes` function from `scanpy` [Wolf et al., 2018], we fit a variance component model for each gene with rank-based inverse normal transformed expression values as the outcome and 64-dimensional slide-level embeddings (image or text) as individual-level random effects. The model was trained on 50% of samples and evaluated on the held-out 50% by computing Spearman correlation between observed and predicted expression values. Genes with Bonferroni-corrected P-values  $< 0.05$  were considered significantly associated. This procedure was performed independently for image and text embeddings, enabling comparison of the biological signal captured by each modality. Genes were ranked by their predictability from image and text embeddings respectively<sup>1</sup>, and the top 100 were considered for pathway enrichment analysis using the GSEA module from Fang et al. [2023], with all analyzed highly variable genes used as the background set.

### 3.2 GWAS Analysis

**Genotype Quality Control and Imputation.** Genotype data for 996 samples were obtained from the Genome Data Commons archive. Initial Birdseed files were converted to VCF format, retaining genotype scores with confidence  $\geq 0.1$  and setting lower-quality scores to missing. Quality control (QC) was applied as follows: individuals were retained if they met genotype missingness ( $\text{mind} \geq 0.1$ ) and heterozygosity thresholds ( $|z_{\text{het}}| < 3$ ). Variants were filtered based on Hardy-Weinberg equilibrium ( $P_{\text{HWE}} > 10^{-6}$ ), missingness ( $\text{geno} \geq 0.01$ ), and minor allele frequency (MAF  $\geq 1\%$ ). Related individuals were excluded based on a relatedness cutoff of 0.125. Genotype imputation was conducted using the SHAPEIT/IMPUTE2 pipeline [Delaneau et al., 2012, Howie et al., 2009], using the 1000 Genomes Project Phase 3 [Consortium et al., 2015] as reference panel. Post-imputation, variants with an imputation quality score ( $\text{INFO} \geq 0.8$ ) and MAF  $\geq 5\%$  were retained, resulting in a final dataset of 753 individuals and 6,059,041 imputed variants after merging with histology data. To account for population structure, we computed the leading genetic principal components on the pre-imputation dataset, restricting this analysis to variants with MAF  $\geq 5\%$ .

**Multivariate GWAS using Linear Mixed Models.** To assess associations between individual genetic variants and image and text embeddings, we employed the linear mixed model framework introduced in [Chaudhary et al., 2024]. Briefly, given a genotype vector  $\mathbf{g}$  across  $N$  individuals, an  $N \times L$  matrix of histological embeddings  $\mathbf{X}$ , and an  $N \times K$  covariate matrix  $\mathbf{F}$ , the model is defined as:

$$\text{link}^{-1}(\mathbf{g}) = \mathbf{F}\boldsymbol{\alpha} + \mathbf{u}, \quad \text{where } \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathcal{K}(\mathbf{X})), \quad (1)$$

where  $\boldsymbol{\alpha}$  represents fixed effects from covariates, and  $\mathcal{K}(\mathbf{X})$  denotes the  $N \times N$  covariance matrix modeling pairwise histological similarity between individuals based on embeddings  $\mathbf{X}$ . We adopted a

---

<sup>1</sup>Using the Spearman correlation P-value as the ranking metric.

Gaussian likelihood, as previously shown to yield robust calibration and power for testing histological embeddings [Chaudhary et al., 2024]. For the covariance  $\mathcal{K}(\mathbf{X})$ , we used a cosine similarity kernel,  $\mathcal{K}_{\text{cosine}}(\mathbf{X})$ , which measures the similarity between histological embeddings by calculating the cosine of the angle between embedding vectors. This kernel effectively captures the alignment of histological features across samples, enhancing interpretability and relevance in genetic association tests [Schütze et al., 2008, Chaudhary et al., 2024]. For all tests, we accounted for sex, age and the leading 30 genetic principal components as covariates in  $\mathbf{F}$ . To evaluate association, we tested  $\sigma_X^2 > 0$  using a score test, similar to sequence kernel association tests for variant-set analyses [Wu et al., 2011]. For further details on the score test P values and on computational efficiency, we refer to [Chaudhary et al., 2024].

### 3.3 Downstream characterization of genome-wide significant loci

**Associated genes and pathways.** We characterized genome-wide significant loci by assessing their impact on gene expression, and subsequently performed pathway enrichment analysis to identify regulated pathways. Specifically, we used a linear model to test associations between the lead variant at each significant locus (predictor) and inverse rank normal transformed expression levels of highly variable genes (response), using matched breast cancer RNA-seq data. Associations were considered significant at an FDR-adjusted threshold of  $P < 0.05$ . For pathway analysis, we used Fisher’s exact test implemented in the gseapy Python module, with gene sets from the MSigDB\_Hallmark\_2020 collection [Liberzon et al., 2015]. The enrichment was evaluated separately for the top 50 positively and top 50 negatively associated genes for each lead variant.

