bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Identification of phenotype-specific networks from paired gene expression-cell shape imaging data

Charlie George Barker[1], Eirini Petsalaki[1], Girolamo Giudice[1], Julia Sero[2], Emmanuel Nsa Ekpenyong[1], Chris Bakal[3], Evangelia Petsalaki[1*]

[1]European Molecular Biology Laboratory-European Bioinformatics Institute, Hinxton CB10 1SD, UK

[2]University of Bath, Claverton Down, Bath BA2 7AY, UK

[3]Institute of Cancer Research, 237 Fulham Road, London, SW3 6JB, UK

[*]Lead Contact/Correspondence: Evangelia Petsalaki, petsalaki@ebi.ac.uk

## Abstract

The morphology of breast cancer cells is often used as an indicator of tumour severity and prognosis. Additionally, morphology can be used to identify more fine-grained, molecular developments within a cancer cell, such as transcriptomic changes and signaling pathway activity. Delineating the interface between morphology and signaling is important to understand the mechanical cues that a cell processes in order to undergo epithelial-to-mesenchymal transition and consequently metastasize. However, the exact regulatory systems that define these changes remain poorly characterised. In this study, we employ a network-systems approach to integrate imaging data and RNA-seq expression data. Our workflow allows the discovery of unbiased and context-specific gene expression signatures and cell signaling subnetworks relevant to the regulation of cell shape, rather than focusing on the identification of previously known, but not always representative, pathways. By constructing a cell-shape signaling network from shape-correlated gene expression modules and their upstream regulators, we found central roles for developmental pathways such as Wnt and Notch as well as evidence for the fine control of NFkB signaling by numerous kinase and transcriptional regulators. Further analysis of our network implicates a gene expression module enriched in the Rap1 signaling pathway as a mediator between the sensing of mechanical stimuli and regulation of NFkB activity, with specific relevance to cell shape in breast cancer.

## Introduction

The study of cancer has long been associated with changes in cell shape as morphology can be a reliable way to subtype cancer and predict patient prognosis [1]. Recent research has implicated cellular morphology in more than just a prognostic role in cancer, with shape affecting tumour progression through the modulation of migration, invasion and overall tissue structure [2][3]. The unique mechanical properties of the tumour tissue (primarily driven by changes in cell shape and the extracellular matrix) are hypothesised to contribute to the 'stem cell niche' of cancer cells that enables them to self-renew as they do in embryonic development [4]. Cell morphology and tumour organisation have been found to be a factor in modulating the intra-cellular signaling state through pathways able to integrate mechanical stimuli from the extracellular environment [5][6][7][8]. The discovery of mechanosensitive pathways in various tissues has revealed a complex interplay between cell morphology and signaling [9]. Further studies have revealed that cell morphology can also be a predictor of tumorigenic and metastatic potential as certain nuclear and cytoplasmic features enhance cell motility and spread to secondary sites [1], aided by the Epithelial to Mesenchymal Transition (EMT). This process is the conversion of epithelial cells to a mesenchymal phenotype, which contributes to metastasis in cancer and worse prognosis in patients [10].

Breast cancer is the most common cancer among women, and in most cases treatable with a survival rate of 99% among patients with a locally contained tumour. However, among those patients presenting with a metastatic tumour this rate drops to 27% [11]. During the development of breast cancer tumours,

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cells undergo progressive transcriptional and morphological changes that can ultimately lead towards EMT and subsequent metastasis [12][13][1]. Breast cancer subtypes of distinct shapes show differing capacities to undergo this transition. For example, long and protrusive basal breast cancer cell lines are more susceptible to EMT [14] with fewer cell-to-cell contacts [15]. Luminal tumour subtypes on the other hand, are associated with good to intermediate outcomes for patients [15] and have a clear epithelial (or 'cobblestone') morphology with increased cell-cell contacts [16]. It is evident that cell morphology plays significant roles in breast cancer and a deeper understanding of the underlying mechanisms may offer possibilities for employing these morphology determinant pathways as potential therapeutic targets and predictors of prognosis.

Signaling and transcriptomic programs are known to be modulated by external cues in the contexts of embryonic development [17], stem-cell maintenance [18][19] and angiogenesis [20]. Numerous studies have flagged NFkB as a focal point for mechano-transductive pathways in various contents [21][22][23][24], but gaps in our knowledge remain as to how these pathways may interact and affect breast cancer development. Sero and colleagues studied the link between cell shape in breast cancer and NFkB activation by combining high-throughput image analysis of breast cancer cell lines with network modelling [25]. They found a relationship between cell shape, mechanical stimuli and cellular responses to NFkB and hypothesised that this generated a negative feedback loop, where a mesenchymal-related morphology enables a cell to become more susceptible to EMT, thus reinforcing their metastatic fate. This analysis was extended by [26], who combined cell shape features collected from image analysis with microarray expression data for breast cancer cell lines to create a shape-gene interaction network that better delineated the nature of NFkB regulation by cell shape in breast cancer. This approach was limited as it only correlates single genes with cell shapes, thus relying on the assumption that a gene's expression is always a useful indicator of its activity [27]. Furthermore, the authors rely on a list of pre-selected transcription factors of interest and as such the approach is not completely data-driven and hypothesis free. Given our knowledge of the multitude of complex interacting signaling pathways in development and other contexts, it is safe to assume that there are many more players in the regulation of cancer cell morphology that have yet to be delineated [28][29][30][31]. Furthermore, how exactly extracellular mechanical cues are 'sensed' by the cell and passed on to NFkB in breast cancer is not clearly understood. From this it is clear that an unbiased approach is needed to identify novel roles for proteins in the interaction between cell shape and signaling.

Here we identify a data-derived cellular signaling network, specific to the regulation of cell shape beyond NFkB, by considering functional co-expression modules and cell signaling processes rather than individual genes. To this end, we have developed a powerful network-based approach to bridge the gap between widely available and cheap expression data, signaling events and large-scale biological phenotypes such as cell shape (Figure 1A). By organizing expression data into context-specific modules we leverage the transcriptome's propensity to be regulated in regulons, thereby aiding the inference of signaling activity. This, combined with the use of feature-correlated modules, allows for the analysis of complex phenotypes, defined by multiple features. The resulting signaling network was validated using data from cell treatments with kinase inhibitors [32] and revealed further regulation of NFkB's utility in modulating mechanotransduction in response to morphological changes in breast cancer.

# Results

## Identification of gene co-expression modules correlated with cell shape features

We first sought to identify gene expression modules (GEMs) that are relevant to the regulation of cell shape. To this end, we used Weighted Gene Correlation Network Analysis (WGCNA) [33] on bulk RNA-seq expression data from 13 breast cancer and one non-tumorigenic epithelial breast cell lines to identify gene co-expression modules correlated with 10 specific cell shape variables [26] (Methods). These described the size, perimeter and texture of the cell and the nucleus (n = 75,653). Of 102 GEMs (Figure S1A), 34 were significantly correlated ($P<0.05$; Student's T-test, Pearson Correlation) with one of 8 cell shape features (Figure 1B). Full list of the genes within the identified modules are summarized in Supplementary Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

