

1 Data-Driven Modelling of Gene Expression States in Breast Cancer and their 2 Prediction from Routine Whole Slide Images

3 Muhammad Dawood¹, Mark Eastwood¹, Mostafa Jahanifar¹, Lawrence Young^{2,3}, Asa Ben-Hur⁴, Kim Branson⁵,
4 Louise Jones⁶, Nasir Rajpoot^{1,7}, Fayyaz ul Amir Afsar Minhas^{1,3}

5
6 ¹Tissue Image Analytics Centre, University of Warwick, Coventry, United Kingdom

7 ²Warwick Medical School, University of Warwick, Coventry, United Kingdom

8 ³Cancer Research Centre, University of Warwick, Coventry, United Kingdom

9 ⁴Department of Computer Science, Colorado State University, Fort Collins CO, United States

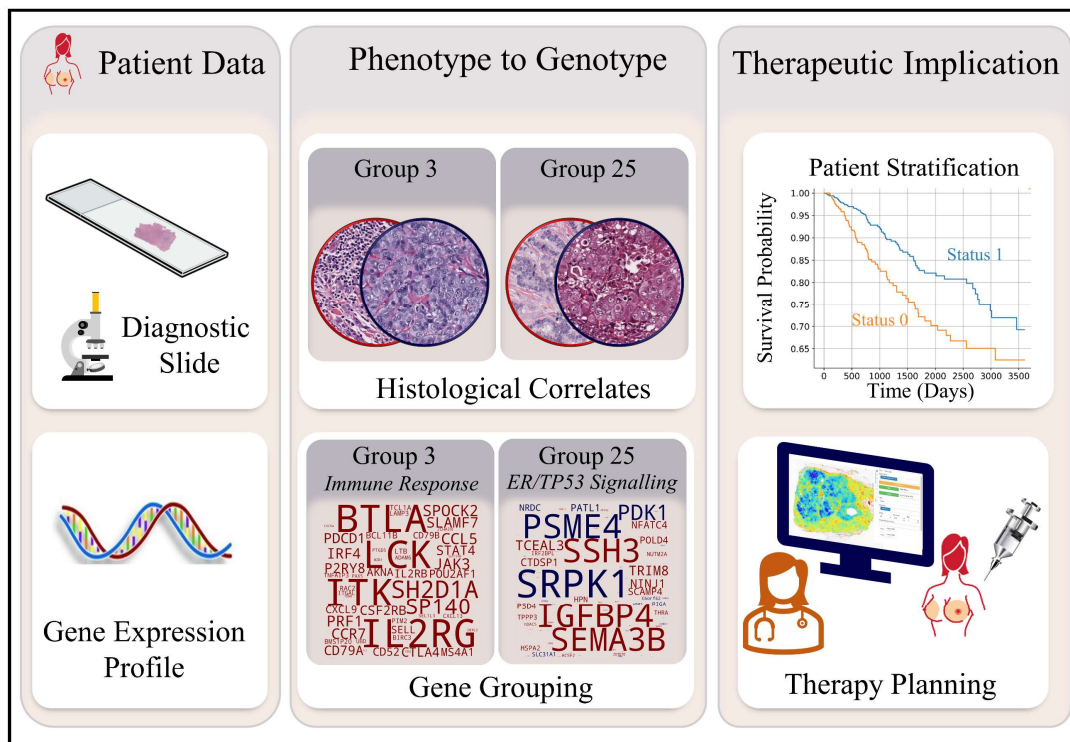
10 ⁵Artificial Intelligence & Machine Learning, GlaxoSmithKline

11 ⁶Barts Cancer Institute, Queen Mary University of London

12 ⁷Department of Pathology, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, United Kingdom

13

14 Graphical Abstract



15

16 Highlights

- 17 • Data-driven discovery of co-expressing gene groups in breast cancer
- 18 • Histological imaging based prediction of gene groups via deep learning
- 19 • Identification of phenotypic correlates of gene-expression in histological imaging
- 20 • Clinical and therapeutic impact of gene groups and their visual patterns identified

21 Summary

22 Identification of gene expression state of a cancer patient from routine pathology imaging and
23 characterization of its phenotypic effects have significant clinical and therapeutic implications.
24 However, prediction of expression of individual genes from whole slide images (WSIs) is
25 challenging due to co-dependent or correlated expression of multiple genes. Here, we use a purely
26 data-driven approach to first identify groups of genes with co-dependent expression and then
27 predict their status from (WSIs) using a bespoke graph neural network. These gene groups allow
28 us to capture the gene expression state of a patient with a small number of binary variables that are
29 biologically meaningful and carry histopathological insights for clinically and therapeutic use
30 cases. Prediction of gene expression state based on these gene groups allows associating
31 histological phenotypes (cellular composition, mitotic counts, grading, etc.) with underlying gene
32 expression patterns and opens avenues for gaining significant biological insights from routine
33 pathology imaging directly.

34 1 Introduction

35 Cancer is a clonal disease in which genetic alterations directly or indirectly alter gene expression,
36 biological pathways, and proteins activity leading to phenotypic changes in the spatial organization
37 of the tumor microenvironment (TME) [1]. Consequently, associating histological and molecular
38 patterns is crucial for understanding disease mechanism and clinical decision-making [2]. Like
39 other cancers, breast tumors also exhibit heterogeneity at both morphological and molecular levels
40 and are divided into several histological and molecular subtypes. During histopathology
41 examination, a tumor section stained with Hematoxylin and Eosin (H&E) is visually examined for
42 features such as mitotic counts, nuclear pleomorphism, epithelial tubule formation, necrosis and
43 tumor-infiltrating lymphocytes, etc., to develop a spatially-informed histological profile of the
44 disease. Similarly, gene expression analysis based on molecular tests such as PAM50 [3], [4],
45 Oncotype-Dx [5] and Mammaprint [6] can also be used for patient subtyping. Gene expression
46 profiling based on such limited gene assays or from Bulk RNA-Seq [7] and single-cell RNA-
47 sequencing (scRNA-seq) [8], [9] plays a key role in understanding the genetic basis of cancer and
48 discovery of novel therapeutic targets. However, such technologies are unable to capture spatial
49 heterogeneity in the expression profile of genes across a tumor section. Spatial profiling of a tumor
50 transcriptome is typically achieved using Spatially resolved Transcriptomics (SpTx) technologies

51 [10]. However, such technologies are generally costly and offer low resolution in terms of spatial
52 details or genes [11], [12]. Consequently, there is a need for cross-linking gene expression and
53 spatial histological imaging profiles to gain a more in-depth understanding of latent factors
54 associated with the disease.

55 In an attempt to achieve this goal, recent advancements in deep learning for computational
56 pathology have demonstrated that prediction of expression profiles of genes is possible from whole
57 slide images (WSIs) of H&E stained tissue sections [13]–[15]. For example, Schmauch et al.
58 proposed a deep learning method called HE2RNA for predicting gene expression profiles from
59 WSIs. Similarly, Wang et al. proposed a deep learning method for predicting the expression profile
60 of 17,695 genes from WSIs [16]. For each of the 17,695 genes, the authors have tiled the WSIs
61 into patches and then trained and optimized an Inception V3 for predicting tile-level and WSI-
62 level expression. Most recently, an attention-based called tRNAsformer has been proposed for
63 predicting the expression level of the individual gene from WSIs in kidney cancer [17].

64 The vast majority of image-based RNA-Seq expression prediction methods focus on associating
65 tissue morphology with the expression level of *individual* genes [15]–[17]. This is typically done
66 by designing a machine learning pipeline in which the input is a WSI, and the output is the
67 expression level of a single gene. However, due to the nature of the biological mechanisms
68 underlying gene expression, genes usually show co-dependent or correlated expression.
69 Consequently, it is, in general, not possible to associate the predicted expression of a single gene
70 from the input WSI to that gene alone. Furthermore, an observed phenotypic effect cannot solely
71 be pinpointed to the known function of a single gene as, typically, it will be a collective effect
72 exhibited by the expression of functionally interrelated genes and a single gene may be associated
73 with multiple functions [18]. Therefore, instead of predicting the phenotypic effect of a single gene
74 from WSIs, it is more meaningful to predict the expression of groups of genes that act
75 concomitantly and exhibit coherent patterns of expression across samples.

76 In contrast to existing research in this domain that focuses on prediction of expression level of
77 individual genes from WSIs, in this work we first characterize the gene expression state of a patient
78 in terms of a small number of binary latent factors or gene groups that are discovered in a purely
79 data-driven manner. These can be viewed as overlapping groups of related genes whose expression
80 shows significant inter-dependence across samples. The motivation behind such gene grouping is

81 that, though co-expression is not causation, co-expressed genes show coordinated responses across
82 a significant subgroup of patients hinting that these genes may be part of an underlying biological
83 pathway, protein complexes or disease subtype [19]. We have shown that the discovered gene
84 groups are clinically and pathologically relevant in terms of their association with survival, breast
85 cancer receptor status, histopathological phenotypes, cancer driver genes mutations, biological
86 pathways enrichment and underlying protein-protein interactions, and also therapeutic decision-
87 making. We then propose a bespoke multi-output graph neural network-based computational
88 pathology pipeline to predict the expression state of a patient in terms of these latent factors from
89 their WSIs. This enables identification of spatial histological patterns associated with individual
90 latent factors as well as the overall gene expression profile of a patient. Finally, we have shown
91 that image-based predicted gene group statuses can be used as a latent representation for the
92 prediction of several other downstream clinical tasks such as patient subtyping, and also driver
93 gene alteration status and pathway alteration status.

94 2 Results

95 2.1 Analytic workflow

96 As shown in **Fig 1**, we performed gene expression analysis of the TCGA breast cancer (TCGA-
97 BRCA) cohort ($n = 1084$) to identify 200 groups of genes such that the expression of genes in the
98 same group is maximally statistically co-dependent. This allows us to capture the inter-dependence
99 between expression profiles of different genes and represent the gene expression state of a given
100 patient in the form of 200 binary variables each corresponding to a single group. To underscore
101 the clinical, therapeutic, and biological significance of each gene group, we computed the
102 association of patient gene group status with survival, enrichment for biological pathways and
103 cancer hallmark processes, and also protein-protein and drug-protein interactions.

104 We then used our bespoke graph neural network-based pipeline that takes a WSI as input and
105 predicts the binary status of 200 gene groups simultaneously in an end-to-end manner. This allows
106 us to model the complete gene expression profile of a patient and identify histological imaging
107 patterns associated with each gene group. Furthermore, the proposed approach allows spatially
108 resolved cross-linking of discovered gene groups with visual information contained in the WSI.
109 The interactive visualization portal for the proposed approach (called Histology Gene Groups

110 Xplorer (HiGGsXplore) is available at:
111 (http://tiademos.dcs.warwick.ac.uk/bokeh_app?demo=HiGGsXplore)

112 2.2 Data-Driven discovery of Gene Groups based on co-dependent expression

113 To capture multivariate nonlinear relationships in gene expression patterns across patient samples,
114 we employed Correlation Explanation (CorEx) on RNA-Seq data of the TCGA-BRCA cohort.
115 CorEx can be used to model the underlying dependency structure of a dataset by identifying groups
116 of random variables that in the context of this application can intuitively be viewed as a
117 manifestation of underlying covarying patterns of gene expression profiles of different genes
118 across patients. The input to CorEx is a 1084×5676 matrix where each row is the normalized
119 gene expression score of 5,676 genes with high expression variance or mutation frequency for
120 each of the 1,084 patients. For this data, CorEx identified 200 gene groups that can explain the co-
121 dependence between gene expression patterns observed in the data without loss of information.
122 This allows us to represent the gene expression state of each patient in terms of these 200 binary
123 variables rather than the expression of all individual genes. As these gene expression groups are
124 identified in a purely empirical manner from gene expression data, the expected impact of any
125 human observation biases on the definition of these gene groups is minimal. Furthermore, a single
126 gene can be associated with multiple gene groups which is desirable from a biological point of
127 view as gene products often perform multiple roles within a cell and can be part of multiple
128 interaction networks [20].

129 The gene composition of a selected number of gene groups is shown as word clouds in **Fig 2A** and
130 **SFig 1**. For example, the binary status of Gene Group 0 (G_0) is defined primarily based on the
131 expression patterns of a set of genes (*MLPH*, *GATA3*, *XBPI*, *FOXA1*, *TFF3*, *ESR1*, etc.). The
132 exhaustive list of genes grouped in all 200 gene groups is provided in supplementary data. **Fig 2B**
133 illustrates the underlying co-dependent expression of genes grouped in a selected gene group along
134 with their group status. The heatmaps clearly show that the expression level of genes in Gene
135 Group 3 (G_3) and Gene Group 25 (G_{25}) are significantly co-dependent across patients. For
136 instance, for patients with $G_3 = 1$, the expression level of *ITK*, *IL2*, *PDCD1* or *PDI*, *ITGAL*,
137 *PDCDILG2* or *PD-L2*, and several other genes are high, whereas, for patients with $G_3 = 0$, these
138 genes show under-expression as evident from the figure. For G_{25} , a consistent trend in gene
139 expression can be seen between status = 0 and 1 patients. For example, for patients with G_{25} status

140 = 1, *MYC*, *CHEK1*, *PSME4*, *YES1*, *NRAS*, *TP53*, and several other genes show high expression
141 levels, whereas, *IGFBP4*, *TCEAL3*, *RORC*, *RETSAT*, and others show low expression. Conversely,
142 for patients with $G25 = 0$, the expression patterns of these genes are reversed.

143 This key result lends support to the motivation of this work, i.e., the expression level of multiple
144 genes is significantly and consistently inter-dependent and the overall gene expression state of a
145 patient can be characterized by a small number of latent factors. It also highlights the fact that it is
146 not possible to disentangle the expression status of individual genes and consequently associate an
147 observed phenotype, say in a WSI, with the status of a single gene. We next investigated the
148 pathological significance of these gene groups and analyze their predictability from WSIs.

149 2.3 Pathological Significance of Gene Groups

150 Here we discuss the clinicopathological significance of gene groups to understand the implications
151 of these latent factors for clinical decision-making before analyzing their predictability from
152 imaging.

153 2.3.1 Association of Gene Groups with Cancer Hallmarks and Biological Pathways

154 Through Gene Set Enrichment Analysis (GSEA) we found genes from several gene groups
155 associated with known cancer hallmark processes and biological pathways. In **Fig 2C** we show
156 the enriched terms for cancer hallmark processes in selected gene groups. For example, genes in
157 Gene Group 0, 10 and 25 show enrichment for Estrogen early and late response, KRAS and
158 mTORC1 signalling, Unfolded Protein Response (UPR), p53 pathway and several other hallmark
159 processes. Similarly, we found genes from Gene Group 3, 15 and 30 associated with Inflammatory
160 response, Interferon Alpha and Gamma response, and several other cancer hallmark processes.
161 Additionally, we found genes from several gene groups associated with several cancer hallmark
162 processes (Epithelial-Mesenchymal Transition (EMT), Myc targets V1 and V2, Mitotic spindle,
163 DNA repair, KRAS up and down signalling, etc.) as shown in **SFig 2**.

164 Apart from cancer hallmark processes, several a number of gene groups has shown enrichment
165 enriched for several biological processes (e.g. T-cell receptor signalling, MAPK cascade, negative
166 regulation of programmed cell death, etc.), KEGG pathways (e.g. *PD-L1* expression and *PD-1*
167 checkpoint pathway, JAK-STAT and PI3K-Akt signalling pathway, Th1, Th2 and Th17 cell
168 differentiation, etc.) and WikiPathways (e.g. DNA damage response, Inflammatory response, B

169 Cell receptor signalling, etc.) as can be seen in **SFig 3**, **SFig 4** and **SFig 5**. For example, G3 and
170 several other gene groups have shown enrichment for *PD-L1* expression and *PD-1* checkpoint
171 pathway in cancer which can be a guiding signal for therapeutic decision-making [21].

172 2.3.2 Gene Groups capture clinically important protein-protein and protein-drug interactions

173 We analyzed the protein-protein interaction (PPI) and protein-drug interaction (PDI) of genes in
174 several gene groups with the end goal of identifying which groups involve proteins that can be
175 targeted with known drugs so that the gene group status can be used as a potential indicator to
176 guide therapeutic decision making. **Fig 2D** shows the PPI and PDI of a selected number of genes
177 from G3 and G25. Regarding G3, interaction between *IL2*, *IL2RB* and *IL2RG* can be seen (left
178 figure), which is expected as *IL2* regulates immunity by teaming up with *IL2RB* and *IL2RG* [22],
179 [23]. Similarly, interaction of tacrolimus, an immunosuppressive and anti-inflammatory macrolide
180 that targets the CD4⁺-cells can be seen with *IL2*. As these genes show high expression when G3
181 = 1, therefore patients with G3 = 1 can be considered a candidate for tacrolimus therapy. In
182 reference to G25 (see right figure), TRIM8 a member of the tripartite motif-containing (TRIM)
183 binding with TP53 can be seen, which has been shown to play a role in regulating TP53/p53-
184 mediated pathway [24]. Similarly, interaction of *YES1*, a targetable oncogene can be seen with
185 drugs such as dasatinib, ponatinib, nintedanib and imatinib. When G25 = 0, *YES1* shows high
186 expression, therefore patients with G25 = 0 could be considered as potential candidates for
187 dasatinib therapy [25]. Apart from this, interaction of *TP53* with several other proteins (*CHEK1*,
188 *MAPK3*, *PLAT*, *NINJI*, *HDAC5*, etc.) and drugs (tamoxifen, doxorubicin, paclitaxel, etc.) can be
189 observed.