**Associated clinical phenotypes from pathology reports.** To evaluate the clinical relevance of genome-wide significant loci, we tested their association with structured phenotypes extracted from diagnostic pathology reports. Binary labels for 11 predefined clinical traits—such as hormone receptor status, lobular or ductal histology, tumor necrosis, and lymphovascular invasion—were derived using LLM prompting with GPT-4.0 via the OpenAI API [OpenAI, 2023]. The model was instructed to return structured JSON responses indicating trait presence, absence, or ambiguity. If the model’s output was uncertain or ambiguous, we conservatively marked the corresponding trait as absent. For prompt details and the full list of traits, see **Supplementary Materials**. For each lead variant, we fit a linear model with binary trait status as the outcome and genotype dosage as the predictor, adjusting for age, sex, and the top 30 genetic principal components. Multiple testing across traits was controlled using Bonferroni correction.

**Associated histological traits through generative modeling.** To decode histological features from latent embeddings, we used a progressive conditional Generative Adversarial Network (cGAN) architecture, as described in Chaudhary et al. [2024]. Once trained, the generator maps 64-dimensional latent representations plus 512-dimensional Gaussian noise to realistic high-resolution ( $256 \times 256$ ) images via a convolutional decoder. To visualize histological changes associated with significant genetic variants, we combined this generator with latent-space interpolation. We first fit a linear mixed model with genotype as the outcome and slide-level embeddings as random effects, adjusting for sex, age, cancer subtype, and the top 30 genetic principal components. The *genetic effect axis* was defined using the leave-one-out BLUP estimator [Mefford et al., 2020]. Patch-level embeddings were then projected onto this axis to assign a genetic effect score. We selected the bottom and top 5% of patches in the score distribution and computed their means,  $\mathbf{x}_m$  and  $\mathbf{x}_M$ . Interpolation was performed via:

$$\mathbf{x}(\alpha) = (1 - \alpha)\mathbf{x}_m + \alpha\mathbf{x}_M, \quad (2)$$

with  $\alpha \in [0, 1]$ . Each  $\mathbf{x}(\alpha)$  was decoded into an image using the generator, producing smooth visualizations of genetic effects on tissue morphology. Full implementation details are provided in Chaudhary et al. [2024].

**Association with survival outcomes.** We evaluated the prognostic relevance of genes associated with significant loci using Cox proportional hazards regression [Cox, 1972]. We focused on genes significantly associated with lead loci at an FDR threshold of 5%, modeling gene expression (as a continuous variable) against disease-specific survival, adjusting for age at diagnosis and cancer grade. Hazard ratios (HRs) and p-values were estimated using the statsmodels library in Python [Seabold and Perktold, 2010]. For visualization, patients were stratified into high and low expression groups based on median expression, and Kaplan-Meier survival curves were plotted. The log-rank test

Table 1: Lead variants of significant loci from multivariate GWAS of image and text embeddings.

<b>RsID</b>	<b>P<sub>image</sub></b>	<b>P<sub>text</sub></b>	<b>Chr</b>	<b>Top Related Genes<sup>1</sup></b>
rs819976	$4.5 \times 10^{-40}$	$1.7 \times 10^{-9}$	1	LINC-PINT, KRT17, FAT2, DSG3, TRIM29
rs17078479	$8.1 \times 10^{-20}$	$5.9 \times 10^{-8}$	8	RN7SL3, RN7SL1, RN7SL4P, AC103591.3, DUSP1
rs861673	$1.5 \times 10^{-8}$	$2.9 \times 10^{-4}$	3	CAP2, RBM24, APOBEC3B, SMC4, AC104695.3
rs10149693	$1.4 \times 10^{-19}$	$3.7 \times 10^{-7}$	14	AC026462.1, LINC01554, RN7SL3, LINC-PINT, RN7SL4P
rs115215744	$4.9 \times 10^{-5}$	$3.3 \times 10^{-9}$	3	ALB, AC011352.3, AFP, ZPLD1, C5orf46
rs11850187	$8.8 \times 10^{-2}$	$9.0 \times 10^{-9}$	14	AP001324.1, SPTSSB, MTND1P23, SLC7A11, SLC5A8

<sup>1</sup> Top 5 genes most strongly associated with each SNP.

was used to assess statistical differences between strata [Kaplan and Meier, 1958]. Disease-specific survival was selected as the endpoint due to its established prognostic relevance in breast cancer, with follow-up available for up to 20 years from the time of diagnosis [Liu et al., 2018].