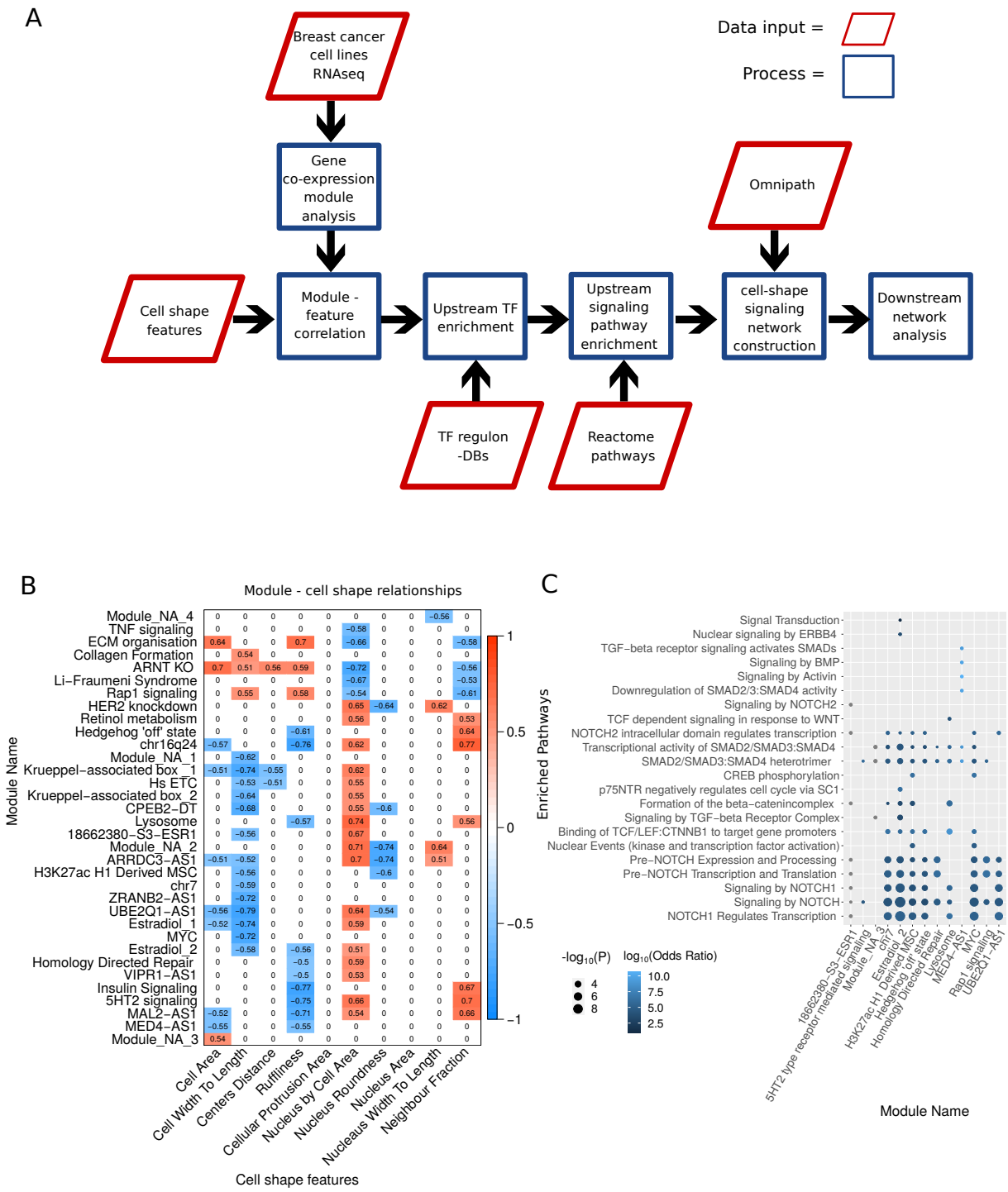



**Figure 1: A.** Schematic illustrating the steps involved in phenotype-specific network construction. Gene expression modules are identified by integrating cell shape variables (derived from imaging data) with RNAseq data from breast cancer cell lines. These gene expression modules are correlated with specific cell shape features to find morphologically relevant modules. Next, transcription factors (TFs) are identified whose targets significantly overlap with the contents of the expression modules. These TFs are used to identify pathways regulating the gene expression modules, which are then integrated to form a contiguous network using PCSF. **B.** Heatmap of significantly correlated gene expression module eigengenes with cell shape features. Non-significant interactions were set to 0 for clarity. **C.** Dot plot illustrating the enrichment of pathways among TFs found to regulate gene expression modules. The x axis shows the module names (as defined by Supplementary Table 2) and the y axis shows the signaling pathways found to be significantly (P<0.01) enriched in the TFs that regulate the given module (as defined by Supplementary Table 3). The y axis is arranged such that the terms with the highest combined odds ratio are at the bottom. Size of the dot represents the -log10(P) and the colour indicates a log 10 transformation of the odds ratio.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We used EnrichR and their suite of gene set libraries [34] to functionally annotate and label some of the modules using enrichment of genes contained within them. We found that the 'Rap1 signaling' module is also enriched for terms such as VEGF signaling and hemostasis, while the 'Insulin Signaling' module is also enriched for cell-cell communication and the 'ECM organisation' module is also enriched in terms such as axon guidance and EPH-Ephrin signaling (Supplementary Table 2). Modules that are most correlated with all features are the 'ARNT KO' module, 'ARRDC3AS1' module and the 'ECM organisation' module (see Figure S1B). Modules that could not be annotated with informative terms were designated 'module non-annotated (NA) 1, 2, 3 etc.

## Transcription factor analysis of cell shape gene co-expression modules reveals the signaling pathways that regulate them

To link these expression modules to the intra-cellular signaling network, we considered both the regulation of modules as transcriptional units as well as the signaling pathways that significantly regulate the identified regulons. Specifically, we first found 17 transcription factor (TF) regulons, as defined in the database TRRUST v2 [35], to be significantly enriched (P<0.1; Fisher's exact test) in our modules (Supplementary Table 3). We therefore consider these TFs as potentially relevant for the regulation of cell shape features and their activity levels as a read-out of cell signaling activity in these cells. These TFs include the EMT antagonist FOXA1 [36], and HOXB7 [37] and ZFP36 [38].

To extend this further, we sought to investigate the pathways responsible for regulating the identified TFs, and by extension the gene expression modules. For this analysis, we also include ENCODE and ChEA Consensus TFs from ChIP-X [39],DNA binding preferences from JASPAR[40][41], TF protein-protein interactions and TFs from ENCODE ChiP-Seq [42] to get a more comprehensive picture of the pathways involved in regulation of cell morphology. Using the identified TFs (Supplementary Table 4) we then used EnrichR [34] to perform a Reactome signaling pathway [43] enrichment analysis. Results from this analysis showed that many identified gene expression modules were regulated by common signaling pathways, with 6 modules sharing pathways associated with downstream signaling and regulation of Notch (Figure 1C).

## Clustering based on morphology reveals distinctive cell-line shapes

To understand key differences in expression patterns and gene regulation between morphologically distinct breast cancer cell lines, we clustered them based on 8 morphological features including area, ruffliness, protrusion area and neighbour frequency and performed differential expression analysis between the identified clusters (Figure 2A). Cluster A is more heterogeneous in its morphology, containing the non-tumorigenic mammary epithelial cell line MCF-10A as well as cell lines from both luminal and basal breast cancer subtypes. Clusters B and C are more distinctly shaped, roughly composed of luminal and basal cell lines respectively with the exception of HCC1954, which was clustered morphologically with luminal subtypes while being characterised as basal. The basal-like cluster is most morphologically distinct from cluster A, but also differs from the luminal-like cluster in that it has a lower nuclear/cytoplasmic area (0.133 ± 0.05 [mean ± SD]), higher ruffliness (0.235 ± 0.12) and lower neighbour fraction (0.258 ± 0.22). The luminal-like cluster had a higher nuclear/cytoplasmic area (0.186 ± 0.1; P<0.001), lower ruffliness (0.213 ± 0.14; P<0.001), and a higher neighbour fraction (0.338 ± 0.26; P<0.001, One-way ANOVA; Tukey HS, n = 75,653). The neighbour fraction feature corresponds to the fraction of the cell membrane that is in contact with neighbouring cells. The lower number of cell-cell contacts in basal-like breast cancer cell lines are indicative of more mesenchymal features associated with worse prognosis due to metastasis. Increased cell-cell contacts in both the luminal-like cluster and the more heterogeneous cluster A correspond to 'cobblestone' epithelial morphology. Interestingly, these groups are closely aligned with the expression of the cell adhesion protein, N-cadherin (Figure 2A), the expression for which is closely associated with a migratory and metastatic phenotype [44]. Representative images of the morphologically clustered cell lines are shown along with the clustering heatmap in Figure 2A (complete set of images provided as Supplementary Data).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

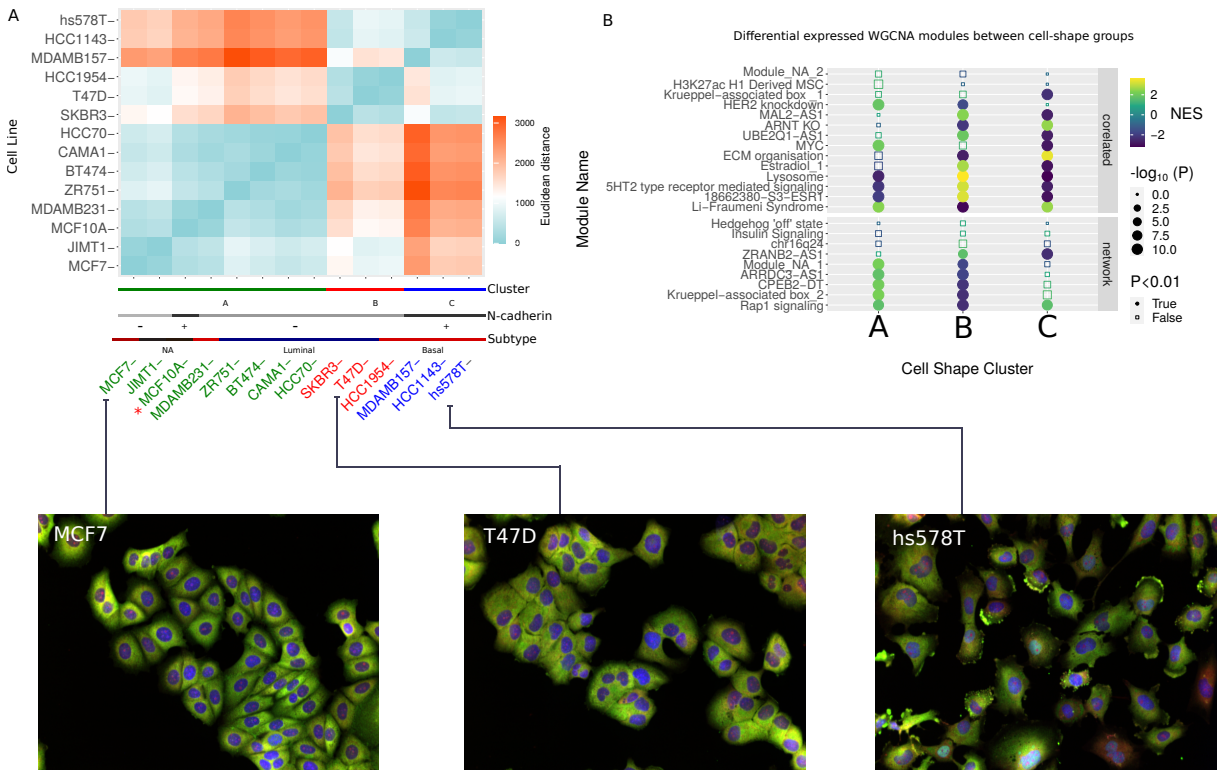



**Figure 2: A.** Heatmap of Euclidean distance between cell lines for shape features to illustrate clusters arising from k-means method. The coloured lines on the bottom show the assigned cluster and the cadherin expression and assigned canonical cancer subtype. **B**. Dotplot showing the enrichment of gene expression modules in the different cell line clusters. Along the y axis are the names of the clusters, faceted by whether they are included in the PCSF derived regulatory network on the bottom and whether they are correlated with cell shape variables, but not included in the network on the top. The x axis shows the cell shape clusters, with letters corresponding to the groups in Figure 2A (A - a heterogeneous mix of breast cancer subtypes, B - luminal-like cell lines and C - basal-like cell lines.). Dots are coloured based on the normalised enrichment value, with down-regulated modules in blue, and the up-regulated modules in yellow. Size corresponds to significance (-log10(P)) with the shape illustrating which changes are significant (adjusted P<0.01, Benjamini-Hochberg). **C.** Images (see Methods) showing morphology of representative cell lines from each respective cluster.