190 2.3.3 Patient stratification into high and low risk using gene groups status

191 We found the binary status of several gene groups associated with overall survival (OS), disease-
192 specific survival (DSS), and progression-free survival (PFS) of patients. **Fig 3A** shows the Kaplan-
193 Meier (KM) survival curves (DSS, PFS and OS) illustrating patients' stratification based on their
194 gene group status. The KM curves indicate that patients can be stratified into high and low risk
195 groups based on their G25 and G195 status with statistical significance (log-rank test FDR-
196 corrected p-value > 0.05). Additionally, from the figure, patients with G3 = 1 have higher survival
197 rates compared to those with G3 = 0 but the stratification is not statistically significant. Our
198 analysis shows that the number of gene groups with statistically significant risk stratification

199 (multiple-hypothesis corrected log-rank p-value < 0.05) is 25, 3 and 2 for DSS, OS and PFS,
200 respectively as shown in **SFig 6**.

201 2.3.4 Association between Gene Groups and breast cancer receptor status

202 We found the status of several gene groups associated with ER, PR and Her2 status as can be seen
203 in **Fig 3B**. For example, from the figure strong positive association of G25 status with ER (Kendall-
204 tau correlation coefficient $\rho_\tau = 0.68$ and $p < 0.01$) and PR ($\rho_\tau = 0.58$ and $p < 0.01$) status can
205 be seen. This correlation was expected as G25 status is defined by IGFBP4 and other relevant
206 genes whose overexpression has previously been found positively associated with ER and PR
207 status [26]. Similarly, we found G35 and G118 status strongly positively associated with her2
208 status as evident from the figure.

209 2.3.5 Association with PAM50 molecular subtypes and immune subtypes

210 We found the status of several gene groups associated with PAM50 molecular subtypes as can be
211 seen in **Fig 3B**. For example, from the figure, strong positive and negative association of G25
212 status can be seen with Luminal A and basal-like subtypes respectively. Since G25 status has also
213 shown strong association with ER and PR status its correlation with Luminal A (ER-positive, PR-
214 positive and Her2 negative) and basal-like (triple negative) subtype is not surprising but highlights
215 the versatility of gene group definitions.

216 Apart from PAM50 subtypes, we found the status of several gene groups associated with immune
217 subtypes (C1, C2, C3 and C4) defined by Thorsson et al [27] as shown in **Fig 3B**. For example,
218 from the figure strong association of Gene Group 15 (G15) can be seen with C2 ($\rho_\tau = 0.72$, $p <$
219 0.01) and C1 ($\rho_\tau = -0.48$, $p < 0.01$) and C3 ($\rho_\tau = -0.31$, $p < 0.01$). This association is expected
220 as majority of G15 genes (*IFIT3*, *OAS3*, *IFI44L*, etc.) are interferon-regulated genes (IRGs) that
221 play a role in the innate immune response and antiviral defense [28]. These results highlight the
222 fact gene group statuses can be utilized as markers for immune activity as well as existing
223 molecular subtyping of breast cancer patients.

224 2.3.6 Association with mutations in cancer genes

225 We found the status of several gene groups associated with gene point mutation status (MUT) and
226 copy number alteration status (CNA) as evident from **Fig 3B**. For example, from the figure, a
227 strong negative correlation of G25 status with TP53 MUT status ($\rho_\tau = -0.59$, $p < 0.01$) and *MYC*

228 CNA status ($\rho_\tau = -0.26$, $p < 0.01$) can be seen. Similarly, the status of several other gene groups
229 can be seen as positively or negatively associated with MUT status (e.g., *CDH1*, *GATA3* and
230 *PIK3A*) and CNA status (e.g., *ERBB2*, *PK2*, *HEY1*, *FGFR* and *F2F2*) of genes.

231 2.3.7 Association of gene groups with pathologist-assigned histological phenotypes

232 We found gene groups status associated with routine clinical features such as histological types
233 (invasive lobular and ductal carcinoma), histological grade (mitotic count, nuclear pleomorphism
234 and epithelial tubule formation) [30] and the spatial fraction of tumor regions with tumor-
235 infiltrating lymphocytes (TIL Regional Fraction) [31] as evident from **Fig 3B**. For example, from
236 the figure, a positive correlation between G3 status and TIL Regional Fraction can be seen.
237 Similarly, the status of G25 can be seen negatively associated with mitosis, necrosis, nuclear
238 pleomorphism, inflammation and tumor grade, whereas positively associated with invasive lobular
239 carcinoma. Association of G3 binary status with TIL Regional Fraction is expected as its status is
240 defined by the expression level of several immune-related genes (e.g., *IL2*, *CD27*, *CCL5*, *PD-1*
241 and *PD-L2*) [27], [32]. Similarly, G25 status negative association with mitotic count is not
242 surprising as previous studies have found that over-expression of *MYC* (G25 = 1 when *MYC* is
243 over-expressed) impairs mitotic spindle formation [33]. This analysis shows that gene group status
244 can be associated with pathologist-assigned histological phenotypes.

245 2.4 Prediction of Gene Groups from histological imaging

246 To explore the association between phenotypic information contained in the WSI and the
247 expression status of a set of genes in a certain gene group we have developed a novel deep learning
248 based multi-task graph neural network pipeline (*SlideGraph*[∞]) that takes a WSI as input and
249 predicts the status of 200 gene groups simultaneously. The workflow of the proposed approach is
250 shown in **Fig 1B**. It builds on our previous work that can model a WSI as a graph to capture
251 histological context but has been significantly expanded and improved [34].

252 2.4.1 Quantitative results of prediction of individual gene group statuses

253 Our predictive analysis shows that the binary status of a significant number of gene groups can be
254 predicted from histology images with high area under the receiver operating characteristic curve
255 (AUROC). **Fig 4A** shows model performance in terms of mean AUROC. The binary status of
256 many gene groups can be predicted with an AUROC of above 0.60. Additionally, the status of

257 around 29 gene groups is predicted with a high AUROC of above 0.80. For the top 20 best-
258 predicted gene groups we show the AUROC distribution across 1,000 bootstrap runs in **Fig 4B**.
259 From the figure, G0, G100 and G25 status can be predicted with an AUC-ROC of above 0.87 with
260 a narrow confidence interval.

261 To analyze the degree to which the complete gene expression profile of a patient can be predicted
262 from imaging alone, **Fig 4C** displays a histogram of patient-wise cosine similarity between
263 histology image-based inferred gene expression state and true gene expression state. From the plot,
264 the similarity score shows a moderate alignment between the true and predicted gene expression
265 states of each patient (average cosine similarity across all patients of 0.27). Of particular interest
266 are patients whose alignment score is either very high or very low. Some example WSI thumbnails
267 of patients whose expression state is best or poorly predicted from histological imaging are shown
268 in **STable 1**. These results point to the fact that although the status of certain groups can be
269 predicted with high accuracy, it is not possible to fully characterize the overall gene expression
270 state of most patients from histological imaging alone. This result is expected due to both technical
271 and underlying biological reasons. For example, histological imaging and gene expression analysis
272 are carried out on different tissue sections and the latter uses “bulk” tissue. Furthermore, not all
273 gene expression changes will have a phenotypic effect that can be observed in a WSI which in turn
274 allows predictive modelling as illustrated in **SFig 7**. This shows that both whole slide imaging and
275 gene expression analysis carry complementary value in understanding disease mechanisms.

276 2.4.2 Spatial Profiling and histological phenotypes of Gene Groups

277 The proposed graph neural network can map WSI-level predictions of a gene group to spatially
278 localized regions or nodes in the input image. This enables the profiling of local histological
279 patterns linked to gene groups based on their node-level predictions. **Fig 5** shows the spatial
280 profiling of gene groups (G3 and G25 as examples) by visualizing node-level prediction scores
281 from *SlideGraph*[∞]. For both gene groups, an example WSI with its corresponding heatmap
282 highlighting node level prediction score is shown against binary status 0 and 1. The heatmap
283 highlights the spatially resolved contribution of different regions of the WSI towards the
284 expression status of a certain gene group being 0 or 1. More specifically, regions highlighted in
285 redder color are indicative of an association with status = 1, whereas regions highlighted in bluish
286 color are indicative of an association with status = 0 of a particular gene group. It is interesting to

287 note that a given gene group exhibits significant variation in prediction score across different
288 regions of the image, which can be linked to the spatial diversity of localized gene expression
289 patterns throughout the tissue. The localized predictions for other gene groups can be viewed in
290 the online portal ([HiGGsXplore](#)).

291 Using node-level prediction score as a guide, we extracted some regions of interest (ROIs)
292 associated with G3 and G25 status = 0 and 1 from their corresponding WSI as shown in **Fig 5**.
293 ROIs representative of G3 = 1 have a relatively high proportion of inflammatory cells compared
294 to G3 = 0 ROIs where tumor cells appear more pleomorphic. Additionally, for the patient with G3
295 (status = 1), the invasive margin of the tumor, which has a higher density of inflammatory cells, is
296 shown to be correlated with G3 status = 1. Given that G3 status is associated with TIL regional
297 fraction (see Fig 3) and immune response related processes and pathways (see Fig 2C, SFig 2 and
298 SFig 4), therefore tumor-infiltrating lymphocytes (TILs) is the likely histological phenotype
299 associated with G3 (status = 1). This also explains the higher survival probability of G3 (status =
300 1) patients as several studies have found TILs associated with good prognosis [35]. Regarding
301 G25, tubule formation, and normal lobule can be seen in ROIs representative of G25 (status = 1),
302 whereas, in ROIs indicative of G25 (status = 0) the obvious feature is necrosis, and more
303 pleomorphic tumor cells. For the patient with G25 (status = 1), regions of the WSI with tubule
304 formation are highlighted as evident from the ROI. However, for patient with G25 (status = 0)
305 tissue regions with normal lobule received higher score since there was no tissue area with tubule
306 formation. The highlighted spatially resolved histological patterns are concordant with their
307 corresponding enriched cancer hallmark processes (Estrogen response, Immune response and p53
308 signalling) and biological pathways (see Fig 2C, SFig 2 and SFig 4).

309 This analysis shows that the proposed deep learning pipeline has identified relevant spatially
310 resolved histological patterns associated with different gene groups (TILs in the case of G3 and
311 tubule formation in the case of G25) in an automated manner as evident from the heatmaps. It is
312 noteworthy, that in cases where no tubule formation is present in the WSI (see G25 = 0 ROIs), it
313 has highlighted normal lobule which is quite remarkable.

314 2.4.3 Mining differential histological patterns associated with each gene group

315 To explore the association between visual patterns contained in WSIs and gene groups status we
316 identified 25 exemplar patches for each status (0 and 1) of a certain gene group. For these patches,

317 we also computed the cellular composition (counts of neoplastic, inflammatory, connective, and
318 epithelial cells), overall cellularity and mitotic counts. **Fig 6A** shows 10 out of 25 representative
319 patches for each of G3 and G25 status = 0 and status = 1. The main difference between G3 = 0 and
320 1 patches, as seen in the figure, is the presence of lymphoid infiltrate and tumor cellularity. More
321 specifically, G3 = 1 patches have more inflammatory cells and fewer neoplastic cells, whereas the
322 opposite is true for G3 = 0 patches. This differential histological pattern across all patients is
323 concordant with the spatially resolved visual pattern we see in G3 = 0 and 1 ROIs (see Fig 5) and
324 can be used as a histological motif. Additionally, G3 = 0 patches have relatively higher number
325 of mitotic counts compared to G3 = 1. Regarding G25, the striking difference between G25 = 0
326 and G25 = 1 patches is the presence of tubule formation (row 2 patch 2 and 3, row 2 image 2 and
327 3) in the tumor area. As G25 status correlates positively with ER and PR status (see Fig 3B) and
328 previous study has also found ER and PR positive cancers enriched in tubule formation [36],
329 therefore, tubule formation could be the histological phenotype associated with G25 = 1. In
330 contrast, G25 = 0 patches have more pleomorphic sheets of cells and areas of necrosis (row 1
331 image 1 and 3, row 2 image 1 and 2). This pattern agrees with the histopathological phenotypes
332 we observed in **Fig 3B** and **Fig 5**. Finally, G25 = 1 patches show higher mitotic and inflammatory
333 cell counts compared to G25 = 0 patches. Though we are not using any histopathological
334 annotations in training, the predictive model has identified relevant morphometric patterns in an
335 automated manner.

336 Apart from G25 and G3, we found patch-level inflammatory cell counts and mitotic counts
337 statistically significantly associated (Wilcoxon test $p < 0.01$) with the binary status of several
338 other gene groups as shown in **Fig 6B** and **Fig 6C**.

339 2.5 Image-based predicted gene group statuses provide latent space for down- 340 stream predictive modeling

341 Gene expression groups allow us to capture the gene expression profile of a given patient in terms
342 of 200 gene status variables and their prediction through a machine learning model allows us to
343 map histological patterns to these gene groups. However, the predicted statuses of gene groups can
344 also be used as a compressed latent space representation for predictive modelling of other
345 histologically important clinical variables. **Fig (7A-F)** show the predictability of clinical variables
346 based on the predicted gene group statuses as latent variables using a simple linear classifier.

347 PAM50 subtypes such as Basal, Luminal A, Luminal B and Her2 can be predicted from these
348 latent variables with a mean AUROC of 0.90, 0.82, 0.78 and 0.75 respectively. Similarly, the latent
349 representation can also predict the status of ER, PR and Her2 with a mean AUROC of 0.88, 0.79
350 and 0.61 respectively. Apart from this, we found the latent variables predictive of several signalling
351 pathways alteration status, immune subtype, and also genes MUT status and CNA status. For
352 example, TP53 pathway alteration status can be predicted with a mean AUROC of 0.75 from these
353 latent variables [37]. The latent variables can also predict MUT status (14 genes) and CNA status
354 (12 genes) with an AUROC of above 0.60 as evident from **Fig 7E** and **Fig 7F**. For example, TP53
355 point MUT status and *ERRB2* CNA status can be predicted with an AUROC of 0.81 and 0.79
356 respectively, which are higher than baseline results of 0.79 for *TP53* MUT status [38] and 0.62
357 for *ERBB2* MUT status [39]. **Fig 7G** shows some example heatmaps demonstrating spatial
358 profiling of these clinical variables. From figures, ER and PR status have similar highlighted
359 regions, while basal subtypes (ER, PR and Her2 negative) have opposite regions. The heatmaps
360 also show the spatial profiling of Luminal B subtype, and TP53 MUT and pathway alteration
361 status. This clearly illustrates the value of the proposed gene groups for downstream predictive
362 modelling.

363 2.6 Clinical and Therapeutic significance of best-predicted gene groups

364 We found that gene groups predicted with high accuracy ($\text{AUROC} \geq 0.75$) from imaging are
365 significantly associated with disease specific survival (DSS), biological pathways and hallmark
366 processes. All 25 gene groups associated with DSS are predicted with high accuracy from imaging.
367 Besides this, some interesting biological pathways (see **Fig 8**) and cancer hallmark processes (see
368 **SFig 8**) can also be inferred from images based predicted gene groups which can guide histology
369 image-based therapeutic decisions by selecting drugs that target a certain biological pathway (e.g.
370 PI3K-Akt) [40].

371 3 Discussion

372 We performed histological and molecular characterization of breast cancer patients using a purely
373 data-driven approach. Highlighting the limitations of previous methods that predict the expression
374 level of individual genes from histology image, we have shown that significant co-dependencies
375 of different genes across samples (see **Fig 2B**) compromises the ability of deep learning models to

376 identify individual gene level genotype to phenotype mapping. To tackle this, we first grouped
377 genes whose expression patterns are significantly dependent and covarying across samples and
378 then proposed a multi-output graph-based deep learning pipeline (*SlideGraph*[∞]) that predicts
379 both WSI-level and spatially resolved expression status of these gene groups in an end-to-end
380 manner. Using the proposed computational pathology workflow, we demonstrated that the status
381 of a significant number of gene groups can be predicted with high accuracy from imaging. This
382 not only overcomes the limitations of existing image-based gene expression prediction models but
383 provides opportunities to gain biological insights from imaging directly. Finally, we showed that
384 histopathological patterns associated with several gene groups in terms of cellular composition,
385 mitotic counts and exemplar patches can be identified using the proposed computational pathology
386 pipeline.

387 A potential advantage of the employed gene grouping approach is the interpretability of gene
388 groups. The method allows a compact representation of a patient's gene expression state (200
389 binary latent variables) without losing interpretability, which is crucial in this context as it provides
390 insight into biological processes and underlying protein-protein and also drug-protein interactions
391 that can motivate new therapies. Through GSEA, we found genes from several gene groups
392 associated with cancer hallmark processes (e.g. EMT, inflammatory response, estrogen early and
393 late response, mTORC1 signalling, Myc targets, p53 signalling, KRAS up and down signaling)
394 and biological pathways (e.g. Inflammatory response, PD-L1 expression and PD-1 checkpoint,
395 cancer immunotherapy by PD-blockade and EGF/EGFR signalling). Additionally, we have shown
396 that genes in a certain gene group are enriched for protein-protein interaction that can be used for
397 the identification of drugs that modulate the activity of a target protein of interest which will
398 subsequently lead to precise diagnosis of patient tumor.

399 Another important observation regarding gene grouping is that, though the gene groups are defined
400 in a completely data-driven manner without any intelligent selection still they carry significant
401 clinical meaning in terms of association with survival (OS, DSS and PFS), routine clinical
402 biomarkers (ER, PR and Her2 status), driver genes mutation statues, and previously defined
403 PAM50 and Immune subtypes. Apart from this, we found the binary status of several gene groups
404 associated with histopathological annotations which enable direct genotypic to phenotype
405 mapping. Additionally, this genotype to phenotype link can further be validated using GSEA and

406 specialized IHC staining. These results not only validate the clinicopathological significance of
407 these gene groups but also provide a broader picture of an individual tumor by illuminating the
408 interplay between patient gene expression state and several other clinical variables of interest.