## 4 Results

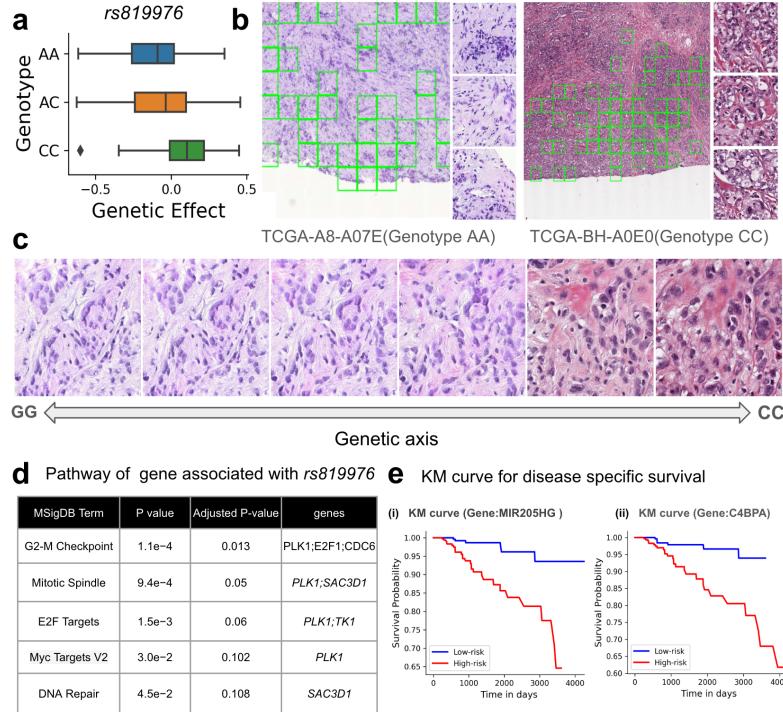


Figure 2: **Biological interpretation of the cross-modal locus rs819976.** (a) Genetic effect scores (from the image-based genetic axis) stratified by genotype at rs819976, showing a dose-dependent shift. (b) Representative H&E-stained whole slide images from individuals with AA and CC genotypes. Patches with extreme genetic effect scores—aligned with the direction of rs819976—are highlighted in green; selected patches illustrate allele-specific morphological differences. (c) Interpolated histology images generated via PGAN along the genetic effect axis of rs819976, moving from the GG to CC genotype. Morphological transitions suggest pleomorphism and necrosis associated with the risk allele. (d) Enriched pathways among the top 50 genes associated with rs819976; genes are annotated from the MSigDB Hallmark 2020 collection. (e) Kaplan–Meier survival curves for two genes most strongly associated with rs819976—MIR205HG and C4BPA—highlighting associations with disease-specific survival.

#### 4.1 Biological Relevance of Image and Text Embeddings

We processed breast tissue slides from 1,054 individuals in the TCGA breast cancer cohort [Ciriello et al., 2015; 13 et al., 2012], defining image and text embeddings as described in **Methods**. To assess the biological relevance of these embeddings, we evaluated their ability to predict gene expression levels across individuals. Among 5,000 highly variable genes, 1,469 were significantly associated with image embeddings and 1,054 with text embeddings (Bonferroni-adjusted  $P < 0.05$ ). Notably, 824 genes were shared across both sets, representing a highly significant overlap (hypergeometric test,  $P < 10^{-300}$ , see **Methods**). Pathway enrichment analysis revealed converging signals across both modalities. Top-ranked genes associated with each embedding type were enriched for hallmark cancer pathways, including TNF- $\alpha$  signaling via NF- $\kappa$ B, G2/M checkpoint, and adipogenesis—all linked to breast cancer proliferation, immune signaling, and clinical prognosis [Cai et al., 2017; Oshi et al., 2020] (**Supplementary Figure A2**). Overall, these results demonstrate that both image and text embeddings capture molecular variation reflective of transcriptional programs in breast cancer.

#### 4.2 Multivariate GWAS and Biological Characterization of Significant Loci

We performed multivariate GWAS on both image and text embeddings derived from H&E-stained slides and diagnostic pathology reports, respectively. This analysis identified six independent genome-wide significant loci in breast cancer, with two loci colocalizing across both modalities (**Figure 1b**).