Using the identified groups of cell lines in the previous step, differential expression analysis and transcription factor activity analysis was used to study gene regulation signatures specific to cell line morphological clusters. The results are shown in Supplementary Table 5, with gene set enrichment analysis showing upregulation of genes involved in the extracellular matrix, collagens, integrins and angiogenesis in the basal-like cluster. Significantly enriched terms (P<0.05) in downregulated genes include 'fatty acid and beta-oxidation' and 'ERBB network pathway'. In the genes upregulated in the luminal-like cluster, we observed enrichment of terms such as 'hallmark-oxidative phosphorylation'. Downregulated genes were enriched in 'integrin-1 pathway', 'core matrisome' and genes linked to 'hallmark epithelial-mesenchymal transition and migration'. For the remaining B/L group, the term with the highest normalized enrichment score was 'targets of the transcription factor Myc' followed by terms associated with ribosomal RNA processing. Down-regulated terms include 'cadherin signaling pathway' (Supplementary Table 5).

We also calculated the differential expression for the WGCNA gene expression modules and found distinct patterns of expression between luminal-like and basal-like clusters of cell lines (Figure 2B). Among these, the Rap1 Signaling module is upregulated in basal-like clusters and downregulated in luminal-like clusters. This is consistent with the fact that this gene expression module is negatively correlated with neighbour fraction, a feature that is observed to decrease in mesenchymal-like cell shapes [15]. Other modules whose expression distinguishes basal-like from luminal-like include the MAL2-AS1 module (enriched in desmosome assembly), ARNT/KO module (enriched in TNF-signaling by NFkB) and ECM organisation module (enriched in focal adhesion proteins - see Supplementary Table 2).

To link the observed gene expression differences to cell signaling we used the tool DOROTHEA [45] to

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

calculate transcription factor activities, as their modulation is one of the main results of cell signaling processes. We corroborated that the heterogenous B/L group had significantly activated Myc levels. In the luminal-like cluster, ESRRA (estrogen-related receptor $\alpha$) is the most significantly overrepresented regulome, followed by EHF, KLF5 and ZEB2. Under-represented regulomes include KLF4, SMAD4, SMAD2, SOX2 and RUNX2. For the basal-like cluster, the regulome with the highest normalised enrichment score is SOX2, as well as Musculin and HOXA9. Downregulated regulomes include ZEB2, Myc, ESRRA and KLF5 (Figure S2D).

## Assembly of a data-driven cell shape regulatory network

In order to integrate our data-driven GEMs with signaling pathways, we used the Prize Collecting Steiner Forest (PCSF) algorithm [46]. This is an approach that aims to maximise the collection of prizes associated with inclusion of relevant nodes, while minimizing the costs associated with edge-weights in a network. This allowed for the integration of the WGCNA modules, the Reactome pathways that regulate them, the TRRUST transcription factors and the differentially expressed DOROTHEA regulons into a contiguous regulatory network describing the interplay between cell shape and breast cancer signaling. The network used for this process was extracted from the database OmniPath [47] to provide a map of the intracellular signaling network described as a signed and directed graph. We incorporated identified GEMs into the network by interlinking them with the relevant TFs and signalling pathways.

The resulting network of 691 nodes included 97.11% of the genes identified by our analysis (Figure 3A). The new proteins that were included by the PCSF algorithm to maximise prize collection showed gene set enrichment of common terms (Pathways from Panther [48]) relative to the original prizes (including Wnt, EGF, Angiogenesis, Ras, Cadherin and TGF-$\beta$ pathways), but also included are some new terms (VEGF, Integrin and Endothelin pathways) (P<0.001; Figure 3B).

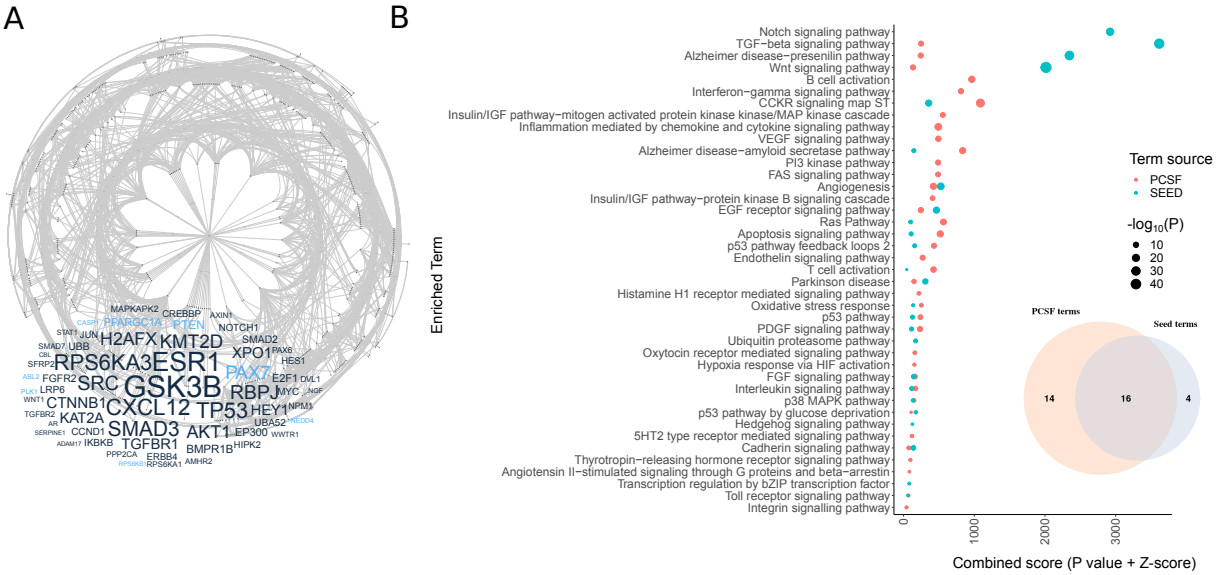



**Figure 3: A.** Network derived from integrating enriched pathways and transcription factor regulons within cell shape correlated gene expression modules. Below is a word cloud illustrating the top 50 nodes with highest betweenness centrality in the network, with those nodes belonging to the original seeds coloured black, and those genes included by the PCSF step coloured light blue. **B.** Gene set enrichment (Panther DB 2016) of original seed nodes and the nodes included by PCSF with the dot size indicating the level of significance (-Log10(P)) of the term enrichment. Inset is a Venn diagram showing the overlap in enriched terms between the seed nodes and the terms included by PCSF.

Studying the network properties of our PCSF-derived regulatory network we find that the degree distribution is typical for a biological network (Figures S3A-C). The proteins in the network can be ranked by betweenness centrality to disseminate them based on network importance. Nodes with high centrality lie between many paths and can control information flow. Proteins with the highest centrality are primarily prizes (GSK3*b*, ESR1, p53, SMAD3 - Figure S3D) indicating that the PCSF solution was not achieved by the inclusion of new hub proteins that are not of interest to our analysis. That being said, a small minority of high centrality nodes were not in the original prizes, including PAX7, PTEN and PPARGC1a.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Studying the network properties of our PCSF-derived regulatory network we find that the degree distribution is typical for a biological network (Figures S3A-C). The proteins in the network can be ranked by betweenness centrality to disseminate them based on network importance. Nodes with high centrality lie between many paths and can control information flow. Proteins with the highest centrality are primarily prizes (GSK3b, ESR1, p53, SMAD3 - Figure S3D) indicating that the PCSF solution was not achieved by the inclusion of new hub proteins that are not of interest to our analysis. That being said, a small minority of high centrality nodes were not in the original prizes, including PAX7, PTEN and PPARGC1a.

## Small-molecule inhibitors targeting kinases in our network significantly perturb cell morphology

To validate our network, we used an independent dataset to evaluate whether perturbing the function of kinases within our predicted network would produce a significant effect on morphological features. For this, we used the Broad Institute's Library of Integrated Network-based Cellular Signatures (LINCS) small molecule kinase inhibitor dataset [32]. Here, they measured morphological changes in the breast cancer cell line HS578T in response to various small molecule kinase inhibitors using high through-put imaging techniques [49]. The morphological variables measured in this data-set are mostly analogous to the ones used to construct the network, however there are some discrepancies which we used as negative controls to ensure our network was phenotype-specific.