409 A striking feature of the proposed approach for mapping patient gene expression status with
410 morphometric patterns contained in the WSIs is its reliability and explainability. Localized
411 histological patterns identified by *SlideGraph*[∞] can be explained in terms of enriched hallmark
412 process, biological pathway and underlying protein-protein interaction, and also through
413 specialized IHC staining and genome sequencing. For example, we found genes from G3 enriched
414 for several immune-related biological processes and pathways including PD-L1 expression and
415 PD-1 checkpoint pathway which in histology images we found associated with a high proportion
416 of TIL. Thought the observation is interesting but still further validation is needed using IHC data.
417 After validation, this will allow the selection of patients for immunotherapy based on routine
418 histology images. Regarding G25 we found tubule formation in majority of G25 = 1 representative
419 patches, which was consistent with IHC ER and PR status and also the associated cancer hallmark
420 process (Estrogen signalling). In contrast, G25 = 0 patches have more pleomorphic sheets of cells
421 several with area of necrosis, which is again concordant with their association with pathologist-
422 assigned phenotypes (necrosis and nuclear pleomorphism), TP53 MUT status and p53 signalling
423 pathway. This show that the proposed deep learning pipeline has identified relevant spatially
424 resolved histological patterns associated with the status of gene groups in an automated manner.

425 Image-based prediction of gene expression state will open doors of gaining biological insights
426 from imaging directly and is expected to be advantageous in both cancer research and clinical
427 setup. In cancer research, the proposed approach can be used for studying the interplay between
428 gene expression and histopathological phenotypes. Additionally, it can also be used by
429 pharmaceutical industries in their drug discovery pipeline when they study the response of lead
430 compounds in early-phase trials. In clinical setup, it will allow cost-effective precision diagnostic
431 from imaging data alone. The proposed computational pathology pipeline not only predicts patient
432 gene expression but also provides a detailed insight in terms of patient survival (OS, DFS and
433 PFS), possible up or downregulated biological processes and their underlying protein-protein
434 interaction, possibly mutated or copy-altered genes, and information about ER, PR and HER2
435 status, PAM50 and immune subtypes. These types of analysis will provide a more detailed insight

436 into an individual tumor in a cost-effective way. It is important to highlight here, that though we
437 managed to predict the expression status of several gene groups with high accuracy and we
438 extensively validated the results, further extensive validation on a large multi-centric dataset is
439 needed before entering into clinical practice.

440 [Acknowledgments](#)

441 M.D. would like to acknowledge the PhD studentship support from GlaxoSmithKline. F.M and
442 N.R supported by the PathLAKE digital pathology consortium which is funded from the Data to
443 Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge
444 Fund, managed and delivered by UK Research and Innovation (UKRI). F.M and M.E also
445 acknowledge funding support from EPSRC EP/W02909X/1.

446 [Author Contributions](#)

447 Conception: FUAAM, NR, KB and MD; Experiment Design: MD, FUAAM, NR, KB, ABH;
448 Bioinformatics analysis: FUAAM, ABH, MD; Pathologist review: LJ; Clinical Review: LJ and
449 LY; Coding and data analysis: MD; Visualization and portal development: ME and MD; Mitotic
450 data analysis: MJ and MD; Writeup: MD and FM with input and review from all authors; Funding
451 acquisition: NR, FUAAM and KB.

452 [Declaration of interests](#)

453 NR is the CSO of Histofy Ltd.

454

455

456

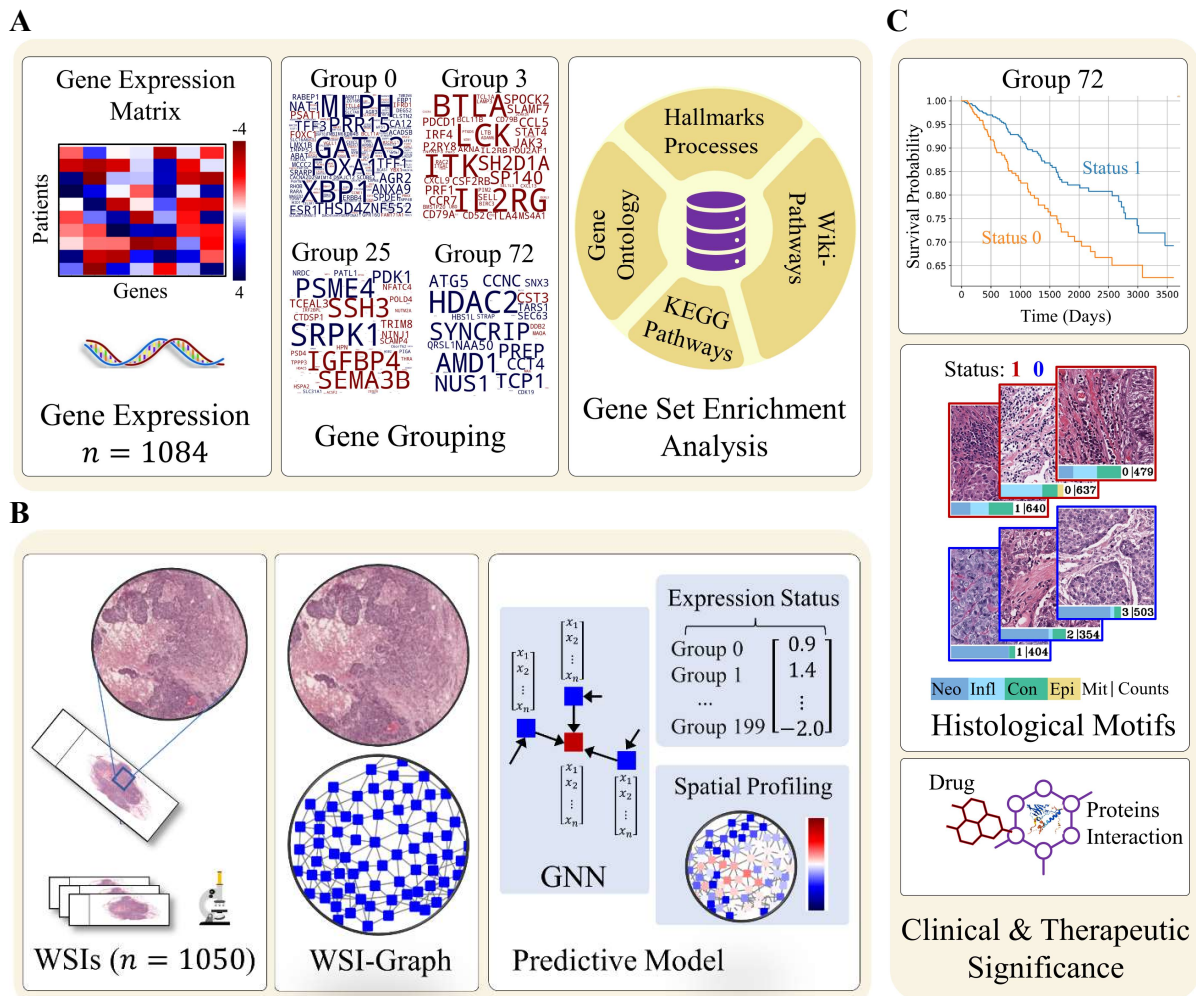


Figure 1: Analytic workflow for patient gene expression state prediction from whole slide images (WSIs). A) Workflow of data-driven discovery of gene groups and their pathological significance is shown. We first identified 200 binary latent factor or gene groups from the gene expression data in a data-driven manner. A gene group can be viewed as overlapping group of genes that exhibit coherent patterns of expression across sample. Word clouds demonstrating the gene composition of different gene groups. The color of the gene indicates whether its median expression across patients is high (red) or low (blue) when gene group status = 1. Afterward, we assessed the biological significance of the genes grouped in different gene groups through gene set enrichment analysis. B) The proposed *SlideGraph*^{oo} pipeline for prediction of gene groups status from WSIs. We first construct graph representation of a WSI and then feed it into a Graph Neural Network (GNN) for predicting WSI-level and spatially resolved expression status of these 200 gene groups. C) Identification of clinically relevant gene groups in term of association with survival and their associated histological motifs. Histology image-based inference of personalized medication by analyzing protein-protein and drug-protein interaction of gene groups.

457

458

459

460

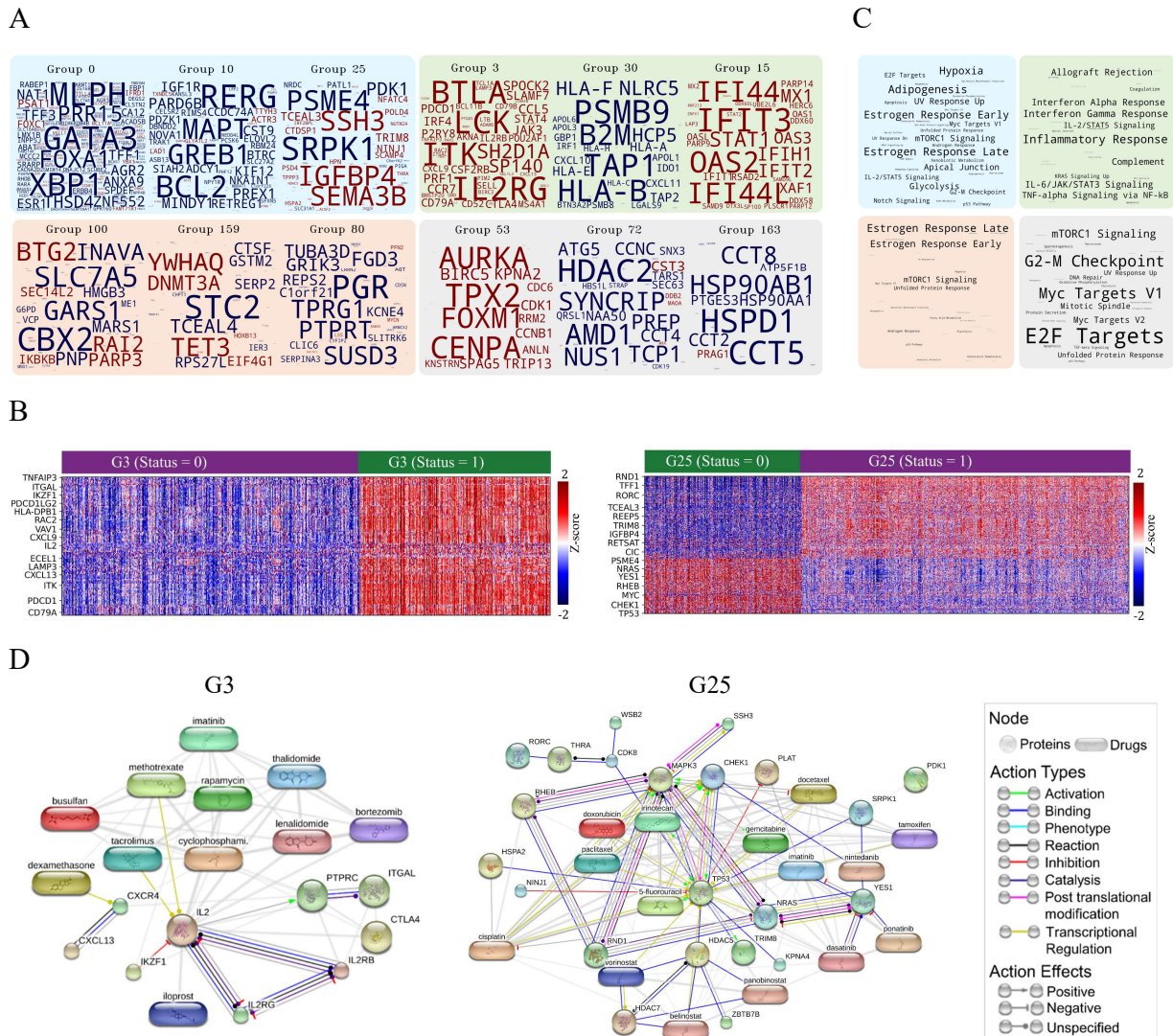


Figure 2: Data Driven Discovery of Gene Groups, their biological and therapeutic significance.

- Word clouds demonstrating the gene composition of different gene groups. The color of the gene indicates whether its median expression across patients is high (red) or low (blue) when gene group status = 1. The font size of gene within a group is proportional to the amount of information that the gene status provides about a particular gene.
- Gene expression profile and group status of genes (one per row) for all patients (one per column) in Gene Group 3 (G3) and Gene Group 25 (G25) are shown.
- Enriched terms for hallmark processes in similar gene groups (note color in A) are shown, with font sizes proportional to the number of gene groups that show enrichment for a certain process.
- Protein-protein and protein-drug interaction of selected genes in G3 (left plot) and G25 (right plot) are shown. Nodes shown in circles represent proteins, while the rounded rectangle shapes represent drugs. The edges between nodes show different types of interaction and potential therapeutic targeting.

461

462

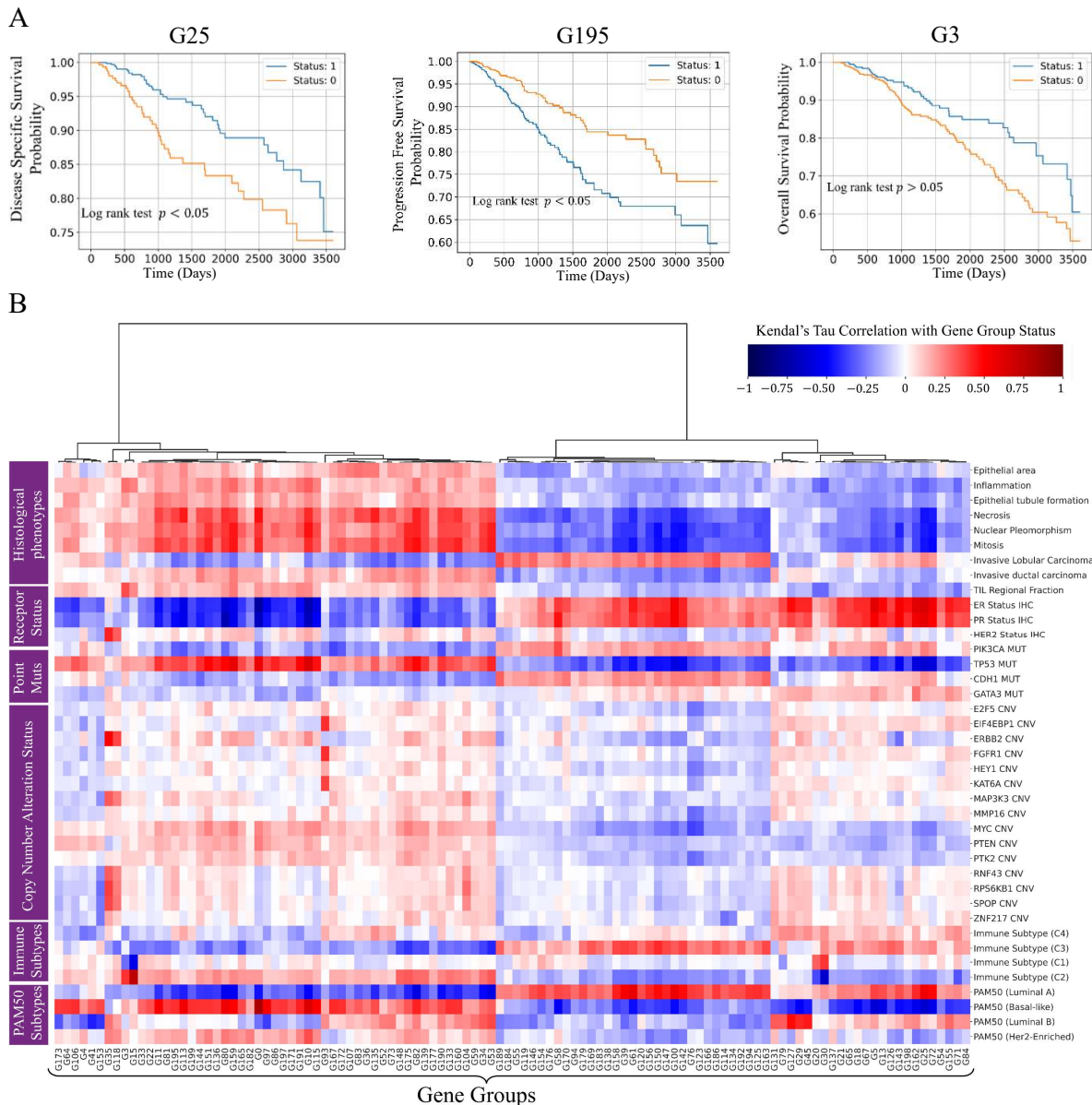


Figure 3: Clinical and pathological significance of gene groups binary status.