The strongest cross-modal signal was observed at *rs819976* ( $P_{\text{image}} < 4.5 \times 10^{-40}$ ;  $P_{\text{text}} = 1.7 \times 10^{-9}$ ), a variant located in the gene body of *ATAD3B*, a mitophagy receptor previously implicated in poor breast cancer prognosis [Shu et al., 2021; Ovaska et al., 2013]. This SNP was associated with expression changes in 190 genes, including *LINC-PINT*, *KRT17*, and *FAT2*, all known markers of breast cancer aggressiveness [Li and Hu, 2024; Tang et al., 2022; Roache et al., 2022]. Pathway enrichment analysis highlighted the p53 signaling cascade, KRAS signaling, and the G2–M checkpoint pathway (**Figure 2**)—hallmarks of tumor proliferation and metastasis. *rs819976* also appears in a SuSiE fine-mapped GTEx eQTL credible set for gene *ATAD3B* in breast epithelium, supporting its regulatory relevance. To assess its clinical significance, we tested *rs819976* for association with 11 structured diagnostic phenotypes extracted from pathology reports via GPT-4.0-based prompting. The variant showed significant associations with Invasive Lobular Carcinoma (ILC), the presence of Lobular Carcinoma In Situ (LCIS), and Estrogen Receptor (ER) positivity, reinforcing its connection to well-established diagnostic features of aggressive breast cancer. To visualize morphological effects, we employed a progressive GAN model conditioned on genotype. Interpolation along the inferred genetic effect axis revealed that the risk allele (C) was associated with pleomorphic and necrotic histological features (**Figure 2**)—traits consistent with aggressive tumor biology as observed in real tissue slides.

The second cross-modal locus, *rs17078479* ( $P_{\text{image}} < 8.1 \times 10^{-20}$ ;  $P_{\text{text}} = 5.9 \times 10^{-8}$ ), is located in a noncoding region on chromosome 8. It was associated with the expression of 13 genes, including *DUSP1*, *FOS*, *LINC-PINT*, and *EGR1*, all implicated in breast cancer progression [Boulding et al., 2016; Bland et al., 1995; Li and Hu, 2024; Saha et al., 2021]. Enriched pathways included TNF- $\alpha$ /NF- $\kappa$ B signaling and cholesterol homeostasis, both of which play central roles in inflammation and metabolic regulation in cancer [Wu and Zhou, 2010; Liu et al., 2021]. No structured clinical traits were significantly associated with this variant after Bonferroni correction. PGAN-based interpolation did not produce clearly interpretable morphological differences for this variant (**Supplementary Figure A3**).

Beyond these two loci with cross-modal support, we identified four genome-wide significant loci supported by a single modality. *rs10149693* ( $P_{\text{image}} < 1.4 \times 10^{-19}$ ;  $P_{\text{text}} = 3.7 \times 10^{-7}$ ), an intronic variant in *OTUB2*, was significantly associated with image-derived embeddings. *OTUB2* is known to promote tumor stemness and metastasis [Zhang et al., 2019]. This variant was associated with expression of four genes, including *LINC01554*, a long non-coding RNA linked to the regulation of *miR-1267*, a marker of tumor progression [Wang et al., 2022b; Torkashvand et al., 2016]. Pathway analysis implicated hypoxia response and TNF- $\alpha$ /NF- $\kappa$ B signaling, both relevant to inflammation and tumor aggressiveness [Xu et al., 2010; Liao and Dickson, 2000; Zhi et al., 2024; Wu and Zhou, 2010]. *rs861673* ( $P_{\text{image}} < 1.5 \times 10^{-8}$ ;  $P_{\text{text}} = 2.9 \times 10^{-4}$ ), located on chromosome 3, also emerged from image-based analysis. Although no single gene was significantly associated, pathway-level analysis highlighted KRAS signaling, a well-established driver of oncogenesis [Kim et al., 2015]. Two additional loci—*rs115215744* ( $P_{\text{text}} = 3.3 \times 10^{-9}$ ;  $P_{\text{image}} = 4.9 \times 10^{-5}$ ) and *rs11850187*

( $P_{\text{text}} = 9.0 \times 10^{-9}$ ;  $P_{\text{image}} = 8.8 \times 10^{-2}$ )—were exclusively identified through the text-based GWAS. *rs115215744* was associated with the expression of *AFP* and *ZPLD1*, while *rs11850187* was linked to *SPTSSB* and *MTND1P23*, among others (see **Table 1**). Although these loci may capture modality-specific signals, their lack of cross-modal replication warrants caution, and larger cohorts will be needed to robustly assess their biological and clinical relevance.