We combined this with data from a target affinity assay [50] describing the binding affinities of small molecules to kinases. This enabled us to sort the kinase inhibitors into those that target proteins we predict regulate cell shape (through their inclusion in the PCSF derived network) and those that do not. Figure 4A illustrates that there is a statistically significant ($n = 37$, Wald test $P<0.05$) deviation from the control between drug treatments targeting kinases within the predicted network and those targeting other proteins for cytoplasmic area, cytoplasmic perimeter, nucleus area, nucleus length, nucleus width and nucleus perimeter. This difference is insignificant for features that were not correlated with gene expression modules in our initial analysis (such as number of small spots in the cytoplasm and nucleus, and nuclear compactness), indicating that our network is phenotype-specific to the features used in network generation. We also repeated this analysis in other cell lines (SKBR3, MCF7 and non-tumorigenic mammary cell line MCF10A) with results with limited statistical significance (S4A-C) suggesting a bias in our network towards certain breast cancer cell lines (see Discussion).

Interestingly, there is greater variance in the effect size for kinase inhibitors targeting proteins contained within the predicted regulatory network than those outside. We hypothesised that it was the network properties of kinases within our network that dictated their effect on morphological features, with some targets being on the periphery of our predicted network and therefore having limited influence over the regulation of cell shape. To test this, we studied the extent to which the effect of a kinase inhibitor was correlated with the combined centrality of its targets as defined by our network. For this we used the centrality algorithm PageRank (Brin and Page, 1998) and accounted for off-target effects of the kinase inhibitors using the Szymkiewicz-Simpson index (describing the overlap of a kinase inhibitor's targets and the proteins that constitute the network - Methods).

Figure 4B shows moderate correlations between target centrality and the effect size for each feature, illustrating that kinase inhibitors targeting proteins with high centrality in our network modulate cell shape more than inhibitors with peripheral targets. As with studying the effect of targeting kinases contained within our network versus those outside of it, this correlation is higher among morphological variables that are the same or similar to those cell shape features correlated with gene expression modules used to construct the network. The correlation between combined centrality and drug absolute effect on cell area ($n=37$) was moderate but significant for cytoplasm area, cytoplasm perimeter, nucleus area, nucleus length, nucleus half-width and nucleus perimeter (with Spearman correlation coefficient between 0.34 - 0.37 for all of them, with $P<0.05$). This correlation in change in morphological features with the centrality of the targeted kinases illustrates the relevance of our constructed network in regulating cell shape. For variables that were not correlated to any gene expression module, we see visibly lower correlation coefficients and insignificant associations (Spearman correlation coefficients of 0.05 - 0.29, $P>0.05$). These results illustrate that the topology of our network explains some of the variation in the effect of kinase inhibitors tested, in a manner that is feature specific to the ones that were used to construct the network model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

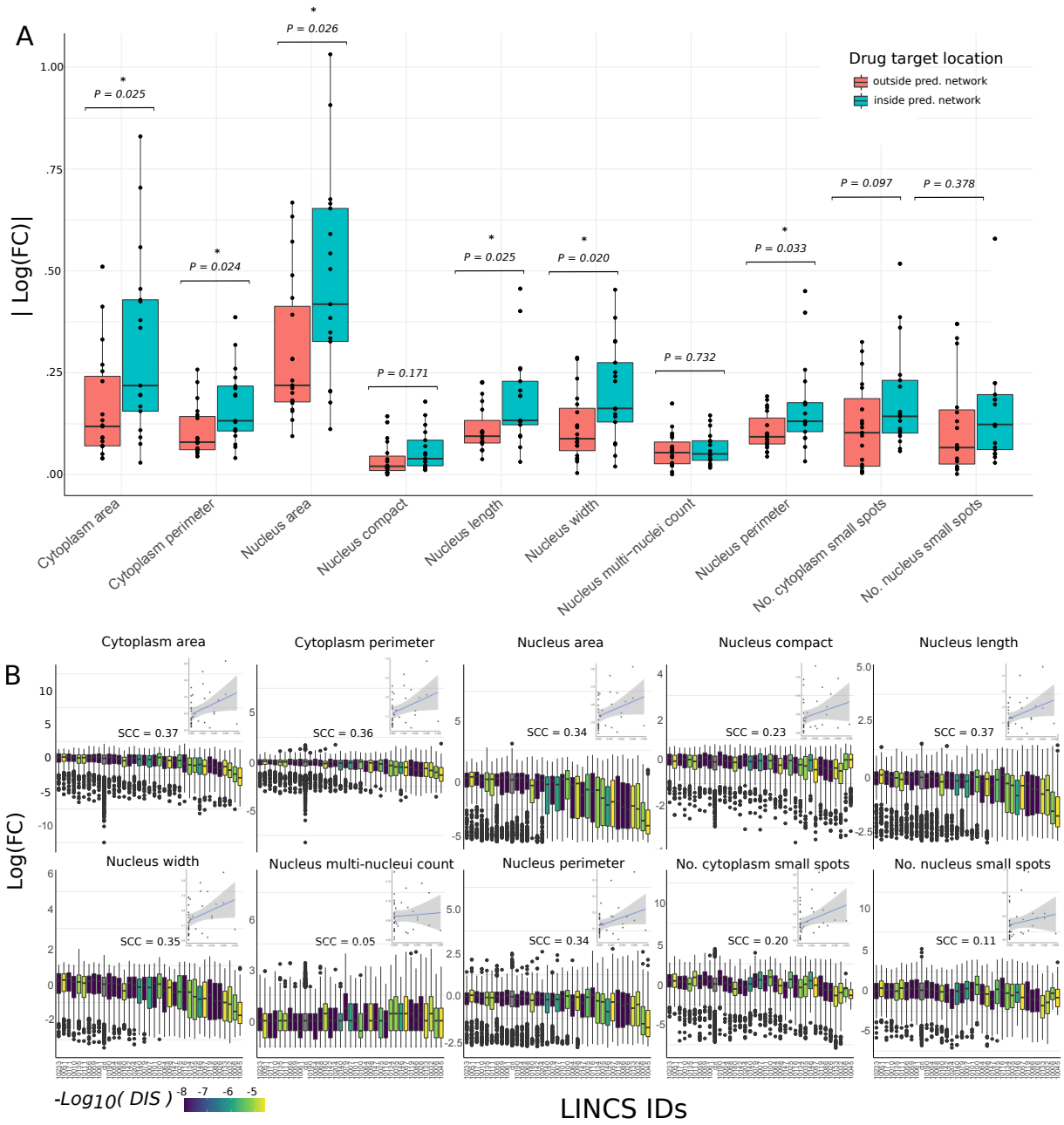



**Figure 4: A.** Box plots showing the absolute log(10) fold changes after treatment with a drug relative to a control for each cell shape variable. The drugs are grouped depending on whether they target kinases within the predicted regulatory network (blue) and those targeting other kinases not predicted to be associated with cell shape (red). P values (Welch Two Sample t-test) are showing with stars indicating significance. **B.** Bar plot showing the absolute difference in log fold changes of cell shape variables after treatment with a drug relative to a control. Here, each drug is shown separately (with the LINCs ID shown on the x axis) and coloured based on the drug influence score (DIS) and each data-point represents a single cell. Inset are plots showing the correlation between this influence score and the difference between mean treated cells and mean control cells in each of the 10 measured cell shape features for each drug. Spearman correlation coefficients are shown above the inset plots.

## Network propagation of activated TFs reveals differentially activated processes in the cell shape regulatory network

As transcription factor activity remains the most reliable indicator of signaling that can be extracted from transcriptomics data [51], we applied network propagation to identify regions of which differentially regulated transcription factors have an effect. The algorithm Random Walk with Restart (RWR [52]) was used to diffuse from activated and inactivated transcription factors in our network reflected by the normalized enrichment scores of transcription factors identified by DOROTHEA [45] (Figure 5A & B; Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 6; Methods).

The most relevant super-node in both luminal and basal diffusions was the gene expression module, Rap1 Signaling, a module which is correlated with several cell shape variables (neighbour frequency, ruffliness, nuclear by cytosolic area and cell width to length) and is enriched in members of the mechanosensitive Rap1 signaling pathway. By performing RWR diffusions on each of the seed nodes separately (Figure S5A-B) we can see that the source of this module's probability is from the transcription factors JARID2 and RUNX2 in luminal-like cell lines, and JARID2 for basal-like.

Specific proteins that were top ranked after performing the network propagation in basal-like cell lines include the orphan nuclear receptor NR0B2. Individual RWR found 3 seed transcription factors responsible for this node's high probability: AR, ESRRA and NR1H3. Other proteins flagged by the propagation were SMAD4, which is regulated by TGF-$\beta$, IKBKB, which is an activator of NFkB and YAP1. GO term enrichment of significantly ranked proteins in basal-like cell lines found significantly enriched terms such as Wnt signaling pathway, Angiogenesis related signaling and TGF-$\beta$ signaling pathway (P<0.01; Figure S5C). For luminal-like cell lines, NR0B2 is also significantly ranked from the network propagation (as a result of ESRRA activity) as well as transcriptional co-activator PPARGC1A. GO terms enriched in the luminal RWR include many shared with basal (TGF-b, Wnt, CCKR signaling and angiogenesis), but also include new terms such as HIF activation, p53 pathway and Toll receptor signaling pathway (P<0.01; see Figure S5D).