- A) Kaplan-Meier curve showing stratification of patient into high and low risk group based on G3 (left plot), G25 (middle plot) and Gene group 195 (G195) binary status. G25 and G195 status is associated with 10-year censored disease specific survival and progression free survival (log rank test FDR corrected p-value < 0.05). G3 status can stratify patient into high and low risk group but FDR corrected p-value is not significant.
- B) Association of gene groups with histological phenotypes, receptor status, genes point mutation status and copy number alteration status, and also immune and PAM50 molecular subtypes. Gene groups are shown along x-axis, and histological phenotypes and other clinical markers are shown along y-axis. Red and blue colors indicate the degree of association between gene groups status and a specific histopathological phenotype or clinical marker. Dark-red color shows strong positive correlation while strong negative correlation is shown using dark-blue color. (Abbreviations - CNV: Copy Number Variations, TIL: Tumor Infiltrating Lymphocytes)

463

464

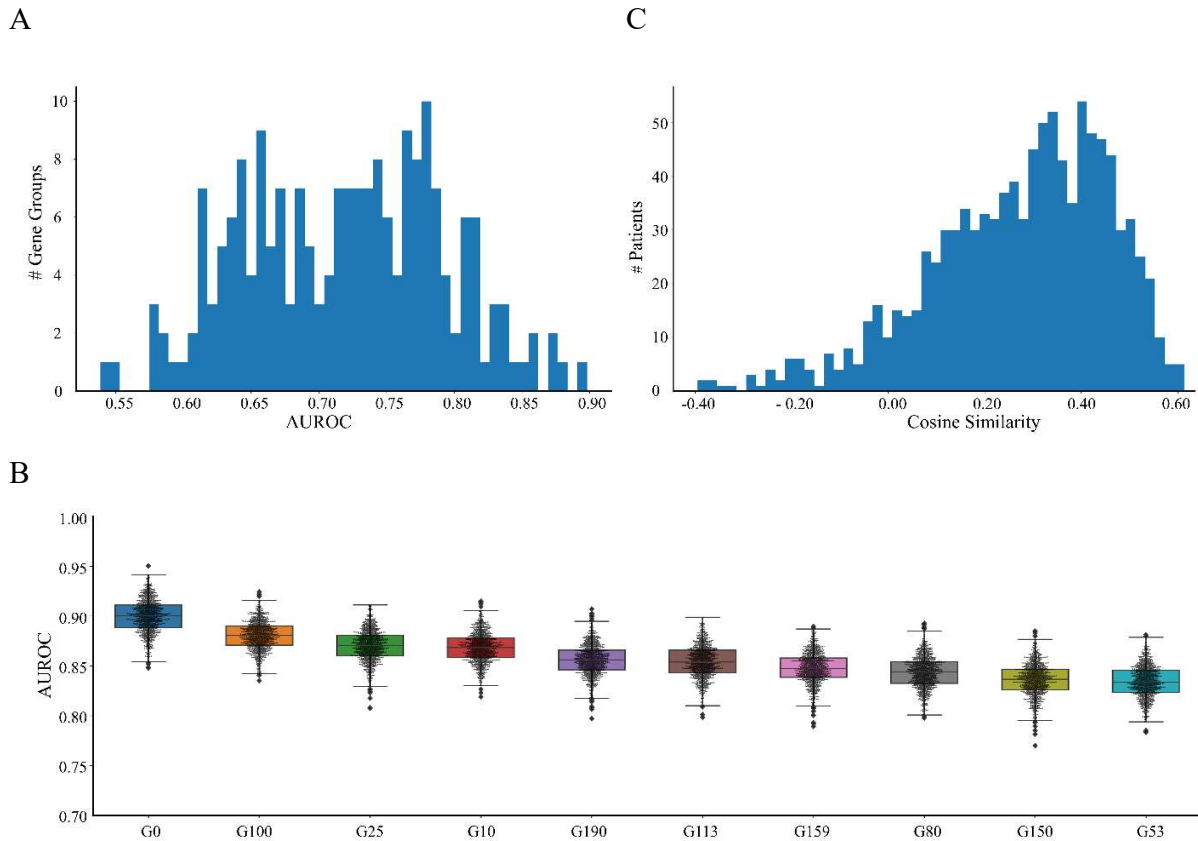


Figure 4 Quantitative result.

- A) Histogram displaying the AUROC at which the binary status of gene groups are predicted from WSIs.
- B) Box plot showing AUROC distribution of top-10 best predicted gene groups across-1,000 bootstrap runs.
- C) Histogram of patient-wise cosine similarity between true and predicted gene expression state.

465

466

467

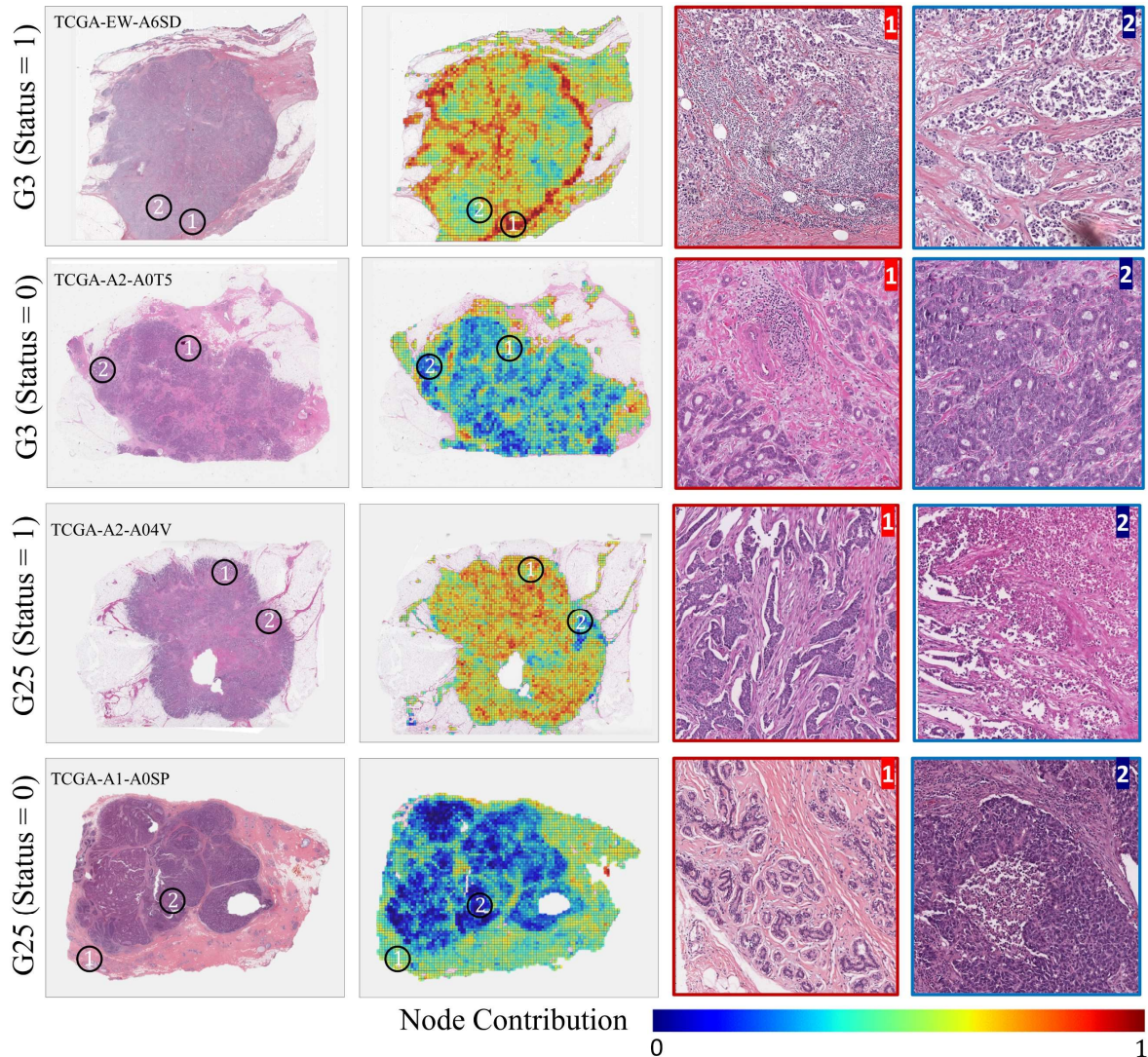


Figure 5: Spatial profiling of gene groups status.

Spatial profiling of gene group 3 (G3) and 25 (G25) is displayed through example WSIs and heatmaps. The heatmaps use pseudo colors (bluish to red) to highlight the spatially resolved contribution of patches to the predicted expression state, with bluish and redder color indicating highly contributing status = 0 and status = 1 regions, respectively. From WSIs we extracted magnified version of highly contributing status = 0 and status = 1 regions (ROIs) outlined by red and blue color, respectively. The black circles highlight regions of WSIs from which ROIs were extracted. For an interactive visualization, please see: tiademos.dcs.warwick.ac.uk/bokeh_app?demo=HiGGsXplore

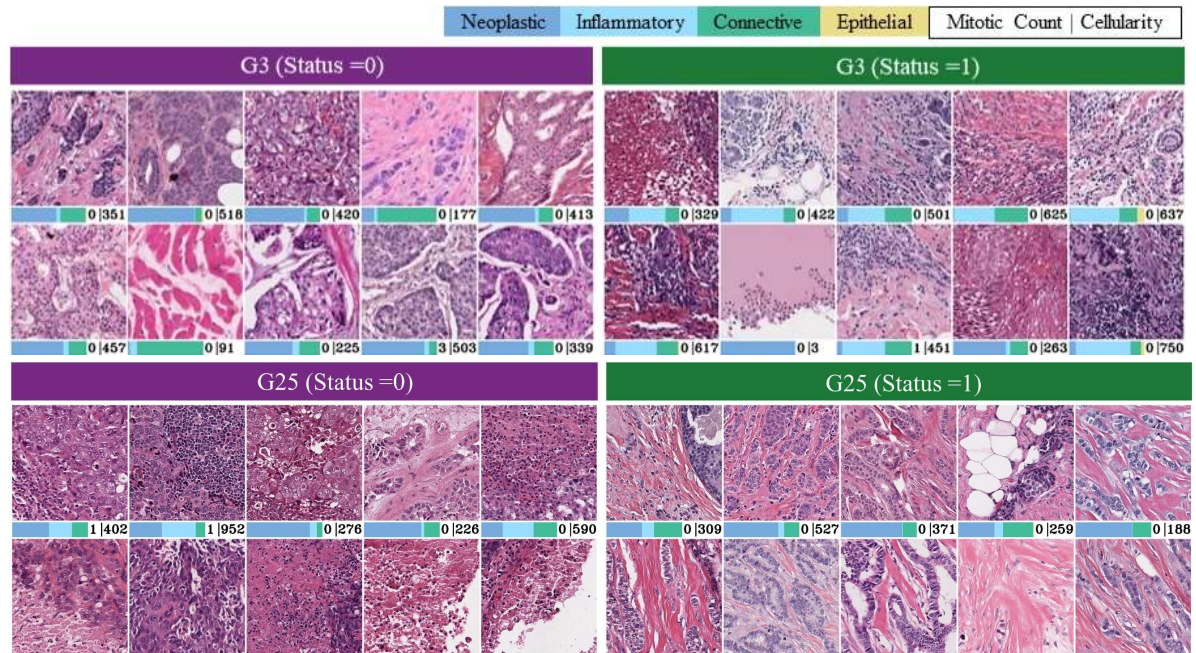
468

469

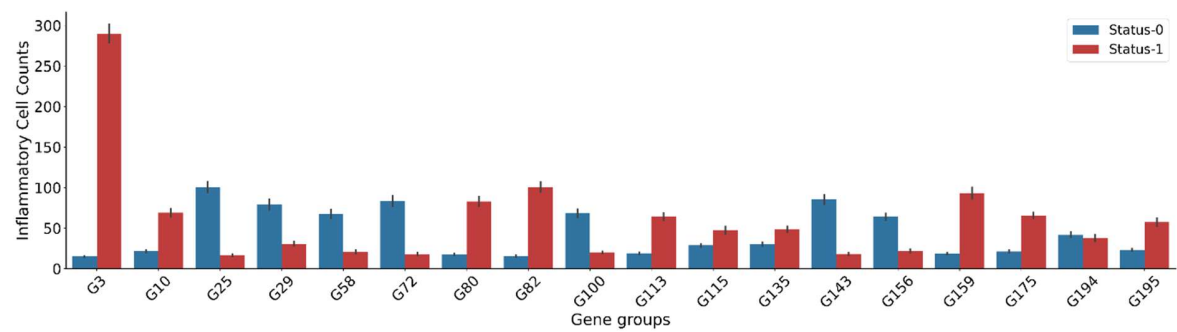
470

471

A



B



C

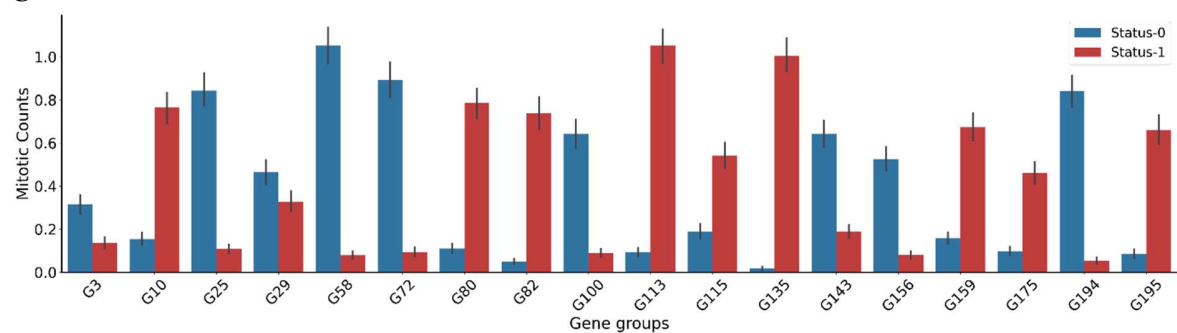


Figure 6: Histological patterns associated with gene groups.

- A) Representative patches of G25 and G3 status 1 and 0 are shown. The bar below the patches shows patch level cellular composition, mitotic counts and cellularity.
 B) Gene groups status (0 and 1) association with patch-level Inflammatory cell counts.
 C) Gene groups status (0 and 1) association with patch-level mitotic cell counts.

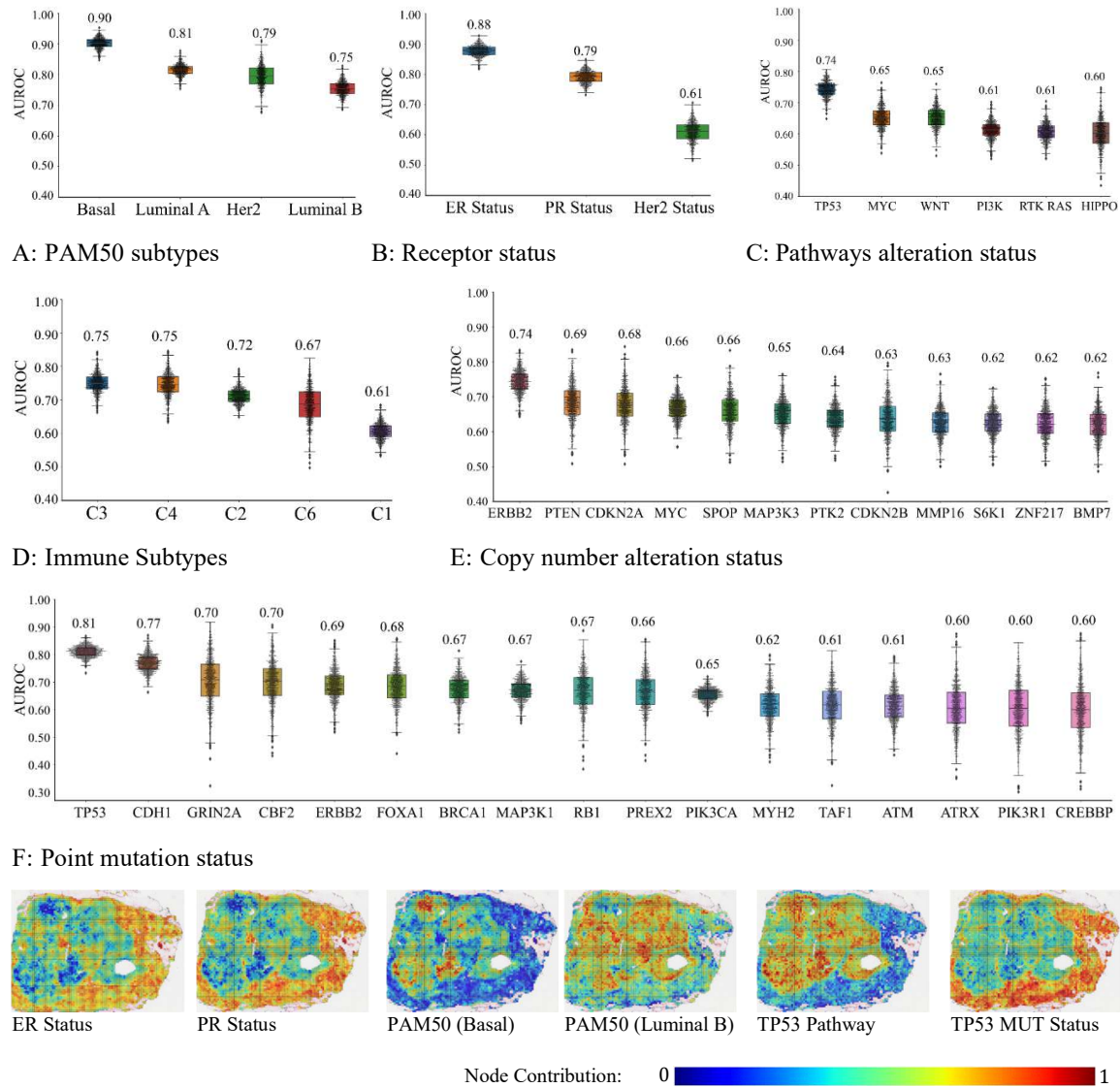


Figure 7: Implication of Image-based predicted gene group statuses for downstream predictive modeling. Prediction of (A) receptor status, (B) PAM50 molecular subtypes, (C) Immune subtypes, (D) pathways alteration status, (E) driver genes copy number alteration status and (F) point mutation status from image-based predicted gene groups status. Each box in the figure shows the AUROC distribution at which a clinical variable is predicted from image-based predicted gene group status across 1,000 bootstrap runs. The scatter plot on top of box plot shows the AUROC values across different bootstrap runs while the numeric value above each box shows the mean AUROC value. (G) Spatial profiling of some routine clinical variables is shown using example heatmaps. The heatmaps use pseudo colors (bluish to red) to highlight the spatially resolved contribution of patches to status = 0 and 1 of a certain clinical variable, with bluish color indicating highly contributing status = 0 regions and red color indicating highly contributing status = 1 regions.