#### 4.3 Association with survival outcomes

We first performed survival analysis on the six genome-wide significant variants identified in our study, adjusting for age at diagnosis and cancer grade; however, no significant associations were observed (**Methods**). We then analyzed the 207 genes whose expression was significantly associated with these variants (FDR  $\leq 5\%$ ) and identified two genes, *MIR205HG* ( $P < 8.8 \times 10^{-6}$ ; hazard ratio (HR) = 0.547) and *C4BPA* ( $P < 3.6 \times 10^{-4}$ ; HR = 0.549), as significantly associated with disease-specific survival (FDR  $\leq 5\%$ ). Both *MIR205HG* and *C4BPA* have been previously implicated as prognostic markers in breast cancer [Xu et al., 2022, Zou et al., 2024]. Kaplan-Meier survival curves stratified patients into high-risk and low-risk groups based on the median risk score derived from the Cox model, showing a clear separation in survival probabilities between the two groups (**Figure 2e**).

### 5 Discussion

This study introduces a multimodal GWAS framework that integrates image-based and text-based phenotypes derived from diagnostic pathology to identify germline variants influencing breast cancer histology. By jointly analyzing tissue morphology and clinical language, we demonstrate that complementary representations of the tumor microenvironment can provide converging evidence for germline associations, increasing interpretability and biological confidence beyond what either modality offers alone.

Across modalities, we identified six genome-wide significant loci, with two exhibiting strong association in both modalities—highlighted as the most robust and interpretable findings. For example, *rs819976* (in *ATAD3B*) was associated with pleomorphic, necrotic morphology and poor prognosis, supported by both embedding modalities, gene expression, pathway enrichment, and disease-specific survival. *rs17078479* similarly demonstrated concordant evidence across modalities, including links to inflammatory and metabolic pathways. These converging signals reinforce the utility of multimodal phenotyping for germline discoveries. Beyond discovery, such loci could inform patient stratification or risk modeling, especially when integrated with somatic alterations and clinical features.

To further contextualize genetic signals, we used large language models to extract structured traits from diagnostic reports. This enabled phenotype association analysis (akin to PheWAS), linking variants such as *rs819976* to diagnostic features including ER status and histological subtype. While these extracted traits were not independently validated, the approach illustrates how LLMs can scale phenotype curation—a direction that warrants further benchmarking. We also explored model-based visual interpretation using progressive GANs. For selected loci, we generated interpolations illustrating allele-specific effects on morphology. While these outputs offer intuitive visual summaries that complement quantitative associations, they are not definitive representations of genetic effects and should be interpreted cautiously alongside real histology.

To fully realize the potential of germline discovery in cancer, future work must address key data and modeling challenges. First, there is a need for independent validation across diverse cohorts with matched histology, clinical text, and germline genotypes. Triaging such datasets will be critical for assessing replicability and generalizability across clinical settings and demographic backgrounds. In our case, analyses were limited to the TCGA-BRCA cohort, underscoring the importance of cohort expansion for future efforts. Second, larger sample sizes will be critical for unlocking the full discovery potential of these methods, as demonstrated using simulations in prior work [Chaudhary et al., 2024]. Third, the presence of strong technical and batch effects poses a challenge for both discovery and interpretability. Future approaches that incorporate joint multimodal representation learning may better isolate biological signals, mitigating these challenges.

## Competing interests

The authors declare that they have no competing interests.

## Acknowledgments

This study used genotype and histology data from The Cancer Genome Atlas (TCGA), accessed via dbGaP ([phs000178.v11.p8](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC37880/)) under project #37880. SC and FPC were funded by the Free State of Bavaria's High-Tech Agenda through the Institute of AI for Health (AIH). SC also acknowledges support from HIDSS-006 – the Helmholtz Information and Data Science School for Health at Helmholtz Munich, TUM, and LMU.

## Contributions

SC implemented the methods and performed the main analyses. AV, SV, and MH provided critical input on data preprocessing, study design, methodology, and interpretation of results. SC and FPC conceived the project and wrote the manuscript with input from all authors. FPC supervised the study.

## Use of Artificial Intelligence

In the preparation of this manuscript, we utilized the large language models GPT-4 and GPT-5 (<https://chat.openai.com/>) for editing assistance, including language polishing and clarification of text. While this tool assisted in refining the manuscript's language, it was not used to generate contributions to the original research, data analysis, or interpretation of results. All final content decisions and responsibilities rest with the authors.

