Supplementary Materials for “AI-based histopathology phenotyping reveals germline loci shaping breast cancer morphology”

Contents

A1 Extraction of structured clinical phenotypes

We derived binary clinical phenotypes for breast cancer using both structured fields and unstructured diagnostic reports from the TCGA-BRCA dataset. Labels for Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC) were obtained directly from structured clinical annotations. The remaining phenotypes were extracted from free-text pathology reports using large language model (LLM) prompting.

Each report was processed using a custom prompt to GPT-4.0 (OpenAI), which was instructed to return a structured JSON object indicating the presence or absence of predefined histopathological features. The model was constrained to binary outputs according to the following prompt:

```
You are an expert medical language model trained in oncologic pathology.
Your role is to extract structured binary breast cancer phenotypes
from unstructured pathology reports, such as those from the TCGA-BRCA cohort.

From the report provided, return a JSON object mapping predefined phenotypes
to binary values, using the following coding:

- 1 if the feature is present, confirmed, or stated as positive.
- 0 if the feature is explicitly absent or negative.
- "NA" if there is no clear information, ambiguous wording, or pending results.
```

```
Use your expertise to resolve synonyms and nuanced terminology where possible.
Do not infer unstated information -- when uncertain, use "NA".

Return only the JSON dictionary in your response.

Phenotype keys: "DCIS_present", "LCIS_present", "ER_positive", "PR_positive",
 "Lymph_node_metastasis", "Lymphovascular_invasion_present",
 "Tumor_necrosis_present", "Multifocal_tumor", "Apocrine_features_present"

Here is the pathology report: {Pathology report}
```

This pipeline was applied to all 753 patients to generate structured clinical phenotypes for downstream variant-trait association analysis.

A2 Supplementary Figures

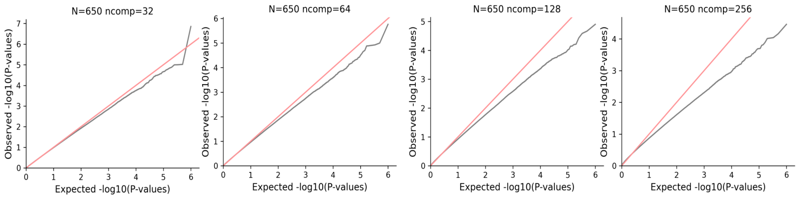
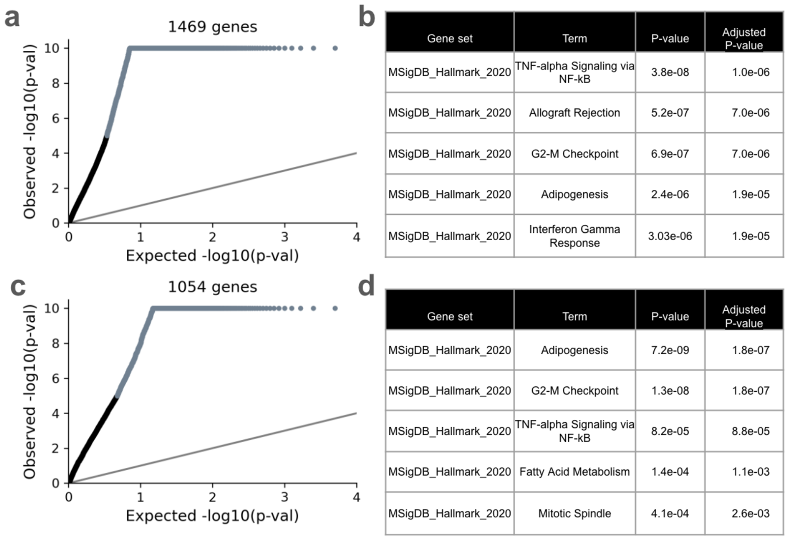


Figure A1: **Calibration of GWAS P-values under null simulation across varying embedding dimensionality**. QQ plots show P-value distributions from HistoGWAS applied to synthetic datasets with no genetic effects, using cohort size $N = 650$ and varying the number of principal components (ncomps) retained. As the number of dimensions increases, P-values become increasingly deflated. These results informed our decision to retain 64 principal components in this study. **Note: Figure adapted from Chaudhary et al. [1]**



Gene set	Term	P-value	Adjusted P-value
MSigDB_Hallmark_2020	TNF-alpha Signaling via NF-kB	3.8e-08	1.0e-06
MSigDB_Hallmark_2020	Allograft Rejection	5.2e-07	7.0e-06
MSigDB_Hallmark_2020	G2-M Checkpoint	6.9e-07	7.0e-06
MSigDB_Hallmark_2020	Adipogenesis	2.4e-06	1.9e-05
MSigDB_Hallmark_2020	Interferon Gamma Response	3.03e-06	1.9e-05

Gene set	Term	P-value	Adjusted P-value
MSigDB_Hallmark_2020	Adipogenesis	7.2e-09	1.8e-07
MSigDB_Hallmark_2020	G2-M Checkpoint	1.3e-08	1.8e-07
MSigDB_Hallmark_2020	TNF-alpha Signaling via NF-kB	8.2e-05	8.8e-05
MSigDB_Hallmark_2020	Fatty Acid Metabolism	1.4e-04	1.1e-03
MSigDB_Hallmark_2020	Mitotic Spindle	4.1e-04	2.6e-03

Figure A2: **Gene expression prediction and pathway enrichment analysis for image and text embeddings.** (**a**) Quantile–quantile (QQ) plot of observed versus expected $-\log_{10} P$-values from gene-wise out-of-sample prediction tests using image embeddings, with 1,469 genes exceeding the Bonferroni significance threshold. (**b**) Top five pathways enriched among significantly predicted genes from image embeddings, based on MSigDB Hallmark 2020 annotations. (**c**) Analogous QQ plot for gene expression predictions using text embeddings, with 1,054 genes significantly associated. (**d**) Analogous pathway enrichment results for text embeddings, highlighting strong overlap with image-derived signals.

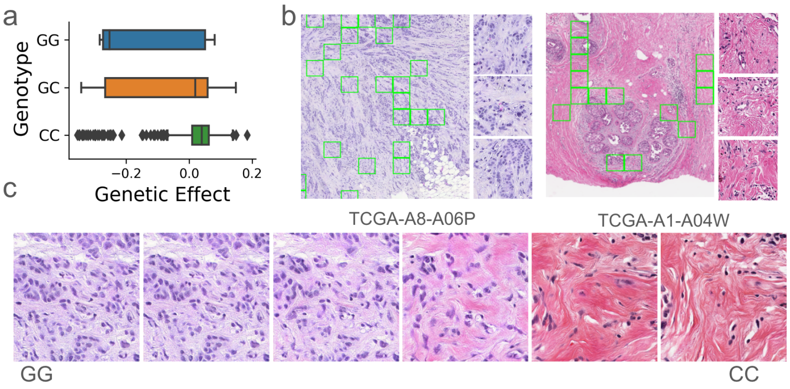


Figure A3: **Impact of the variant *rs*17078479** (**a**) The box plots displays the variation of genetic axis scores for $rs17078479$.(**b**) Whole slide images showcasing phenotypes associated with different allele, the patches closely linked to specific alleles (defined by the genetic axis of $rs17078479$) are highlighted in green. For each slide three patched are highlights to provide the detailed view of the histological differences.(**c**) Histology image demonstrating the allele effect by projecting interpolated embeddings along the direction of $rs17078479$ genetic axis score

References

[1] Shubham Chaudhary, Almut Voigts, Michael Bereket, Matthew L Albert, Kristina Schwamborn, Eleftheria 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.