473

474

475

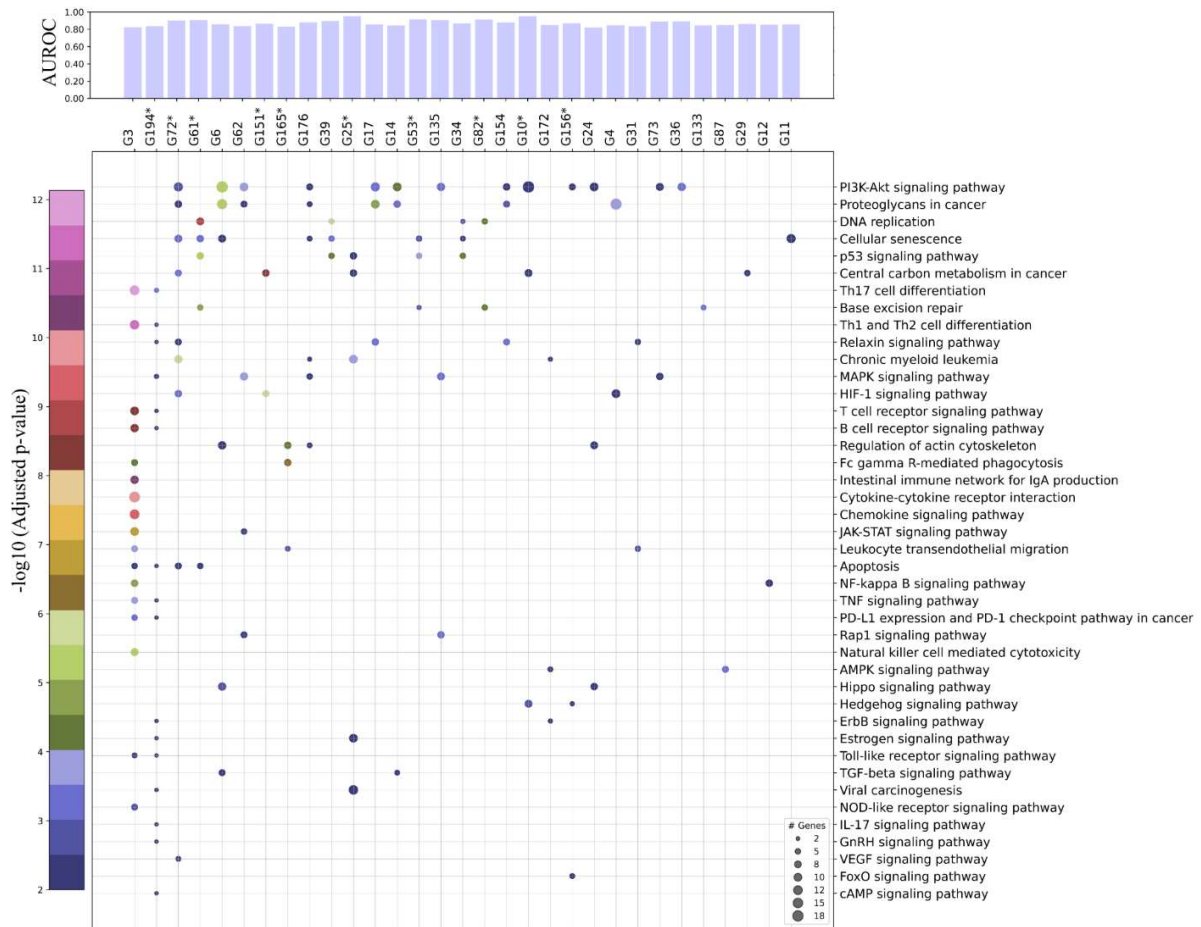


Figure 8: Clinical and Therapeutic significance of best predicted gene group. The scatter plot shows association of gene groups with biological pathways with gene group shown along x-axis (one per column) and corresponding enriched pathways on y-axis (one per row). The size of scatter shows the number of genes from a particular gene group that has shown statistically significant association (FDR adjusted p-value < 0.01) with a certain biological pathway. In the plot the p-value is represented by the color of scatter dots. The top bar plot shows the prediction accuracy (AUROC) at which the status of these gene groups is predicted from histology images. Gene groups that show statistically significant association with disease specific survival are annotated with a * next to the gene group name.

476

477

478

479

480

481

482 4 STAR Methods

483 4.1 Dataset

484 4.1.1 Acquisition and preprocessing of RNA-Seq data

485 We collected RSEM (RNA-Seq by Expectation and Maximization) normalized RNA-Seq data of
486 1084 TCGA breast cancer patients from cBioportal [41], [42]. The gene expression data was
487 obtained using log₂ normalized z-score values of the expression of 5,596 genes having high
488 variance in expression across patient samples along with known oncogenes.

489 4.1.1.1 Acquisition of whole slide images and survival data

490 We collected 1,133 Whole Slide Images (WSIs) of Formalin-Fixed Paraffin-Embedded (FFPE)
491 Hematoxylin and Eosin (H&E) stained tissue section of 1084 patients having breast cancer from
492 the Cancer Genome Atlas (TCGA) [43], [44]. For patients with multiple slides, we selected the
493 one with best visual quality. Additionally for robust analysis, we ignored WSIs with missing
494 baseline resolution information. After slide filtering, we used 1,050 WSIs each belonging to an
495 individual patient to avoid any overlap between training and testing over the same patient. For
496 these patients, we used the survival data from the TCGA standardized clinical dataset called Pan-
497 Cancer Clinical Data Resource (TCGA-CDR) [45] and other clinical data from cBioportal. For
498 these patients we obtained annotation of 11 histopathologic features scored by pathologist from
499 the data released by Thennavan et.al [30].

500 4.2 Data Driven Discovery of Gene Groups with CorEx

501 To model associations between expression profile of different genes we used Total Correlation
502 Explanation (CorEx) on the gene expression matrix M of size $m \times n$ where m and n are the number
503 of patient samples, and genes, respectively [46]. As the expression of different genes is
504 significantly inter-dependent and correlated, CorEx allows us to represent the gene expression
505 state of a patient in terms of a small number of binary variables or gene groups that can capture
506 information contained in the expression of all genes of a given patient with minimal loss. For a
507 detailed mathematical formulation underlying CorEx, the interested reader is referred to the CorEx
508 paper [46]. Given $M_{m \times n}$ as input, the output of CorEx is a matrix $G_{m \times d}$ with each column of
509 G corresponds to a binary latent factor G_k ($k = 1 \dots d$ with $d \ll n$) so that the mutual information
510 between the expression level of genes is minimized after conditioning on G_1, \dots, G_d . In other

511 words, the latent factors identified by CorEx can “*explain away*” the association between
512 expression of various genes. Akin to “loadings” in principal component analysis (PCA), the
513 definition of each binary latent factor G_k is based on mutual information between the expression
514 score of a certain gene and the binary status of G_k across patient samples. This allows us to model
515 each of the latent factors as a ranked (by mutual information) collection or group of genes.
516 However, unlike PCA (or other linear or kernelized dimensionality reduction techniques based on
517 covariance), CorEx can capture non-linear statistical relationships and dependencies between input
518 variables (genes) directly due to its use of mutual information (see comparative analysis in [46]).
519 Furthermore, CorEx produces binary latent factors which can be easier to interpret as the status of
520 a certain gene group for a given patient will either be 0 or 1. We run the algorithm for 100 iterations
521 on the z-score expression of TCGA-BRCA patients for discovering 200 binary latent factors. The
522 number of latent factors were decided based on the TC distribution shown in **SFig 9**. The
523 distribution demonstrates that the overall TC (sum of TCs of all latent factor) plateaus and
524 approaches zero after selecting 200 latent factors. Therefore, we selected 200 latent factors. The
525 binary statuses of these 200 latent variables define the expression state of a patient, where the
526 binary value of each latent variable is defined by the group of genes whose gene expression
527 patterns are substantially co-dependent across samples as shown in **Fig 2B**.

528 4.2.1 Analysis of Biological and Therapeutic Significance of gene groups

529 Hallmark processes and KEGG pathways enrichment for genes in different gene groups were
530 obtained using Enrichr [47]. In line with previous work [20], we selected a maximum of top 400
531 genes from each gene group whose mutual information is greater than 0.002. We passed the gene
532 set to Enrichr which returns the enriched terms across a selected library (in our case KEGG
533 pathway and MSigDB hallmarks) coupled with their statistical significance (FDR-adjusted p-value
534 using Benjamini-Hochberg methods). We used a cutoff value of $p < 0.01$ on the adjusted p-value
535 for statistical significance of an enriched term across the selected library. The protein-protein and
536 drug-protein interactions are analyzed using STITCH [48].

537 4.3 WSI Analysis Pipeline with *SlideGraph*[∞]

538 4.3.1 Preprocessing of whole slide images

539 We segment the tissue regions of WSIs using a tissue segmentation model and ignore regions with
540 tissue artefacts (pen-marking, tissue folding, etc.). Each WSI is then tiled into patches of size

541 512×512 pixels at a spatial resolution of 0.50 microns-per-pixel (MPP). Patches capturing less
542 than 40% of informative tissue area (pixels with intensity higher than 200) are discarded, and the
543 remaining patches (both tumor and non-tumor) are used.

544 4.3.2 WSI-graph Construction

545 A graph $= (V, E)$ is defined by a vertex set V , and an edge set E . The set $V = \{v_i | i = 1, \dots, N\}$
546 defines nodes in a graph (in our case is the set of patches in a WSI) while connectivity between
547 nodes is defined by the edges E . Each node $v_i = (g_i, h_i)$ captures the spatial location (g_i), and
548 feature representation (h_i) of a patch in the WSI. We obtain the feature representation $h_i \in \mathcal{R}^{1024}$
549 of a patch x_i by extracting latent representation from ShuffleNet [49] pretrained on ImageNet
550 [50]. The edge set E is obtained by connecting nodes to the neighboring nodes (distance less than
551 4000 pixels) using Delaunay triangulation. If two nodes v_i and v_j are connected, then there will
552 be an edge $e_{ij} \in E$.

553 4.4 Gene expression state prediction using Graph Neural Network

554 We pass the graph representation of a WSI through a Graph Neural Network (GNN) for predicting
555 the node-level and WSI-level expression status of all gene groups simultaneously. In this work,
556 we have developed a custom multi-output GNN that predicts the patch-level and WSI-level
557 expression statuses of different gene groups in an end-to-end manner. Node level representation is
558 passed through EdgeConv layers $L = \{1, 2, 3\}$. Each EdgeConv layer [51] updates the
559 representation of each node in the graph by aggregating the information from their neighboring
560 node and generates embedding for successive layers. For a node in layer l at index m the output
561 embedding of EdgeConv layer can mathematically be written as follows:

$$562 \quad h_m^l = \sum_{k \in \mathfrak{N}(m)} \mathcal{H}^l (h_m^{l-1} \parallel h_k^{l-1} - h_m^{l-1})$$

563 In the above equation $h_m^0 = h_m$, $\mathfrak{N}(m)$ represents the neighboring nodes of m , and \mathcal{H}^l denote a
564 neural network. EdgeConv operation is trying to combine information of a node h_m^l and
565 neighboring nodes $\mathfrak{N}(m)$. Since we are using three EdgeConv layers, each node is expected to
566 capture information from the neighboring nodes that are less than 5-hops apart in the WSI-graph.

567 For spatial profiling for gene expression groups, the feature representation \mathbf{h}_m^l of a node $\mathbf{v}_m =$
568 $(\mathbf{g}_j, \mathbf{h}_j) \in V$ is passed as input to a multilayer perceptron $f_l(\mathbf{v}_m) = f(\mathbf{h}_m^l)$ for generating node
569 level prediction score which is then aggregated across all layers for getting patch level prediction
570 score for all gene groups.

571
$$f(\mathbf{v}_m) = \sum_{l=0}^L f_l(\mathbf{h}_m^l)$$

572 The WSI-level score for the expression status of all gene groups is obtained by pooling and
573 aggregating node-level prediction scores as follows:

574
$$F(G) = \sum_{\forall m \in V} f(\mathbf{v}_m)$$

575 The trainable parameters of the EdgeConv layers and node-level classifiers are learned in an end-
576 end manner using backpropagation. In a training batch of size N , the model predicted score for
577 $k = \{1 \dots K\}$ binary latent factors are compared with their ground truth value using pairwise
578 ranking loss [34], mathematically formulated as follows:

579
$$\mathcal{L} = \sum_k \sum_{(a,b) \in P_k} \max\left(0, 1 - (f^k(X_a) - f^k(X_b))\right)$$

580 Here $P_k = \{(a, b) | y_a^k > y_b^k, a, b = 1 \dots N\}$ is the set of all pair of patients (a, b) where the
581 expression status of patient a is greater than patient b for latent factor k . Minimization of the loss
582 function $\mathcal{L}(\cdot)$ will enforce the model to rank status = 1 patients higher than status = 0 for all latent
583 factors.

584 4.5 Training and evaluation of *SlideGraph*[∞]

585 We trained and evaluated the performance of *SlideGraph*[∞] using 5-fold cross-validation, in which
586 the dataset is subsampled into five 80/20 non-overlapping splits. The model is trained on 80% of
587 the data and 20% data is held out for testing. From the training data we randomly select 10% of
588 the data for parameter tuning and optimization. We train *SlideGraph*[∞] on the training set for 300
589 epochs using the Adam optimizer with an initial learning rate and weight decay of 0.001 and
590 0.0001, respectively. In each epoch, the training set is sampled into mini-batches of size 8, and the

591 learnable parameters of *SlideGraph*[∞] are updated using adaptive momentum based optimizer.
592 To avoid overfitting, we stop the model training early, if performance over the validation set does
593 not improve for 20 consecutive epochs. During training, we maintain a queue of size 10 for
594 tracking the best models based on their performance over the validation set. More specifically, we
595 insert the model into the queue if the validation loss at epoch n is less than the loss at epoch $n - 1$.
596 For test set inference, we ensemble the prediction score of all the models in the queue by averaging
597 the prediction score and using that as the final prediction. For quantitative performance assessment,
598 we report area under the receiver operating characteristic curve (AUROC) over the test set.

599 4.6 Spatial Profiling of Gene Groups and visualization

600 For a given WSI, the spatially resolved contribution of different tissue regions toward the
601 expression status of a certain gene groups can be visualized. We developed an online portal
602 (http://tiademos.dcs.warwick.ac.uk/bokeh_app?demo=HiGGsXplore) which can assist user in
603 spatially resolved cross-linking of genotype-phenotype mapping in terms of these gene groups.
604 More specifically, the portal uses WSI coupled with node level prediction of different gene group
605 and then show the node level prediction in the form of an interactive heatmap. Additionally, the
606 tool can also show different histological features when the user hover over a node in the graph.

607 4.7 Identification of Histological motifs

608 To uncover cellular and morphometric patterns associated with the expression status (0, or 1) of a
609 particular gene group we divided patients into two groups (status = 0 and status = 1). For each
610 group, we select 50 patients whose expression statuses are accurately predicted from their WSIs.
611 From each of these WSIs, for patients with status = 1, we extract the highest scoring (based on
612 node-level score) 1% patches, while for status = 0, we extract the lowest scoring patches and then
613 cluster the patches within each group for getting representative patterns. Within each group (status
614 = 0, and 1) we cluster the patches using 25-medoid clustering. After clustering, we get 25 visual
615 patterns (histological motifs) representative of expression status = 0 and status = 1 of a certain
616 gene group.

617 4.8 Cellular composition estimations

618 We estimated the counts of neoplastic, inflammatory, connective, and normal epithelial cells
619 present in a patch using our in-house cellular composition predictor ALBRT. ALBRT takes a patch
620 of size 256×256 at a spatial resolution of 0.25 MPP and predicts the counts of the
621 aforementioned types of cells present in it. We extracted patches of size 256×256 at 0.25 MPP
622 using (x, y) of coordinates of 512×512 at 0.50 MPP. For each 512×512 patch, we obtained the
623 cellular composition estimates by aggregating ALBRT-predicted cellular estimates of around 16
624 256×256 patches. The cellularity was computed by summing the counts of neoplastic,
625 inflammatory, connective and epithelial cells present in a 512×512 patch.

626 4.9 Estimation of mitotic counts

627 Mitosis detection has been done using the state-of-the-art “mitosis detection: fast and slow”
628 (MDFS) method [52]. MDFS is a two-stage method where mitotic candidates are first detected
629 using a fully convolutional neural network and then refined by a deeper CNN classifier. Several
630 techniques have been incorporated during the training of the MDFS to make it robust against
631 domain shift problems seen in histology images and generalize better to unseen images. After
632 detecting mitotic figures, we estimate the patch-level mitotic counts by counting all the detected
633 mitoses in the patch.