## References

- Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.
- Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023.
- Kirby I Bland, Manos M Konstadoulakis, Michael P Vezeridis, and Harold J Wanebo. Oncogene protein co-expression value of ha-ras, c-myc, c-fos, and p53 as prognostic discriminants for breast carcinoma. *Annals of surgery*, 221(6):706–720, 1995.
- Tara Boulding, Fan Wu, Robert McCuaig, Jennifer Dunn, Christopher R Sutton, Kristine Hardy, Wenjuan Tu, Amanda Bullman, Desmond Yip, Jane E Dahlstrom, et al. Differential roles for dusp family members in epithelial-to-mesenchymal transition and cancer stem cell regulation in breast cancer. *PloS one*, 11(2):e0148065, 2016.
- G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- Xiaoli Cai, Can Cao, Jiong Li, Fuquan Chen, Shuqin Zhang, Bowen Liu, Weiyang Zhang, Xiaodong Zhang, and Lihong Ye. Inflammatory factor tnf- $\alpha$  promotes the growth of breast cancer via the positive feedback loop of tnfr1/nf- $\kappa$ b (and/or p38)/p-stat3/hbxip/tnfr1. *Oncotarget*, 8(35):58338, 2017.
- Rui Cao, Fan Yang, Si-Cong Ma, Li Liu, Yu Zhao, Yan Li, De-Hua Wu, Tongxin Wang, Wei-Jia Lu, Wei-Jing Cai, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics*, 10(24):11080, 2020.

- Francesco Paolo Casale, Barbara Rakitsch, Christoph Lippert, and Oliver Stegle. Efficient set tests for the genetic analysis of correlated traits. *Nature methods*, 12(8):755–758, 2015.
- Ajay Chatrath, Roza Przanowska, Shashi Kiran, Zhangli Su, Shekhar Saha, Briana Wilson, Takaaki Tsunematsu, Ji-Hye Ahn, Kyung Yong Lee, Teressa Paulsen, et al. The pan-cancer landscape of prognostic germline variants in 10,582 patients. *Genome medicine*, 12:1–18, 2020.
- Shubham Chaudhary, Almut Voigts, Michael Bereket, Matthew L Albert, Kristina Schwamborn, Elefheria Zeggini, and Francesco Paolo Casale. Histogwas: An ai-enabled framework for automated genetic analysis of tissue phenotypes in histology cohorts. *bioRxiv*, pages 2024–06, 2024.
- Richard J Chen and Rahul G Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*, 2022.
- Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.
- Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature biomedical engineering*, 9(4):483–493, 2025.
- Giovanni Ciriello, Michael L Gatz, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.
- 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.
- David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.
- Maria Escala-Garcia, Qi Guo, Thilo Dörk, Sander Canisius, Renske Keeman, Joe Dennis, Jonathan Beesley, Julie Lecarpentier, Manjeet K Bolla, Qin Wang, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. *British journal of cancer*, 120(6):647–657, 2019.
- Maria Escala-Garcia, Sander Canisius, Renske Keeman, Jonathan Beesley, Hoda Anton-Culver, Volker Arndt, Annelie Augustinsson, Heiko Becher, Matthias W Beckmann, Sabine Behrens, et al. Germline variants and breast cancer survival in patients with distant metastases at primary breast cancer diagnosis. *Scientific reports*, 11(1):19787, 2021.
- Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.
- Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer*, 1(8):800–810, 2020.
- Nicholas A Furlotte and Eleazar Eskin. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics*, 200(1):59–68, 2015.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

- Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.
- Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023.
- Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.
- Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.
- Jenna Kefeli and Nicholas Tatonetti. Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. *Patterns*, 5(3), 2024.
- Rae-Kwon Kim, Yongjoon Suh, Ki-Chun Yoo, Yan-Hong Cui, Hyeonmi Kim, Min-Jung Kim, In Gyu Kim, and Su-Jae Lee. Activation of kras promotes the mesenchymal features of basal-type breast cancer. *Experimental & molecular medicine*, 47(1):e137–e137, 2015.
- Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph Lippert. transfrgwas: Gwas of images using deep transfer learning. *Bioinformatics*, 38(14):3621–3628, 2022.
- Alexis Lamiable, Tiphaine Champetier, Francesco Leonardi, Ethan Cohen, Peter Sommer, David Hardy, Nicolas Argy, Achille Massougbedji, Elaine Del Nery, Gilles Cottrell, et al. Revealing invisible cell phenotypes with conditional generative modeling. *Nature Communications*, 14(1):6386, 2023.
- Tristan Lazard, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. Giga-ssl: Self-supervised learning for gigapixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023.
- Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.
- Daohong Li and Aixia Hu. Linc-pint suppresses breast cancer cell proliferation and migration via meis2/ppp3cc/nf- $\kappa$ b pathway by sponging mir-576-5p. *The American Journal of the Medical Sciences*, 367(3):201–211, 2024.
- DJ Liao and RB Dickson. c-myc in breast cancer. *Endocrine-related cancer*, 7(3):143–164, 2000.
- Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.
- Neal I Lindeman, Philip T Cagle, Mary Beth Beasley, Dhananjay Arun Chitale, Sanja Dacic, Giuseppe Giaccone, Robert Brian Jenkins, David J Kwiatkowski, Juan-Sebastian Saldivar, Jeremy Squire, et al. Molecular testing guideline for selection of lung cancer patients for egfr and alk tyrosine kinase inhibitors: guideline from the college of american pathologists, international association for the study of lung cancer, and association for molecular pathology. *Journal of Thoracic Oncology*, 8(7):823–859, 2013.
- Christoph Lippert, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle. Limix: genetic analysis of multiple traits. *BioRxiv*, page 003905, 2014.
- Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