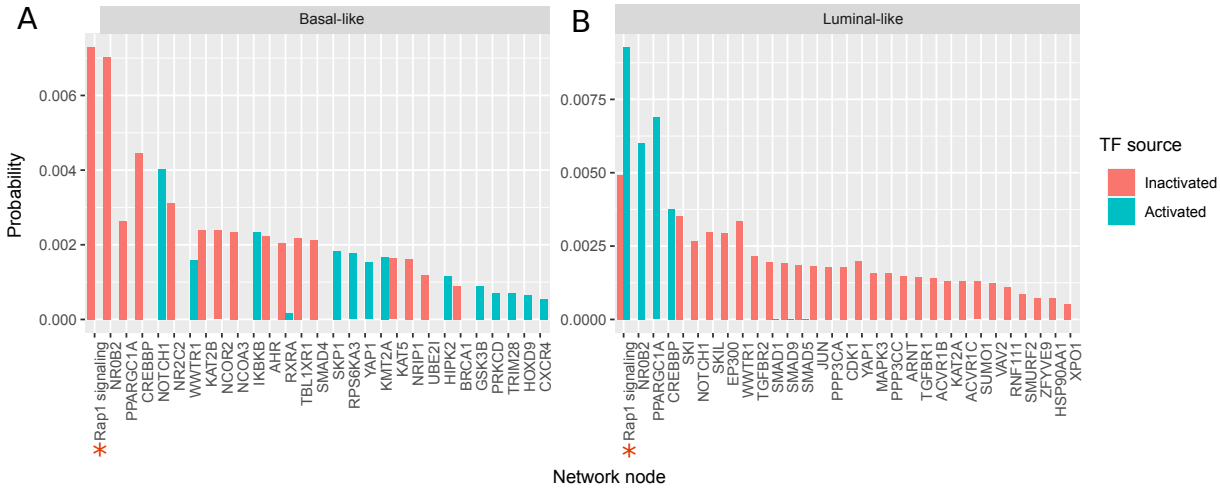



**Figure 5:** Bar plot showing network propagation in predicted cell shape network from activated and inactivated transcription factors in basal-like cell lines **(A)** and luminal **(B)**. The y axis is a steady state probability (or the 'heat' of the nodes in the network after the diffusion) over the graph imposed by the starting seeds, ordered by size. Red bars represent propagation from transcription factor seeds that are predicted to be in-activated, and blue bars show propagation from transcription factor seeds that are predicted to be activated. Red stars along the x axis indicate supernodes that represent gene expression modules. Only those nodes with combined probability > 0.0001 are shown, with the full results available in Supplementary Table 4.

## Rap1 gene expression module correlates with known morphologically-relevant TFs in both cell culture and clinical samples

To explore the significance of the Rap1 gene expression module in breast cancer we measured its activity (Methods) in 78 BRCA cell lines. This enabled the correlation of its combined activity with the activity of known transcription factors predicted by DOROTHEA [53] (Figure S6). We find that Rap1 GEM activity was significantly correlated (Kendall; P<0.01, FDR adjusted) with the activity of 19 TFs. Among these are RUNX2 (consistent with the results from our network propagation), TEAD1 (TF mediating the function of YAP1/TAZ) and NFkB. We also correlated transcription factor activity using the same method on tumour samples extracted from TCGA (*https://www.cancer.gov/tcga*). Using this publicly available dataset, we studied 1,090 BRCA tumours and performed differential expression on each sample. We found 40 TFs significantly correlated (Kendall; P<0.01, FDR adjusted) with Rap1 GEM (Figure 6A-B). The intersection of

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

this analysis between in cell lines and the clinical data were the TFs: SP3, NFKB1, ZNF589, ZC3H8, HIF1A, STAT1, ZNF584, ZNF175 and KLF5.

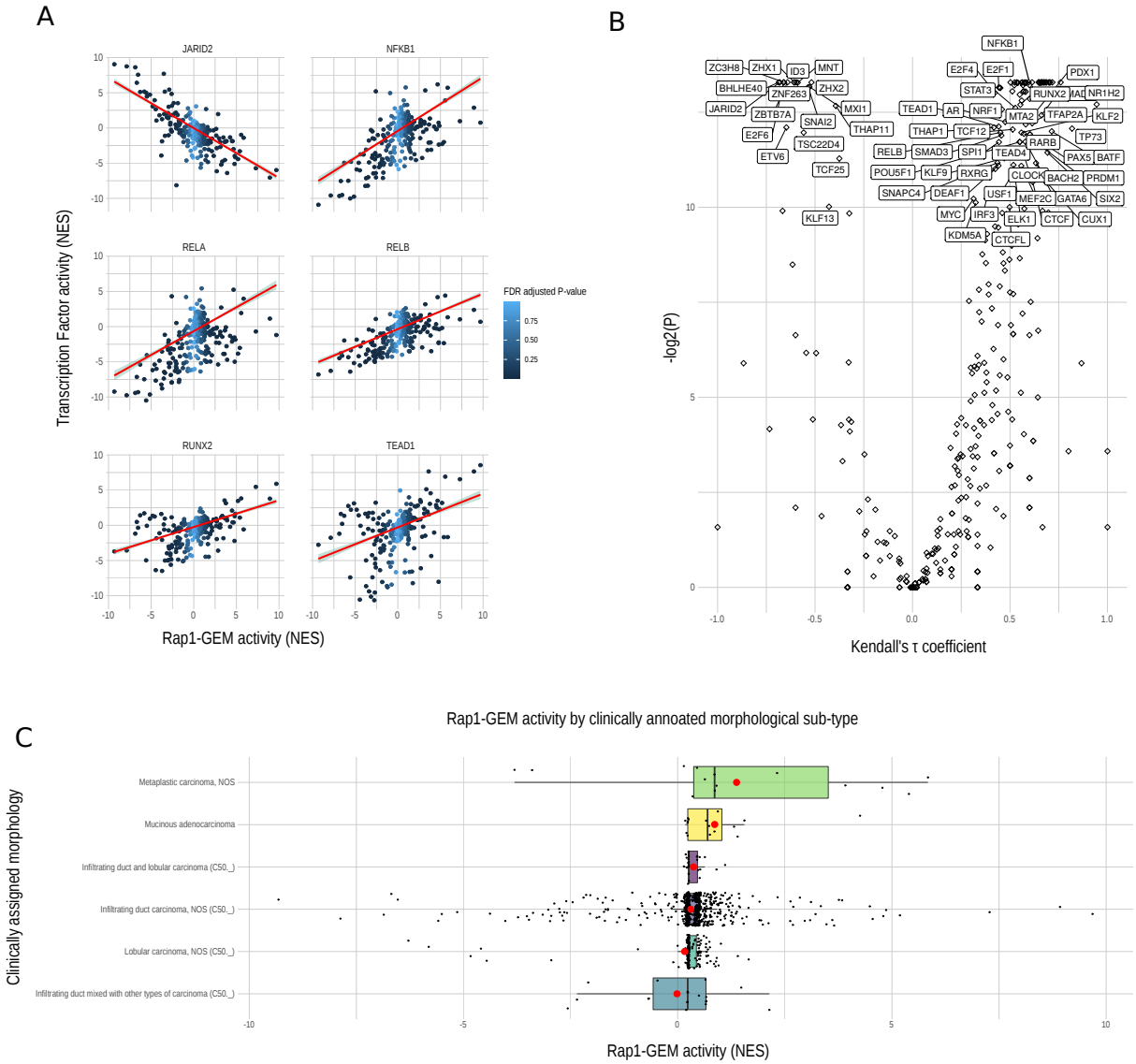



**Figure 6:** **A** Plots showing the correlation between the Rap1 gene expression module activity (Normalised enrichment score - see methods) and the activity (NES) of various transcription factor (JARID2, NFKB1, RELA, RELB, RUNX2 and TEAD1) . Line of best fit according to linear regression is shown in red, with confidence interval in grey. Colour of the points in the plot represents the FDR adjusted P-value of the Rap1 NES as calculated by DOROTHEA. **B.** Volcano plot illustrating the correlation (Kendall's rank correlation) between activity of Rap1 gene expression module and transcription factor activity, with Kendall's tau coefficient along the x axis and -log2(FDR adjusted P) along the y axis. **C.** Barplot showing Rap1-GEM activity across different breast cancer samples, separated by clinically assigned morphology. The y axis shows Rap1 gene expression module activity as calculated by DOROTHEA in NES. Mean values for each group are shown by a red dot.

We studied the expression of the module in tumour samples and compared different groups of clinically annotated morphological subtypes. The morphological subtype with the highest overall Rap1-GEM activity was metaplastic carcinoma, a subtype characterised by poorly cohesive sheets [54] and a high propensity to metastasize [55] (Figure 6C). This morphological subtype has a distribution significantly greater ($P < 0.005$; Two sample Kolmogorov-Smirnov test) than the most frequently assigned morphological subtype (Infiltrating duct carcinoma, NOS). This subtype is a common and homogenous breast cancer grouping characterised by its failure to exhibit morphological features that might allow it to be classified as anything more specific [56].