634 4.10 Training and evaluation of Downstream predictors

635 We train separate multi-output perceptron for predicting the receptor status, PAM50 molecular
636 subtypes, Immune subtypes, pathways alteration status, genes point mutation status and copy
637 number alteration status using *SlideGraph*[∞] predicted gene groups status as features. The
638 classifier for each downstream task is trained and evaluated using same loss function and training
639 and validation protocol employed for *SlideGraph*[∞] training and evaluation. After cross-
640 validation, we get the downstream classifier prediction score for a particular clinical variable of
641 interest for all patients. For performance we subsample 67% of the patients 1,000 times with
642 replacement, and compute the AUROC between ground truth and model predicted score.

643 Data and code availability

644 Whole slides images (WSIs) and corresponding genomic data and clinical data of all TCGA
645 patients used in the study can be downloaded from NIH Genomic Data Common Portal at this link:

646 <https://portal.gdc.cancer.gov/> . All genomic and histological analysis was performed in python.
647 The deep learning model *SlideGraph*^{oo} was developed using PyTorch Geometric library. Code
648 and documentation of all python script used in the study can be found at:
649 <https://github.com/engrodawood/HiGGsXplore>.

650 5 References

- 651 [1] D. Hanahan, “Hallmarks of Cancer: New Dimensions,” *Cancer Discovery*, vol. 12, no. 1, pp.
652 31–46, Jan. 2022, doi: 10.1158/2159-8290.CD-21-1059.
- 653 [2] Y. J. Heng *et al.*, “The molecular basis of breast cancer pathological phenotypes,” *J Pathol*,
654 vol. 241, no. 3, pp. 375–391, Feb. 2017, doi: 10.1002/path.4847.
- 655 [3] J. S. Parker *et al.*, “Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes,”
656 *J Clin Oncol*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.
- 657 [4] C. Sweeney *et al.*, “Intrinsic Subtypes from PAM50 Gene Expression Assay in a Population-
658 Based Breast Cancer Cohort: Differences by Age, Race, and Tumor Characteristics,” *Cancer*
659 *Epidemiology, Biomarkers & Prevention*, vol. 23, no. 5, pp. 714–724, May 2014, doi:
660 10.1158/1055-9965.EPI-13-1023.
- 661 [5] J. A. Sparano and S. Paik, “Development of the 21-gene assay and its application in clinical
662 practice and clinical trials,” *J Clin Oncol*, vol. 26, no. 5, pp. 721–728, Feb. 2008, doi:
663 10.1200/JCO.2007.15.1068.
- 664 [6] M. Buyse *et al.*, “Validation and Clinical Utility of a 70-Gene Prognostic Signature for
665 Women With Node-Negative Breast Cancer,” *JNCI: Journal of the National Cancer*
666 *Institute*, vol. 98, no. 17, pp. 1183–1192, Sep. 2006, doi: 10.1093/jnci/djj329.
- 667 [7] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,”
668 *Nat Rev Genet*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.
- 669 [8] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nat Methods*, vol.
670 6, no. 5, pp. 377–382, May 2009, doi: 10.1038/nmeth.1315.
- 671 [9] S. Picelli *et al.*, “Smart-seq2 for sensitive full-length transcriptome profiling in single cells,”
672 *Nat Methods*, vol. 10, no. 11, pp. 1096–1098, Nov. 2013, doi: 10.1038/nmeth.2639.
- 673 [10] V. Marx, “Method of the Year: spatially resolved transcriptomics,” *Nat Methods*, vol. 18, no.
674 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41592-020-01033-y.
- 675 [11] P. L. Ståhl *et al.*, “Visualization and analysis of gene expression in tissue sections by spatial
676 transcriptomics,” *Science*, vol. 353, no. 6294, pp. 78–82, Jul. 2016, doi:
677 10.1126/science.aaf2403.
- 678 [12] C. R. Merritt *et al.*, “Multiplex digital spatial profiling of proteins and RNA in fixed tissue,”
679 *Nat Biotechnol*, vol. 38, no. 5, Art. no. 5, May 2020, doi: 10.1038/s41587-020-0472-9.
- 680 [13] M. Dawood, K. Branson, and N. M. Rajpoot, “All You Need is Color: Image based Spatial
681 Gene Expression Prediction using Neural Stain Learning,” in *Joint European Conference on*
682 *Machine Learning and Knowledge Discovery in Databases*, Springer, 2021, pp. 437–450.
- 683 [14] B. He *et al.*, “Integrating spatial gene expression and breast tumour morphology via deep
684 learning,” *Nature Biomedical Engineering*, vol. 4, no. 8, Art. no. 8, Aug. 2020, doi:
685 10.1038/s41551-020-0578-x.

- 686 [15] B. Schmauch *et al.*, “A deep learning model to predict RNA-Seq expression of tumours from
687 whole slide images,” *Nature Communications*, vol. 11, no. 1, Art. no. 1, Aug. 2020, doi:
688 10.1038/s41467-020-17678-4.
- 689 [16] Y. Wang *et al.*, “Predicting Molecular Phenotypes from Histopathology Images: A
690 Transcriptome-Wide Expression–Morphology Analysis in Breast Cancer,” *Cancer Research*,
691 vol. 81, no. 19, pp. 5115–5126, Oct. 2021, doi: 10.1158/0008-5472.CAN-21-0482.
- 692 [17] A. Alsaafin, A. Safarpoor, M. Sikaroudi, J. D. Hipp, and H. R. Tizhoosh, “Learning to predict
693 RNA sequence expressions from whole slide images with applications for search and
694 classification,” *Commun Biol*, vol. 6, no. 1, Art. no. 1, Mar. 2023, doi: 10.1038/s42003-023-
695 04583-x.
- 696 [18] A.-R. Carvunis, F. Roth, M. Calderwood, M. Cusick, G. Superti-Furga, and M. Vidal,
697 “Interactome Networks,” in *Handbook of Systems Biology Concepts and Insights*, 2013, pp.
698 45–63. doi: 10.1016/B978-0-12-385944-0.00003-4.
- 699 [19] P. Paci, G. Fiscon, F. Conte, R.-S. Wang, L. Farina, and J. Loscalzo, “Gene co-expression in
700 the interactome: moving from correlation toward causation via an integrated approach to
701 disease module discovery,” *npj Syst Biol Appl*, vol. 7, no. 1, Art. no. 1, Jan. 2021, doi:
702 10.1038/s41540-020-00168-0.
- 703 [20] S. Pepke and G. Ver Steeg, “Comprehensive discovery of subsample gene expression
704 components by information explanation: therapeutic implications in cancer,” *BMC Medical
705 Genomics*, vol. 10, no. 1, p. 12, Mar. 2017, doi: 10.1186/s12920-017-0245-6.
- 706 [21] C.-J. Li, L.-T. Lin, M.-F. Hou, and P.-Y. Chu, “PD-L1/PD-1 blockade in breast cancer: The
707 immunotherapy era (Review),” *Oncology Reports*, vol. 45, no. 1, pp. 5–12, Jan. 2021, doi:
708 10.3892/or.2020.7831.
- 709 [22] T. A. Waldmann, “The biology of interleukin-2 and interleukin-15: implications for cancer
710 therapy and vaccine design,” *Nat Rev Immunol*, vol. 6, no. 8, Art. no. 8, Aug. 2006, doi:
711 10.1038/nri1901.
- 712 [23] P. Durda *et al.*, “Plasma Levels of Soluble Interleukin-2 Receptor α ,” *Arteriosclerosis,
713 Thrombosis, and Vascular Biology*, vol. 35, no. 10, pp. 2246–2253, Oct. 2015, doi:
714 10.1161/ATVBAHA.115.305289.
- 715 [24] F. Mastropasqua *et al.*, “TRIM8 restores p53 tumour suppressor function by blunting N-MYC
716 activity in chemo-resistant tumours,” *Molecular Cancer*, vol. 16, no. 1, p. 67, Mar. 2017, doi:
717 10.1186/s12943-017-0634-7.
- 718 [25] I. Garmendia, E. Redin, L. M. Montuenga, and A. Calvo, “YES1: A Novel Therapeutic
719 Target and Biomarker in Cancer,” *Molecular Cancer Therapeutics*, vol. 21, no. 9, pp. 1371–
720 1380, Sep. 2022, doi: 10.1158/1535-7163.MCT-21-0958.
- 721 [26] K. Mita *et al.*, “Prognostic Significance of Insulin-like Growth Factor Binding Protein
722 (IGFBP)-4 and IGFBP-5 Expression in Breast Cancer,” *Japanese Journal of Clinical
723 Oncology*, vol. 37, no. 8, pp. 575–582, Aug. 2007, doi: 10.1093/jjco/hym066.
- 724 [27] V. Thorsson *et al.*, “The Immune Landscape of Cancer,” *Immunity*, vol. 48, no. 4, pp. 812-
725 830.e14, Apr. 2018, doi: 10.1016/j.immuni.2018.03.023.
- 726 [28] S. Piera-Velazquez, F. A. Mendoza, S. Addya, D. Pomante, and S. A. Jimenez, “Increased
727 expression of interferon regulated and antiviral response genes in CD31+/CD102+ lung
728 microvascular endothelial cells from systemic sclerosis patients with end-stage interstitial
729 lung disease,” *Clin Exp Rheumatol*, vol. 39, no. 6, pp. 1298–1306, 2021, doi:
730 10.55563/clinexprheumatol/ret1kg.

- 731 [29] M. Wang, J. Li, J. Huang, and M. Luo, “The Predictive Role of PIK3CA Mutation Status on
732 PI3K Inhibitors in HR+ Breast Cancer Therapy: A Systematic Review and Meta-Analysis,”
733 *Biomed Res Int*, vol. 2020, p. 1598037, May 2020, doi: 10.1155/2020/1598037.
- 734 [30] A. Thennavan *et al.*, “Molecular analysis of TCGA breast cancer histologic types,” *Cell*
735 *Genom*, vol. 1, no. 3, p. 100067, Dec. 2021, doi: 10.1016/j.xgen.2021.100067.
- 736 [31] J. Saltz *et al.*, “Spatial Organization and Molecular Correlation of Tumor-Infiltrating
737 Lymphocytes Using Deep Learning on Pathology Images,” *Cell Reports*, vol. 23, no. 1, pp.
738 181-193.e7, Apr. 2018, doi: 10.1016/j.celrep.2018.03.086.
- 739 [32] T. Jiang, C. Zhou, and S. Ren, “Role of IL-2 in cancer immunotherapy,” *Oncoimmunology*,
740 vol. 5, no. 6, p. e1163462, Apr. 2016, doi: 10.1080/2162402X.2016.1163462.
- 741 [33] J. Rohrberg *et al.*, “MYC Dysregulates Mitosis, Revealing Cancer Vulnerabilities,” *Cell*
742 *Reports*, vol. 30, no. 10, pp. 3368-3382.e7, Mar. 2020, doi: 10.1016/j.celrep.2020.02.041.
- 743 [34] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, and F. Minhas, “SlideGraph+: Whole
744 slide image level graphs to predict HER2 status in breast cancer,” *Medical Image Analysis*,
745 vol. 80, p. 102486, Aug. 2022, doi: 10.1016/j.media.2022.102486.
- 746 [35] N. AiErken *et al.*, “High PD-L1 Expression Is Closely Associated With Tumor-Infiltrating
747 Lymphocytes and Leads to Good Clinical Outcomes in Chinese Triple Negative Breast
748 Cancer Patients,” *Int J Biol Sci*, vol. 13, no. 9, pp. 1172–1179, 2017, doi: 10.7150/ijbs.20868.
- 749 [36] B. Weigelt, F. C. Geyer, and J. S. Reis-Filho, “Histological types of breast cancer: How
750 special are they?,” *Mol Oncol*, vol. 4, no. 3, pp. 192–208, Jun. 2010, doi:
751 10.1016/j.molonc.2010.04.004.
- 752 [37] F. Sanchez-Vega *et al.*, “Oncogenic Signaling Pathways in The Cancer Genome Atlas,” *Cell*,
753 vol. 173, no. 2, pp. 321-337.e10, Apr. 2018, doi: 10.1016/j.cell.2018.03.035.
- 754 [38] P. Keller, M. Dawood, and F. ul A. A. Minhas, “Maximum Mean Discrepancy Kernels for
755 Predictive and Prognostic Modeling of Whole Slide Images.” arXiv, Jan. 23, 2023. doi:
756 10.48550/arXiv.2301.09624.
- 757 [39] H. Qu *et al.*, “Genetic mutation and biological pathway prediction based on whole slide
758 images in breast carcinoma using deep learning,” *npj Precis. Onc.*, vol. 5, no. 1, Art. no. 1,
759 Sep. 2021, doi: 10.1038/s41698-021-00225-9.
- 760 [40] Y. He *et al.*, “Targeting PI3K/Akt signal transduction for cancer therapy,” *Sig Transduct*
761 *Target Ther*, vol. 6, no. 1, Art. no. 1, Dec. 2021, doi: 10.1038/s41392-021-00828-5.
- 762 [41] E. Cerami *et al.*, “The cBio cancer genomics portal: an open platform for exploring
763 multidimensional cancer genomics data,” *Cancer Discov*, vol. 2, no. 5, pp. 401–404, May
764 2012, doi: 10.1158/2159-8290.CD-12-0095.
- 765 [42] J. Gao *et al.*, “Integrative analysis of complex cancer genomics and clinical profiles using the
766 cBioPortal,” *Sci Signal*, vol. 6, no. 269, p. p11, Apr. 2013, doi: 10.1126/scisignal.2004088.
- 767 [43] K. A. Hoadley *et al.*, “Cell-of-Origin Patterns Dominate the Molecular Classification of
768 10,000 Tumors from 33 Types of Cancer,” *Cell*, vol. 173, no. 2, pp. 291-304.e6, Apr. 2018,
769 doi: 10.1016/j.cell.2018.03.022.
- 770 [44] D. C. Koboldt *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*,
771 vol. 490, no. 7418, Art. no. 7418, Oct. 2012, doi: 10.1038/nature11412.
- 772 [45] J. Liu *et al.*, “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality
773 Survival Outcome Analytics,” *Cell*, vol. 173, no. 2, pp. 400-416.e11, Apr. 2018, doi:
774 10.1016/j.cell.2018.02.052.
- 775 [46] G. V. Steeg and A. Galstyan, “Maximally Informative Hierarchical Representations of High-
776 Dimensional Data.” arXiv, Jan. 30, 2015. doi: 10.48550/arXiv.1410.7404.

- 777 [47] M. V. Kuleshov *et al.*, “Enrichr: a comprehensive gene set enrichment analysis web server
778 2016 update,” *Nucleic Acids Res*, vol. 44, no. Web Server issue, pp. W90–W97, Jul. 2016,
779 doi: 10.1093/nar/gkw377.
- 780 [48] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, “STITCH 5:
781 augmenting protein–chemical interaction networks with tissue and affinity data,” *Nucleic*
782 *Acids Res*, vol. 44, no. Database issue, pp. D380–D384, Jan. 2016, doi: 10.1093/nar/gkv1277.
- 783 [49] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional
784 Neural Network for Mobile Devices,” in *2018 IEEE/CVF Conference on Computer Vision*
785 *and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 6848–6856. doi:
786 10.1109/CVPR.2018.00716.
- 787 [50] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int J Comput*
788 *Vis*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
- 789 [51] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic Graph
790 CNN for Learning on Point Clouds.” arXiv, Jun. 11, 2019. Accessed: Feb. 03, 2023. [Online].
791 Available: <http://arxiv.org/abs/1801.07829>
- 792 [52] M. Jahanifar, A. Shephard, N. Zamanitajeddin, S. E. A. Raza, and N. Rajpoot, “Stain-Robust
793 Mitotic Figure Detection for MIDOG 2022 Challenge.” arXiv, Aug. 26, 2022. doi:
794 10.48550/arXiv.2208.12587.
795

796

797

798

799

800

801

802

803

804

805

806

807

808

809 **Supplementary Materials: Data-Driven Modelling of Gene Expression States**
810 **in Breast Cancer and their Prediction from Routine Whole Slide Images**