Wen Liu, Binita Chakraborty, Rachid Safi, Dmitri Kazmin, Ching-yi Chang, and Donald P McDonnell. Dysregulated cholesterol homeostasis results in resistance to ferroptosis increasing tumorigenicity and metastasis in cancer. *Nature communications*, 12(1):5103, 2021.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Joel Mefford, Danny Park, Zhili Zheng, Arthur Ko, Mika Ala-Korpela, Markku Laakso, Päivi Pajukanta, Jian Yang, John Witte, and Noah Zaitlen. Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models. *Journal of Computational Biology*, 27(4):599–612, 2020.

Michael P Menden, Francesco Paolo Casale, Johannes Stephan, Graham R Bignell, Francesco Iorio, Ultan McDermott, Mathew J Garnett, Julio Saez-Rodriguez, and Oliver Stegle. The germline genetic component of drug sensitivity in cancer cell lines. *Nature communications*, 9(1):3385, 2018.

Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Ricardo Mokhtari, Azam Hamidinekoo, Daniel Sutton, Arthur Lewis, Bastian Angermann, Ulf Gehrmann, Pal Lundin, Hibret Adissu, Junmei Cairns, Jessica Neisen, et al. Interpretable histopathology-based prediction of disease relevant features in inflammatory bowel disease biopsies using weakly-supervised deep learning. *arXiv preprint arXiv:2303.12095*, 2023.

Anna Morra, Maria Escala-Garcia, Jonathan Beesley, Renske Keeman, Sander Canisius, Thomas U Ahearn, Irene L Andrulis, Hoda Anton-Culver, Volker Arndt, Paul L Auer, et al. Association of germline genetic variants with breast cancer-specific survival in patient subgroups defined by clinic-pathological variables related to tumor biology and type of systemic treatment. *Breast cancer research*, 23(1):86, 2021.

OpenAI. Gpt-4 technical report. <https://openai.com/research/gpt-4>, 2023. Accessed: [insert date here].

Masanori Oshi, Hideo Takahashi, Yoshihisa Tokumaru, Li Yan, Omar M Rashid, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. G2m cell cycle pathway score as a prognostic biomarker of metastasis in estrogen receptor (er)-positive breast cancer. *International journal of molecular sciences*, 21(8):2921, 2020.

Kristian Ovaska, Filomena Matarese, Korbinian Grote, Iryna Charapitsa, Alejandra Cervera, Chengyu Liu, George Reid, Martin Seifert, Hendrik G Stunnenberg, and Sampsa Hautaniemi. Integrative analysis of deep sequencing data identifies estrogen receptor early response genes and links atad3b to poor survival in breast cancer. *PLoS computational biology*, 9(6):e1003100, 2013.

Alessandro Palma, Fabian J Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *bioRxiv*, pages 2023–07, 2023.

Thomas Roache, Megan Sumera, Zalaila Laird, and Amrita Datta. The role of protocadherin, fat2 in breast cancer. *Cancer Research*, 82(12\_Supplement):854–854, 2022.

Hege G Russnes, Ole Christian Lingjærde, Anne-Lise Børresen-Dale, and Carlos Caldas. Breast cancer molecular stratification: from intrinsic subtypes to integrative clusters. *The American journal of pathology*, 187(10):2152–2162, 2017.

Subbroto Kumar Saha, SM Riazul Islam, Tripti Saha, Afsana Nishat, Polash Kumar Biswas, Minchan Gil, Lewis Nkenyereye, Shaker El-Sappagh, Md Saiful Islam, and Ssang-Goo Cho. Prognostic role of egr1 in breast cancer: A systematic review. *BMB reports*, 54(10):497, 2021.