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

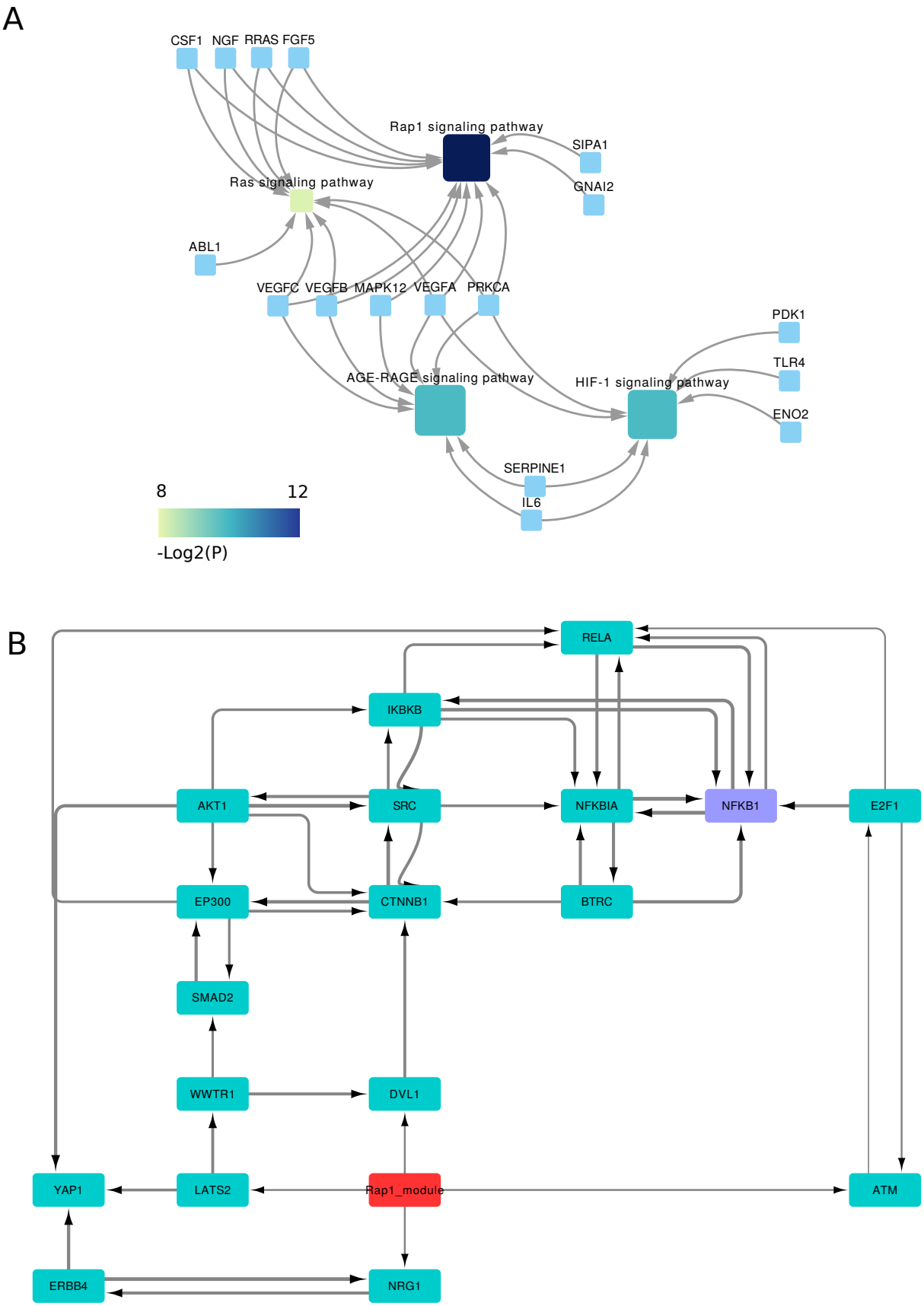



**Figure 7:** **A.** Network showing gene set enrichments of the contents of Rap1 gene expression module. Genes are shown in pale blue, and pathways are shown by nodes whose colour indicates significance of the associated term (-Log2(P)). **B.** Sub-network showing the top flow-carrying edges (99th percentile) calculated using the maximum-flow algorithm between Rap1 gene expression module and NFKB1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

### Content of Rap1 GEM and its network-neighbourhood allow us to explore potential signaling events relevant in the regulation of cell shape

To understand latent processes driven by components within our gene expression module, we also studied interactions between the gene ontology terms enriched within Rap1 GEM (Figure 7A). This revealed that, as well as Rap1 signaling, the GEM is enriched in AGE-RAGE signaling pathway and HIF1 signaling pathway (consistent with HIF1A's activity correlating highly with Rap1 GEM in both cell line patient data). HIF1 is known to be regulated by Rap1 [57], although not explicitly in breast cancer.

NFkB has been previously linked to the regulation of cell shape in breast cancer. To explore the interface of Rap1 GEM with NFkB in terms of intra-cellular signaling, we identified a subnetwork of our network responsible for mediating 'information-flow' between those two nodes, using the algorithm maximum flow (Figure 7B). By studying the flow of information from Rap1 signaling, we can see that a LATS2/WWTR1/DVL1 (all of WNT signaling) lies between the target and source nodes with much of the flow being carried via these edges. This implicates YAP1/TAZ as being a key effector of the identified gene expression module. This finding is supported by TEAD1 (mediating gene expression of YAP1 and WWTR1/TAZ) being among the most highly correlated of TFs with Rap1 GEM (Figure 6B and S6A).

## Discussion

We present a method that uses transcriptomics and phenotypic data to derive a concise sub-network describing the signaling involved in the regulation of cell shape. This analysis recovered known processes like 'adherens junction proteins', 'cadherin' [58]and 'integrin' [59] [60] as well as pathways responsible for the regulation of cell shape in development, such as Wnt [61][62], TGF-$\beta$ [63] and NOTCH [64]. All of these pathways have previously been linked to the development of metastatic phenotypes in breast cancer cells [65] [64] [66]. Importantly, this analysis also sheds light on processes with less characterised associations with cell shape in cancer. We found that a gene expression module enriched in Rap1 signaling, is significantly correlated with cell shape, and is the most differentially expressed module between Luminal-like and Basal-like cell line clusters. We found that it was upregulated in basal-like cell lines while downregulated in luminal, consistent with its negative correlation with neighbour fraction; a cell shape feature most contributing to the 'cobblestone' like features of an epithelial and non-metastatic cell type. This gene expression module was also an important node in our identified signaling network, being at the network confluence of 2 activated transcription factors (JARID2 and RUNX2). We also showed this gene expression module to be expressed in patient data, with its activity being correlated with known developmental and morphologically related transcription factors. This illustrates that our gene expression module identified in only 14 cell lines has wider significance in other contexts. Rap1 is a small GTPase in the Ras-related protein family that has been shown to be involved in the regulation of cell adhesion and migration [67][68]. Specifically, Rap1 has been shown to modulate and activate NFkB activity in response to TNF-$\alpha$ stimulation in mesenchymal stem cells [68] and modulate migration and adhesion [69]. Rap1 is able to regulate IKKs (IB kinases) in a spatio-temporal manner [70], and is crucial for IkBK to be able to phosphorylate NFkB subunit p65 to make it competent [71] Here, we used our network-centric methodology to highlight a transcriptomic module, characterised in part by Rap1 signaling and that this is a key node in our phenotype-specific signaling network. We show that it is significantly correlated with cell-shape relevant transcription factors and that it interacts with our derived network through the developmental transcription factor YAP1. It is possible that our observations of the significance of Rap1 are as a result of more 'direct' interaction between Rap1 and the cytoskeleton. However, the transcriptomic module which we observe accounts for a much larger system-wide rewiring than simply the modulation of cytoskeletal proteins. This implies more complex transcriptional changes that are more long-lasting and robust to change.

The Rap1 signaling GEM identified in the network analysis represents a subset of the transcriptome observable among our analysed cell lines. While it is enriched in Rap1 signaling, it is important to note that it represents a collection of latent biological processes rather than a single pathway assigned to it by gene set enrichment. From our network analysis we hypothesise that it is able to interact with intracellular signaling pathways in order to modulate transcription factor activity and consequently cell shape. Other pathways enriched in the expression module include HIF-1 signaling pathway, which is known to be activated by Rap1 in melanoma [57], but this has not been shown in breast cancer. Also, AGE-RAGE

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

signaling was also enriched in our module of interest. AGE-RAGE signaling pathway has recently been shown to overlap with Rap1 signaling pathway in cardiac fibroblasts to alter the expression of NFkB [72], although this crosstalk has also not been illustrated in breast cancer. Here, we observe genes of all these pathways functioning as a cohesive unit in breast cancer, with their expression correlated with morphological traits associated with a negative prognosis.

We also observe that our gene expression module of interest is significantly correlated with NFkB in both clinical samples and cell culture. Other authors have flagged the direct effect of Rap1 on the cytoskeleton and NFkB [73][74], but here we go further, using a systems approach to link Rap1 signaling with multiple transcription factors and pathways. Based on known functions of Rap1, along with the functions of pathways that we find interact with it, we hypothesize that the identified transcriptomic unit is key in relaying information from a cell's physical environment to modulate and maintain the cancer stem cell niche [75].