811

812

813

814

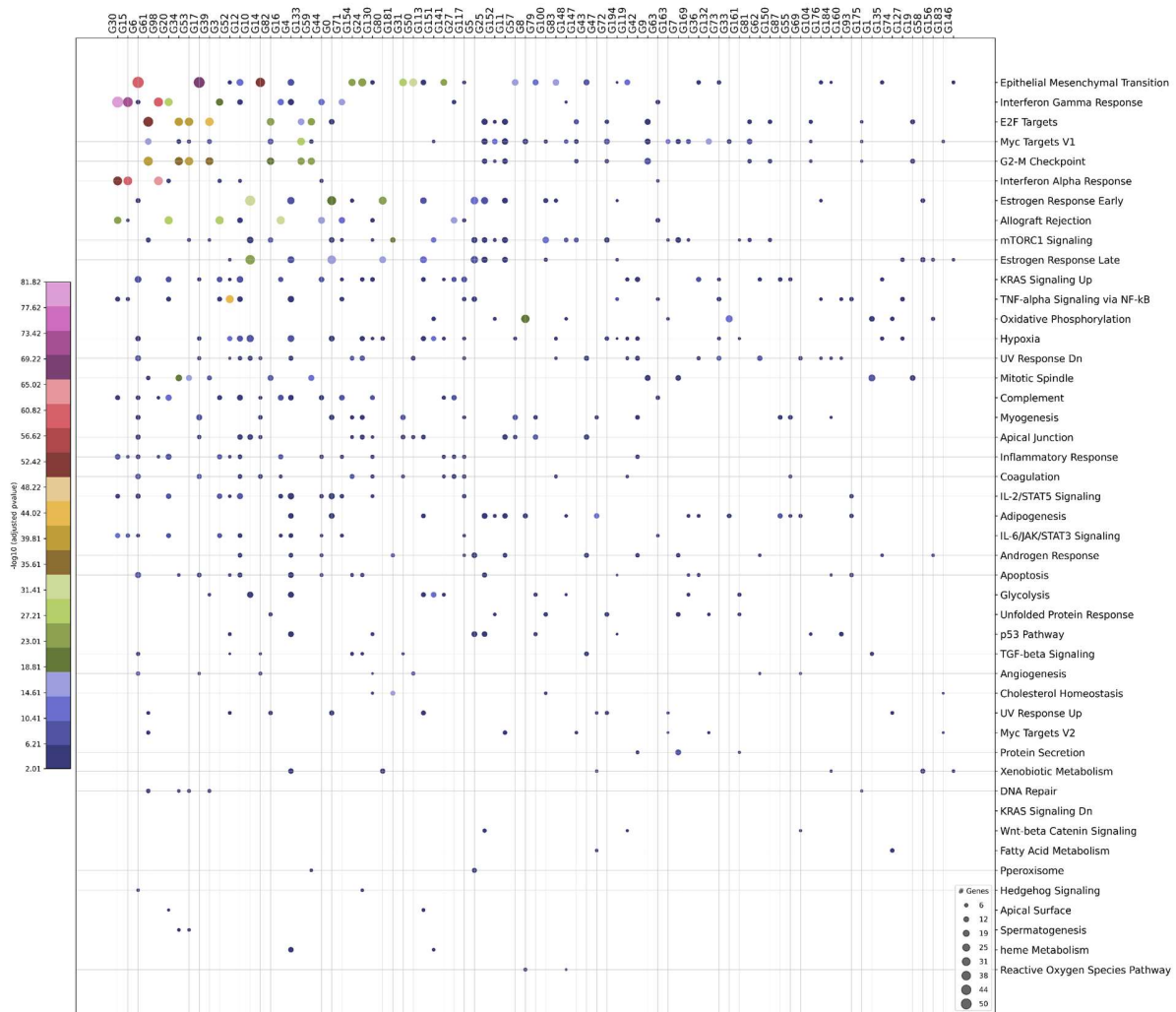
815

816

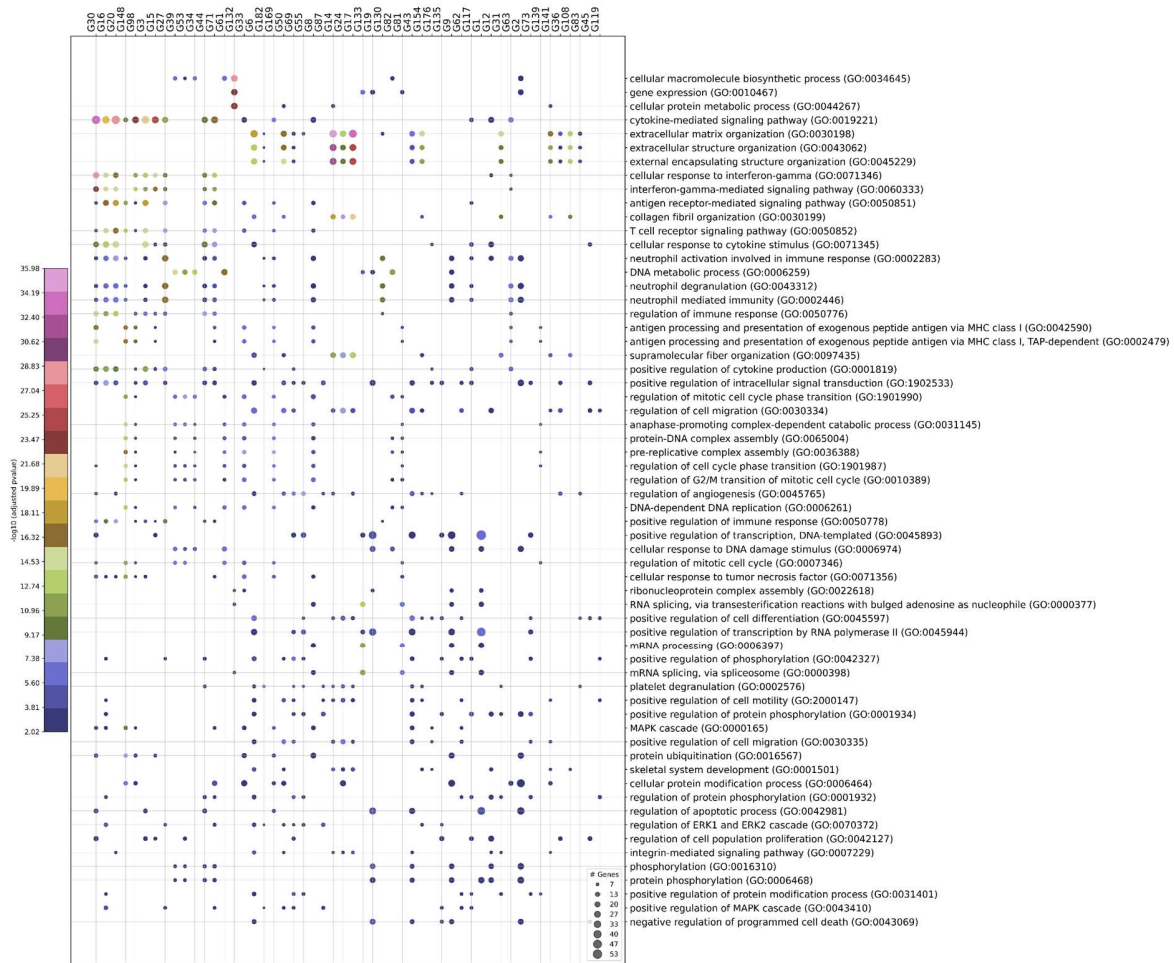
817



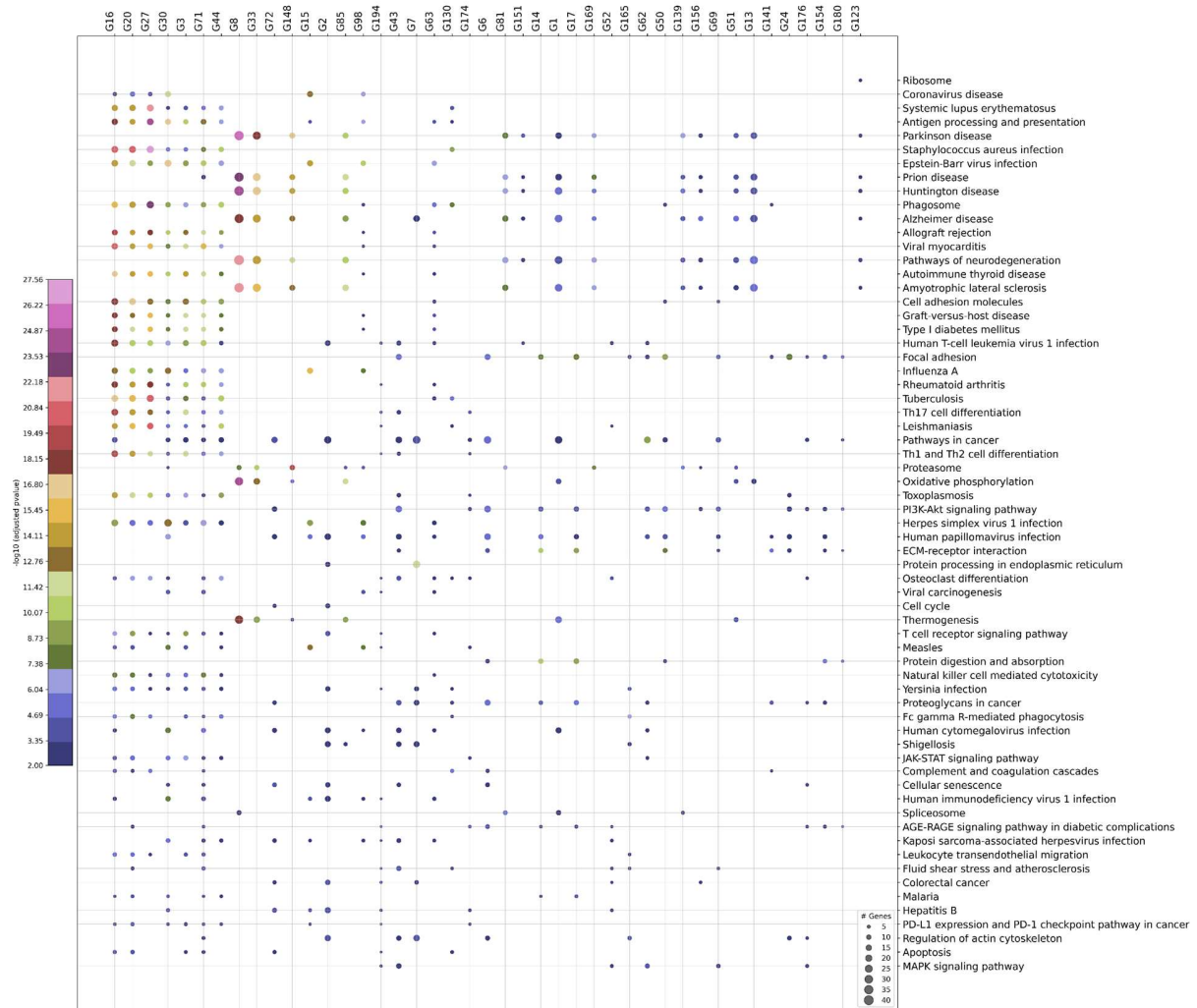
Sfig 1: Word clouds demonstrating the gene composition of different gene groups. The color of the gene indicates whether its median expression across patients is high (red) or low (blue) when gene group status = 1. The font size of gene within a group is proportional to the amount of information that the gene status provides about a particular gene.



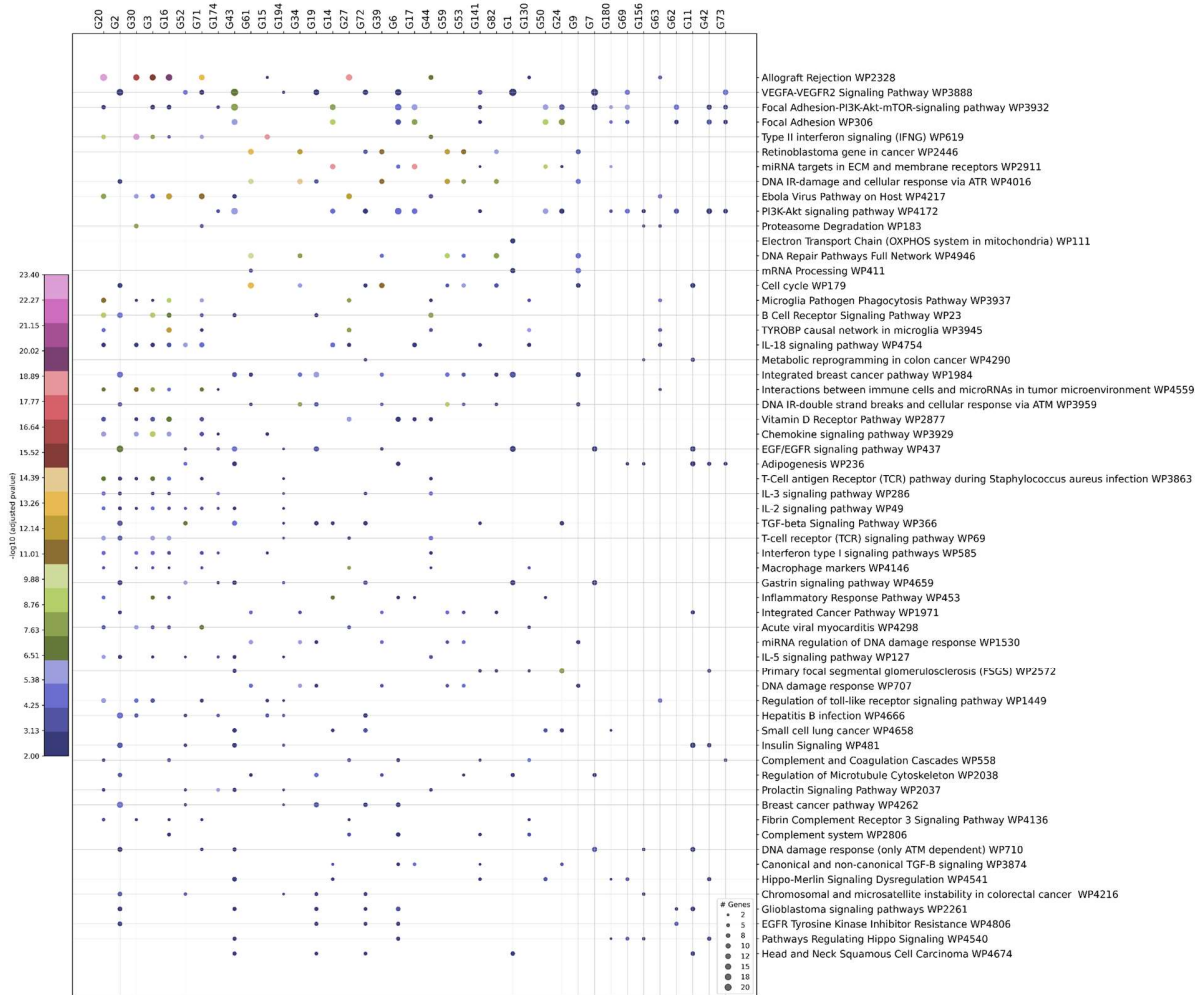
Sfig 2: Enrichment of gene groups for cancer Hallmark processes is illustrated as 2D scatter plot with the gene group displayed along x-axis and the corresponding enriched biological pathways on y-axis. The size of the dot represents the number of genes from a specific gene group that has shown enrichment for a particular hallmark process while its color represents the statistical significance of association in terms of FDR-corrected p-value.



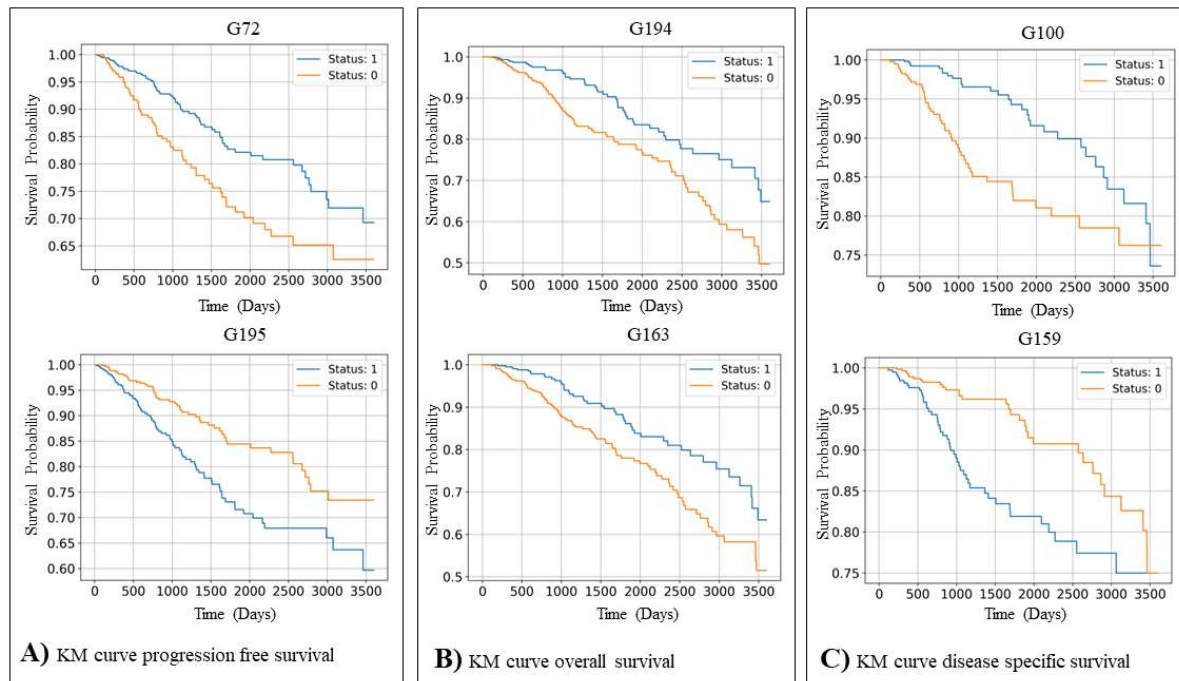
Sfig 3: Enrichment of gene groups for GO (Gene ontology) biological processes is shown as 2D scatter plot with the gene groups displayed along x-axis and the corresponding enriched biological processes on y-axis. The size of scatter dot represents the number of genes from a specific gene group that has shown enrichment for a particular biological process while its color represents the statistical significance of the association in terms of FDR-corrected p-value.



Sfig 4: Enrichment of gene groups for KEGG Pathways is presented as 2D scatter plot with the gene group displayed along x-axis and the corresponding enriched biological pathways on y-axis. The size of the dot represents the number of genes from a specific gene group that has shown enrichment for a particular biological pathway while its color represents the statistical significance of association in terms of FDR-corrected p-value.



Sfig 5: Enrichment of gene groups for cancer WikiPathways is illustrated as 2D scatter plot with the gene group displayed along x-axis and the corresponding enriched pathways on y-axis. The size of the dot represents the number of genes from a specific gene group that has shown enrichment for a particular pathway while its color represents the statistical significance of association in terms of FDR-corrected p-value.