Oliver Lester Saldanha, Chiara ML Loeffler, Jan Moritz Niehues, Marko van Treeck, Tobias P Seraphin, Katherine Jane Hewitt, Didem Cifci, Gregory Patrick Veldhuizen, Siddhi Ramesh, Alexander T Pearson, et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ Precision Oncology*, 7(1):35, 2023.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Li Shu, Chao Hu, Meng Xu, Jianglong Yu, He He, Jie Lin, Hongying Sha, Bin Lu, Simone Engelender, Minxin Guan, et al. Atad3b is a mitophagy receptor mediating clearance of oxidative stress-induced damaged mitochondrial dna. *The EMBO Journal*, 40(8):e106283, 2021.

Shasha Tang, Wenjing Liu, Liyun Yong, Dongyang Liu, Xiaoyan Lin, Yuan Huang, Hui Wang, and Fengfeng Cai. Reduced expression of krt17 predicts poor prognosis in her2high breast cancer. *Biomolecules*, 12(9):1183, 2022.

S Torkashvand, Z Damavandi, B Mirzaei, M Tavallaei, M Vasei, and Seyed Javad Mowla. Decreased expression of bioinformatically predicted piwil2-targetting micrornas, mir-1267 and mir-2276 in breast cancer. *Archives of Iranian medicine*, 19(6):420–425, 2016.

Patrick Turley, Raymond K Walters, Omeed Maghzian, Aysu Okbay, James J Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A Furlotte, et al. Multi-trait analysis of genome-wide association summary statistics using mttag. *Nature genetics*, 50(2):229–237, 2018.

Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, Piet A van den Brandt, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41(9):1650–1661, 2023.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022a.

Zhenchuan Wang, Qiuying Sha, and Shuanglin Zhang. Joint analysis of multiple traits using " optimal" maximum heritability test. *PloS one*, 11(3):e0150975, 2016.

Zizong Wang, Bin Yang, Jin Zhang, and Xiangyang Chu. Long noncoding rna linc01554 inhibits the progression of nsclc progression by functioning as a cerna for mir-1267 and regulating ing3/akt/mTOR pathway. *BioMed Research International*, 2022(1):7162623, 2022b.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Scott E Woodman, Alexander J Lazar, Kenneth D Aldape, and Michael A Davies. New strategies in melanoma: molecular testing in advanced disease. *Clinical Cancer Research*, 18(5):1195–1200, 2012.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

Ya-di Wu and BP Zhou. Tnf- $\alpha$ /nf- $\kappa$ b/snail pathway in cancer cell migration and invasion. *British journal of cancer*, 102(4):639–644, 2010.

Ziqian Xie, Tao Zhang, Sangbae Kim, Jiaxiong Lu, Wanheng Zhang, Cheng-Hui Lin, Man-Ru Wu, Alexander Davis, Roomasa Channa, Luca Giancardo, et al. igwas: Image-based genome-wide association of self-supervised deep phenotyping of retina fundus images. *PLoS genetics*, 20(5): e1011273, 2024.

Jinhua Xu, Yinghua Chen, and Olufunmilayo I Olopade. Myc and breast cancer. *Genes & cancer*, 1 (6):629–640, 2010.

Yaqian Xu, Chenwei Yuan, Jing Peng, Liheng Zhou, Yanping Lin, Yaohui Wang, Jie Zhang, Jiayi Ma, Wenjin Yin, and Jinsong Lu. Lncrna mir205hg expression predicts efficacy of neoadjuvant chemotherapy for patients with locally advanced breast cancer. *Genes & Diseases*, 9(4):837, 2022.

Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, et al. Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. *Nature Genetics*, 56(8):1604–1613, 2024.

Zhengkui Zhang, Jinjin Du, Shuai Wang, Li Shao, Ke Jin, Fang Li, Bajin Wei, Wei Ding, Peifen Fu, Hans van Dam, et al. Otub2 promotes cancer metastasis via hippo-independent activation of yap and taz. *Molecular cell*, 73(1):7–21, 2019.

Shijiao Zhi, Chen Chen, Hanlin Huang, Zhengfu Zhang, Fancai Zeng, and Shujun Zhang. Hypoxia-inducible factor in breast cancer: role and target for breast cancer treatment. *Frontiers in Immunology*, 15:1370800, 2024.

Qianqian Zhu, Emily Schultz, Jirong Long, Janise M Roh, Emily Valice, Cecile A Laurent, Kelly H Radimer, Li Yan, Isaac J Ergas, Warren Davis, et al. Uaca locus is associated with breast cancer chemoresistance and survival. *NPJ breast cancer*, 8(1):39, 2022.

Juan Zou, Yaokun Chen, Zeqi Ji, Danyi Liu, Xin Chen, Mengjia Chen, Kexun Chen, Haojia Lin, Yexi Chen, and Zhiyang Li. Identification of c4bpa as biomarker associated with immune infiltration and prognosis in breast cancer. *Translational Cancer Research*, 13(1):25, 2024.