Previous studies have established a connection between the NFkB signaling pathway and regulation of cell shape in breast cancer [26][25]. Our findings also illustrate the significance of this pathway in the regulation of cell morphology, with multiple NFkB regulators and transcriptional co-activators being flagged in our results. Some morphology-correlated gene expression modules were significantly differentially expressed between cell shape subtypes with the ARNT KO module being significantly upregulated in basal-like cell shapes relative to luminal. We also found this gene expression module to have the highest total correlation with all of the morphological features (see Figure 2B), indicating a strong association with cell shape. By studying terms enriched in this module from the EnrichR library, we find both 'TNF-$\alpha$ signaling via NFkB' to be enriched as well as genes downregulated during AhR nuclear translocator (ARNT) shRNA KO. Signaling by TNF is able to activate NFkB, a transcription factor known to control the expression of many EMT related genes [76] which has shown to be more sensitive to TNF-a stimulation in mesenchymal-like cellular morphologies than epithelial. This was hypothesised to generate a negative feedback which reinforces a metastatic phenotype of breast cancer cells [25]. Here we observe also that an ARNT KO/TNF module is upregulated in basal-like cell lines, consistently with these findings. ARNT is a protein shown to be involved in regulating tumour growth and angiogenesis along with its binding partner aryl hydrocarbon receptor (AhR) [77]. Previous studies have also shown its ability to modulate NFkB signaling with the activated form possibly interfering with the action of activated p65 [78]. Our findings that the upregulation of a gene expression module that is associated with ARNT knockdown further gives credence to NFkB being positively regulated in mesenchymal-like cell morphologies. Furthermore, the results of our network propagation yielded activators and transcriptional coactivators of NFkB (IKBKB [71], NR0B2 [79] and PPARGC1A [80]. These findings provide evidence that NFkB is modulated by both phosphorylation, spatial-temporal location and transcriptional co-activation in breast cancer to sense mechanical stimuli and modulate cell morphology.

Aside from the biological findings of this study, we illustrate a method for network analysis of a specific course-grained phenotype through expression; a notoriously poor (if cheap and widely available) proxy for gauging intracellular signaling [81]. In contrast to existing methods that use gene expression as a direct proxy for signaling [82][83][84][85], our approach infers transcription factor activities from the expression data and uses these as an anchor to infer upstream signaling networks relevant to the regulation of our phenotypes. Transcription factor activities can represent the outcome of a signal transduction process compared to the expression profiles and are thus a better proxy for cell signaling activities of the cell [86]. Such an approach has been previously used, for example by the tool CARNIVAL [87]. However, this and other available tools neglect the propensity for the transcriptome to be regulated in a highly context specific and modular structure [88][89]. Moreover, their reliance on annotated pathways to describe cell signalling undermines their ability to spot novel functional units, specific to a given phenotype. Here, using context-specific gene expression modules, we produced a network connecting the genes of interest from diverse analyses and used a network propagation algorithm to further focus on signaling proteins of novel interest. While there inevitably remains a level of bias stemming from the transcription factor regulon and pathway annotations, our bottom-up approach seeks to identify latent modular structure within transcriptomic data first. This puts the emphasis on data-driven gene expression modules, rather than literature-derived regulons and pathways. This approach takes an important step towards reducing the bias associated with previously annotated pathways and allows the identification of important regulatory units and their function with respect to cell shape from a systems biology point of view. Our network approach allows us to map the interface between two graphically presented systems in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cell; the transcriptome and intracellular signaling. Both can be easily combined with complex, multivariate phenotypic data which here has revealed a clearer picture of how signaling regulates cell morphology in breast cancer.

The interoperability of this approach is obvious, with any number of continuous variables measured with gene expression able to be correlated with module eigengenes using WGCNA. Here, we used OmniPath as a base network, but other network-based representations of the cellular environment can be used based on the appropriate context. Thus, our method represents a data-driven, network-based approach compatible with many different multi-scale phenotypes that are driven by intracellular signaling.

Overall, our unbiased network-based method highlights potential 'missing links' between sensing extracellular cues and transcriptional programmes that help maintain the cancer stem cell niche, and ultimately push breast cancer cells into EMT and metastasis. These represent starting points for further experimental studies to understand and therapeutically target the links between cell shape, cell signaling and gene regulation in the context of breast cancer.

# Materials and methods

### WGCNA analysis

Using Weighted correlation network analysis, we performed co-expression module identification using the R package WGCNA [33]. We used bulk RNA-seq data from Expression Atlas (in FPKM - E-MTAB-2770 and E-MTAB-2706) acquired from commonly used cancer cell lines of various cancer types [90]. We collated 13 breast cancer and 1 non-tumorigenic cell line for which imaging data was available (BT474, CAMA1, T47D, ZR75.1, SKBR3, MCF7, HCC1143, HCC1954, HCC70, hs578T, JIMT1, MCF10A, MDAMB157 and MDAMB231 [25]. We acquired representative images of each cell line from Sero et al., 2015 (*https://datadryad.org/stash/dataset/doi:10.5061/dryad.tc5g4*). Using Ensembl-Biomart, we filtered genes to only include protein-coding genes [91] and genes whose FPKM was greater than 1, leaving a total of 15,304 genes.

We created a signed, weighted adjacency matrix using log2 transformed gene expression values and a soft threshold power ($\alpha$) of 9. We translated this adjacency matrix (defined by Eq.1) into a topological overlap matrix (TOM; a measure of similarity) and the corresponding dissimilarity matrix (TOM - 1) was used to identify modules of correlated gene expression (minimum module size of 30).

$$Eq.1a \quad ij = (1 + cor(x_i, x_j))/2^{\beta}$$

We took morphological variables referring to breast cancer cell lines from Sero et al., 2015, which include 10 significant features shown to be predictive of TF activation [25]. We correlated these features with module eigengenes using Pearson correlation and we tested these values for significance by calculating Student asymptotic p-values for given correlations. For the modules that correlated with morphological features (Pearson Correlation Coefficient 0.5 and Student P<0.05), we identified enriched signaling pathways using the R package EnrichR [34], and the signaling database Reactome [43]. Using the database TRRUST v2 [35], we identified TF regulons that significantly overlap (Exact Fisher's test, P<0.1) with the gene expression module contents. This was done separately for inhibitory and activatory expression regulons for each transcription factor, with regulatory relationships of unknown sign being used in the significance calculations for both.

We named gene expression clusters using significantly enriched terms identified by the EnrichR analysis (Supplementary Table 4). As some clusters were very obscure, we utilized the entire EnrichR list of libraries (*https://maayanlab.cloud/Enrichr/#stats* for full list) with precedence going to the signaling databases of KEGG, Reactome, Panther and Wikipathways [92][93][48]. Some modules could not be assigned informative terms and so were named 'not annotated' (NA).

### Clustering and differential expression

Using the k-means algorithm, we classified the 14 breast cancer cell lines by the median values of each of their shape features (k=3, see Figure S2A). We performed differential expression analysis using the

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

R package DESeq2 [94]. We filtered genes so that only protein coding genes and those with more than 0.5 counts per million in at least 8 cell lines were included. We calculated Log2 fold changes with the cluster of interest as the numerator and the remaining cell lines acting as a control. Using the R package FGSEA [95], we performed gene set enrichment analysis of the differentially regulated proteins using the complete pathways gene set (Release 01 April 2020) from MSigDB [96] and the WGCNA gene expression modules identified in previous analysis. We calculated transcription factor regulon enrichment using the software DOROTHEA [53].

**Network Generation**

Using a Prize Collecting Steiner Forest (PCSF) algorithm, we generated a cell-shape regulatory network implemented through the R package PCSF [46]. For the prize-carrying nodes to be collected by the PCSF algorithm, we used the transcription factors significantly regulating the WGCNA modules using TRRUST ($p<0.1$), the differentially activated transcription factors identified by DOROTHEA ($p<0.1$), and the signaling proteins included in the REACTOME pathways that were enriched in transcription factors identified ($p<0.05$). We identified these pathways by using the TRRUST TFs identified in the previous steps, as well as ENCODE and ChEA Consensus TFs from ChIP-X [39], DNA binding preferences from JASPAR [40][41], TF protein-protein interactions and TFs from ENCODE ChiP-Seq [42]. Using EnrichR, we identified pathways that were enriched in the identified TFs, and the proteins that were included in these pathways were extracted from Pathway Commons using the R package paxtoolsr [97].

The 'costs' associated with each edge in the regulatory network were the inverse of the number of sources linked to each regulatory connection scaled between 1 and 0, such that the more the number of citations for an edge, the lower the cost. For the base network used by the algorithm, we used the comprehensive biological prior knowledge database, Omnipath [98], extracted using the R package OmnipathR [47]. We set each prize for significant TFs or signaling pathways to 100 and used a random variant of the PCSF algorithm with the result being the union of subnetworks obtained on each run (30 iterations) after adding random noise to the edge costs each time (5%). The algorithm also includes a hub-penalisation parameter which we set to 0.005. Other parameters include the tuning of node prices (set to 1) and the tuning of trees in the PCSF output (40).