SFig 6: Kaplan-Meier (KM) survival curves of progression-free survival (PFI), overall survival (OS), and disease-specific survival (DSS) of patients stratified based on gene group statuses. A) KM survival curve of PFI of patients based on G72 and G195 status showing that patients can be stratified into high and low risk groups based on G72 and G195 statuses with a significant p-value (log-rank test FDR-corrected p-value < 0.05 as shown in KM survival curve. B) KM overall survival curve of gene groups G194 and G163 are shown. Overall, we found that the binary status of 3 gene groups (G72, G194 and G163) can stratify patients into high and low risk groups with a significant p-value (log-rank test FDR-corrected p-value < 0.05). C) KM-curve of 2 (out of 25) gene groups that shows statistically significant association (log-rank test FDR-corrected p-value < 0.05) with disease-specific survival are shown. Other gene groups that show statistically significant association with DSS are (G72, G163, G194, G80, G123, G82, G10, G76, G165, G156, G120, G53, G142, G25, G144, G61, G113, G150, G151, G175 and G189).

823

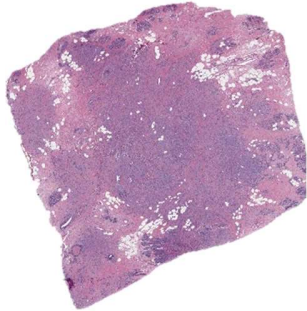
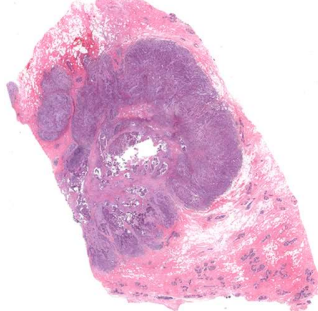
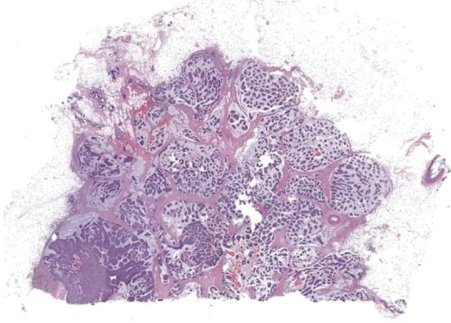
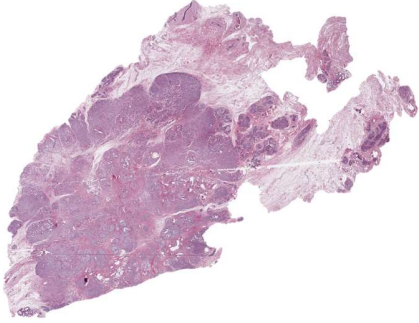
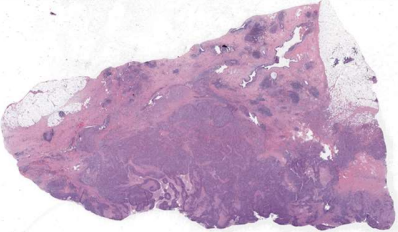
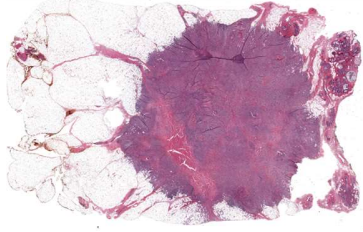
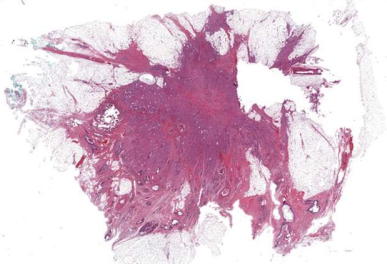
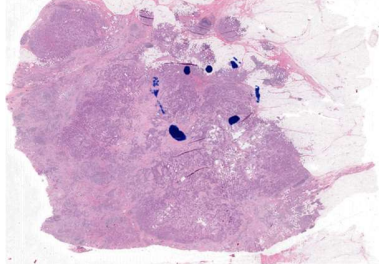
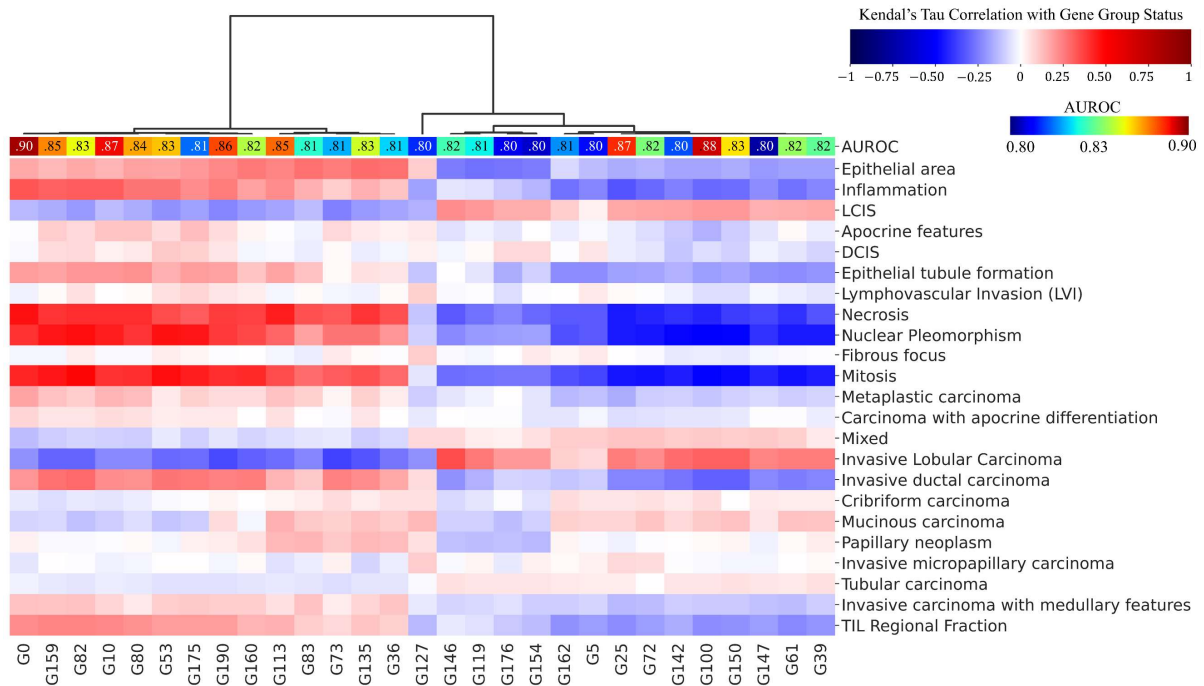
Best Predicted		Worst Predicted	
Thumbnail	S_c	Thumbnails	S_c
	0.61		0.00054
	0.58		-0.0087
	0.56		-0.0098
	0.55		-0.0082

Table 1: Thumbnails of patient WSIs whose gene expression states are best or poorly predicted from histology images using cosine similarity (S_c) between ground truth and predicted gene expression states as performance metric.

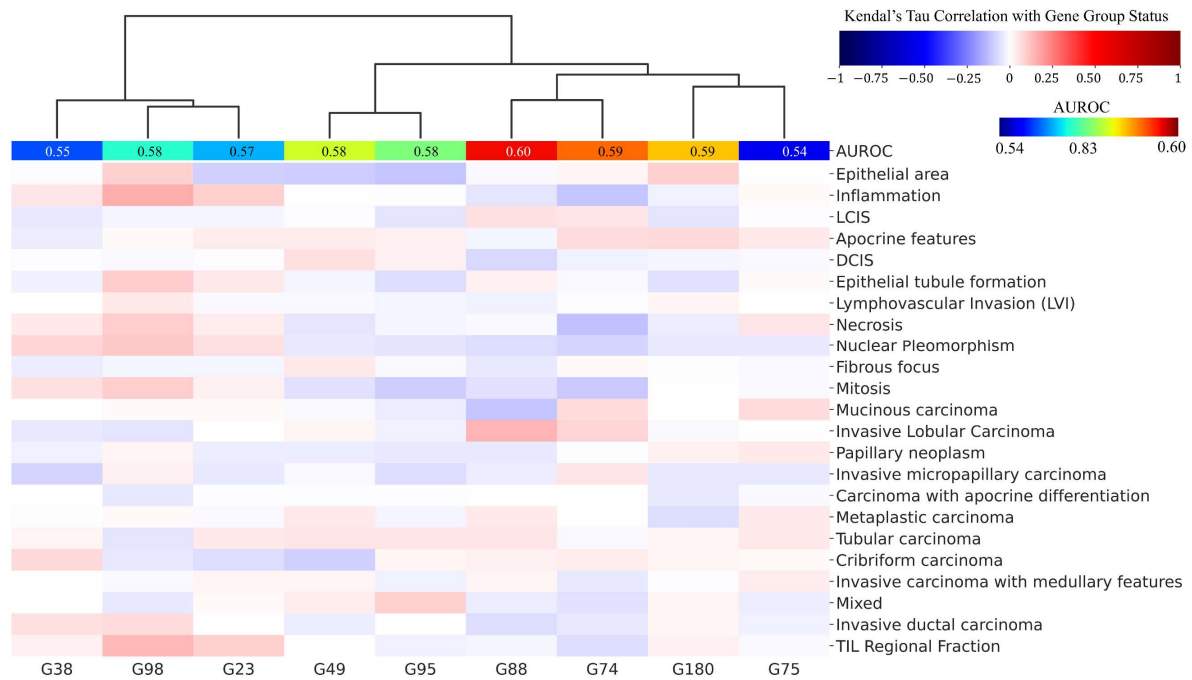
824

825

826



A: Gene groups predicted with high accuracy (AUROC \geq 0.80).



B: Gene groups predicted with poor accuracy (AUROC \leq 0.60).

SFig 7: Association of binary statuses of best and worst predicted gene groups with pathologist-assigned histological phenotypes. The plot uses two color bands one for AUROC and one for Kendall's Tau correlation. The AUROC is illustrated using the jet colormap representing the prediction accuracy of gene group binary status from imaging, while Kendall's Tau correlation between gene group binary status and various histological phenotypes is shown using the seismic colormap. We also annotated the AUROC colormap with the numeric value representing the mean AUROC value across test folds.

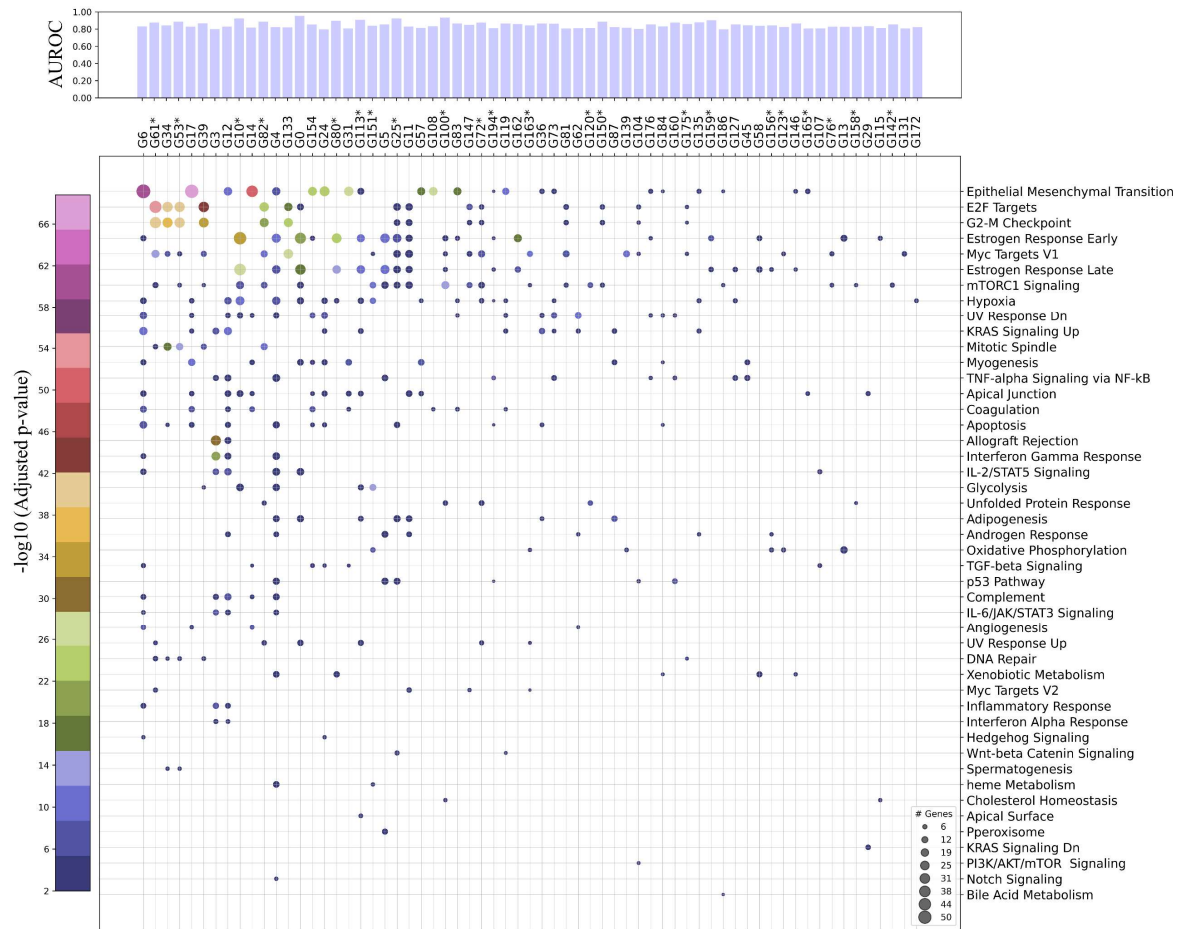
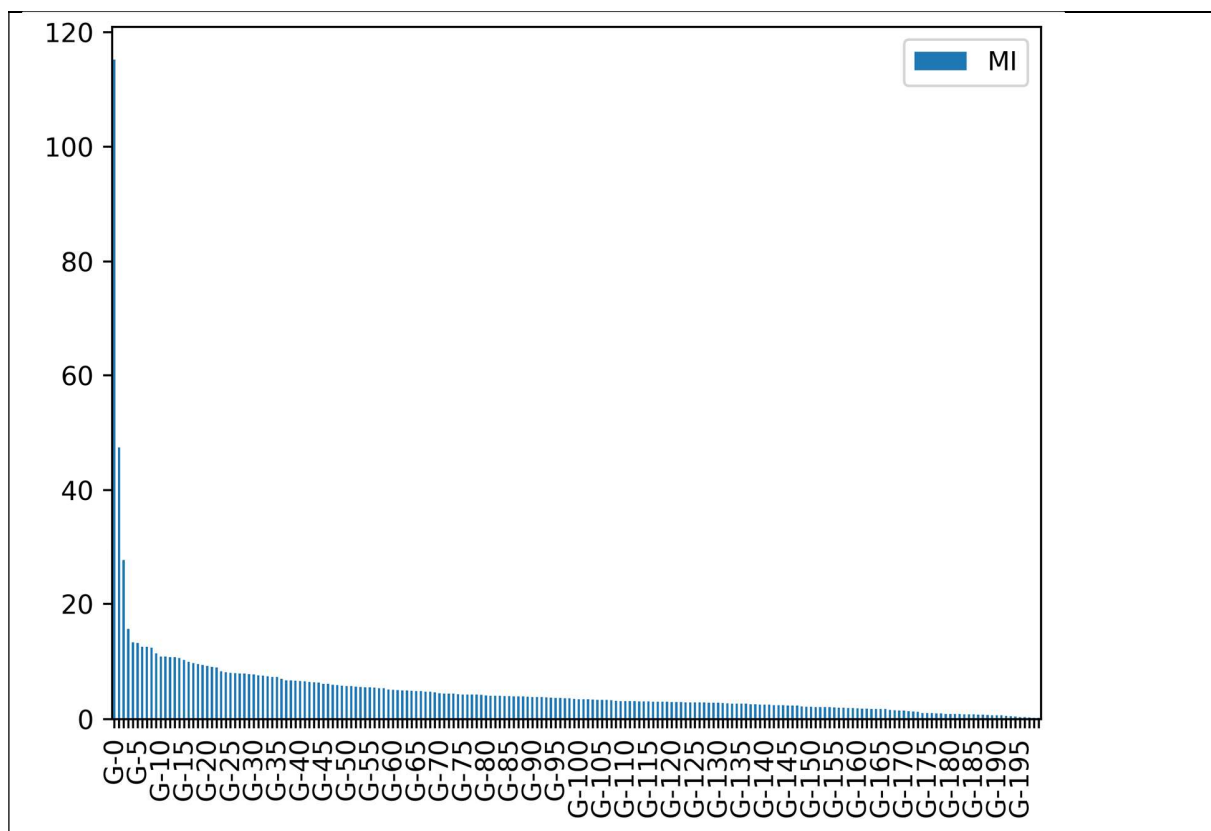


Fig 8: Association of best-predicted gene groups (AUROC ≥ 0.75) with cancer Hallmark processes and disease-specific survival. An example 2D scatter plot showing gene groups (one per column) with hallmark processes (one per row). The size of the scatter dot shows the number of genes in a gene group that has shown statistically significant association (FDR adjusted p-value < 0.01) with a certain biological pathway. In the plot, the p-value is represented by the color of the scatter dots. The top bar plot shows the prediction accuracy (AUROC) at which the binary statuses of these gene groups are predicted from histology images. Furthermore, gene groups that show statistically significant association with disease-specific survival are annotated with a * next to the gene group name.



SFig 9: Plot showing the proportion of total correlation explained by each latent factor.

828

829