We included the WGCNA modules themselves as super-nodes in the network, by adding incoming edges from the transcription factors contained within the regulatory network whose regulomes (as described in TRRUST [35]) significantly overlap (Fisher's exact test; $P<0.1$) with the gene content of the module in question. We represented the respective cell-shape phenotypes as nodes in a similar fashion, by including undirected edges from expression modules and phenotypes where there was significant correlation ($|PCC| > 0.5$ & $P<0.05$) between them. To account for expression modules' effect on upstream signaling, we added edges from the WGCNA modules back up to proteins that were themselves included within the modules. We set the edge weight of these to 1, such that any predicted activity of the gene expression module would be translated directly into its constituent signaling proteins and thus account for feedback between cell shape signaling networks, and the context-specific expression modules identified in the first step. We identified enriched terms in the network using the 2016 release of the database Panther [48] and GSE package EnrichR [34].

**Network propagation of functional TFs**

We examined the potential effect of significantly activated (FDR < 0.05), and deactivated TFs in different cell line clusters using network propagation in our generated network. We replaced edge weight with Resnik Best Match Average (BMA) semantic similarity [99] between the biological process GO terms of the two interacting pairs, with the sign of the interaction being inherited from Omnipath [47]. We then scaled the semantic similarity edge weights between 1 and -1.

We used the differentially activated transcription factors identified using DOROTHEA ($P<0.05$) as seeds for a Random Walk with Restart (RWR) algorithm using the R package diffuseR (available at: https://github.com/dirmeier/diffusr). We judged a node to be significantly ranked if its affinity score relative to the inputted seeds was greater than the same node's affinity score with a random walk simulation performed with randomised seeds. We performed this randomised simulation 10,000 times, from which a p-value was determined to judge significance ($P<0.1$). We performed this propagation by RWR for

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

both luminal-like and basal-like morphological clusters on significantly activated and deactivated transcription factors separately, in addition to simulations where each seed was considered in isolation. We implemented these simulations with a restart probability of 0.95 and performed GO term enrichment on significantly ranked nodes using EnrichR [34] and the signaling database Panther [48].

### Breast cancer cell morphology following kinase inhibitor treatment

We used small molecule kinase inhibition data from Harvard Medical School (HMS) Library of Integrated Network-based Cellular Signatures (LINCS) Center [100], which is funded by NIH grants U54 HG006097 and U54 HL127365 (available from: `https://lincs.hms.harvard.edu/mills-unpubl-2015/`). This dataset is derived from the treatment of 6 cell lines with a panel of 105 small molecule kinase inhibitors. They measured textural and morphological variables following treatment by high through-put image analysis [49][101]. We combined this assay with another dataset from HMS-LINCS; a Target Affinity Spectrum (TAS) for compounds in the HMS-LINCS small molecule library measuring the binding assertions based on dose-response affinity constants for particular kinase inhibitors (`https://lincs.hms.harvard.edu/db/datasets/20000/`). Using this dataset, we filtered for only molecule-binding target pairs with a binding class of 1 (representing a Kd <100nM affinity). Further to this, we removed molecules which had more than 5 targets with a Kd of 100nM. Consequently, the remaining kinase inhibitors were relatively narrow spectrum, thus simplifying analysis of their phenotypic effect. We expressed these results as batch-specific log fold changes of 10um drug treatment relative to the mean of the control set (untreated and DMSO treated cells). We did not present features which were functions of other features (such as ratios) as these were redundant. Correlations analysis was performed on the mean log fold change for a drug subtracted from the log fold change

The morphological data in the kinase inhibition screen was measured using two dyes (DRAQ5 and TMRE), the intensity of which we used to normalise textural features and the measurement of cytoplasmic and nuclear small spots. We reported counts for small nuclear or cytoplasmic spots as a mean of the individually normalised readings from both dyes. We considered a treatment perturbing our network if at least one of the kinase inhibitors targeted a protein that was represented by a node within the network.

### Quantifying kinase inhibitor influence

We incorporate information from the Target Affinity Spectrum assay, as well as graph-based properties of kinase inhibitor targets, using the product of the Szymkiewicz-Simpson similarity (measured between the cell shape network nodes and the drug targets) and the centrality of the targeted nodes in the predicted network with semantic similarity edge weights. The product of these generates, for a given kinase inhibitor the statistic:

$$Eq.2 \sum_{x \in K \cap N} PR(x) \times \frac{K \cap N}{min(|K|, |N|)}$$

Where K is the set of kinases an inhibitor is predicted to target, N is the nodes of the network and the function PR() is the centrality of a particular node in the network as defined by the PageRank algorithm [102]. This centrality measure has been shown to be effective in prioritizing proteins by relative importance in signaling or protein-protein interaction networks [103]. We used this statistic as a measure of a kinase inhibitor's influence on cell shape.

### Analysis of BRCA cell line and TCGA sample RNAseq data

For the cell lines, we used RNA-seq data from Expression Atlas (in FPKM - E-MTAB-2770 and E-MTAB-2706) [90]. This was analysed using DESeq2 [94] as per the methodology in the subsection entitled "Clustering and differential expression". Both TF and module activity was calculated using the algorithm VIPER [104]. For patient data, the results shown here are based upon data generated by the The Cancer Genome Atlas (TCGA) Research Network:*https://www.cancer.gov/tcga*. For computational efficiency, we use Gamma-Poisson models to predict differentially expressed genes from our samples using the package glmGamPoi [105]. We use the sample of interest as the numerator with the remaining tumour samples acting as a control. For quantifying correlation between Rap1 - GEM and different transcription factors we remove samples with insignificant activation of either the TF in question or Rap1-GEM (FDR

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

adjusted P value < 0.05). Correlation was quantified using Kendall rank correlation coefficient. Differences in distributions of morphological subtypes was quantified by Kolmogorov-Smirnov test.

## Maximum-flow network analysis

For maximum-flow calculations, we used the Resnik BMA semantic similarity [99] as the maximum 'carrying capacity' of an edge in the network. To visualise the optimised solution (as implemented by the R package igraph [106] we selected only those edges in the 99th percentile of the flow-carrying edges in the network. Visualisation was performed using the software, Cytoscape [107]. Maximum flow was performed with the R package igraph [106].

## Quantification and Statistical Analysis

Statistical tests were performed in base R unless otherwise mentioned in the methods and p-value cutoffs are shown in parentheses after reporting an effect as significant. Weighted Pearson correlation with t-test for significance was used to correlate eigengenes and cell shape features using the RNA package WGCNA. We used a one-way ANOVA test for comparing the means of the shape variables among the identified 3 cell line clusters (n = 75,653) and a Tukey honest significant differences test to perform multiple-pairwise comparison among the means of the groups. The same tests were performed on the differences in 10 cell shape variables when HS578T was treated with 37 kinase inhibitors (n = 23,128). Fisher exact test was used to test significance of overlap between TRRUST regulons and identified gene expression modules (Supplementary Table 3 shows the size of the overlap).

Enrichment of gene sets was performed by EnrichR, an enrichment library that utilises a hyper-geometric test to identify significantly enriched terms in a gene list. This tool (described in [108] calculates a score combining the Fisher exact test p-value of the enrichment with the z-score deviation from the expected rank. The pre-ranked gene-set enrichment algorithm FGSEA was used for the identification of enriched terms in the differentially expressed genes allowing for accurate estimation of arbitrarily low P-values that occur in expression datasets.

Spearman rank correlation was used to measure the strength of the association between target network centrality and the measured effect of its perturbation by inhibition. Spearman was chosen because the centrality (combined with Szymkiewicz-Simpson) according to equation 2 does not follow an exact normal distribution. Kendall rank correlation coefficient was used when calculating the correlation between TF activity and Rap1-GEM activity because confidence intervals for Spearman's $r_S$ are less reliable and less interpretable than confidence intervals for Kendall's -parameters. When trying to distinguish between many correlations of similar quality, this becomes more important. FDR adjustment for multiple testing correction was always used when multiple tests were performed in the same analysis.

Kolmogorov-Smirnov test was used to measure differences in distributions of clinically assigned tumour morphologies. This was because clinical groupings are mixed (i.e. Infiltrating duct and lobular carcinoma) and others are characterized by an absence of features over their presence. This means that the assumption of normality required for a t-test is not fulfilled.

For differential expression analysis the DESeq2 R package [94] was used. DESeq2 fits negative binomial generalized linear models for each gene and uses the Wald test for significance testing. The package then automatically detects count outliers using Cooks's distance and removes these genes from analysis.

Significance was determined for RWR network propagation by randomising seed nodes (preserving their values) 10,000 times and selecting only the non-seed nodes that were significantly ranked relative to the randomised simulations (P<0.1).

## Data and Software Availability

The complete R scripts used for this methodology are available on Gitlab; `https://gitlab.ebi.ac.uk/petsalakilab/phenotype_networks`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## Acknowledgements

We would like to thank EMBL-EBI for funding the project. We would also like to thank Bishoy Wadie and Vivian Robin for critical reading of the manuscript.

## Author Contributions

**CB:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft**. EiP:** Conceptualization, Methodology, Software. **GG:** Conceptualization, Supervision. **JS:** Resources. **ENE:** Validation, Writing - Original Draft**. ChrB:** Writing - Review & Editing, Resources. **EP:** Writing - Review & Editing, Conceptualization, Supervision, Methodology, Project administration, Funding acquisition.

## Competing interests

The authors declare no conflict of interest.

## References

[1] Pei-Hsun Wu et al. "Single-cell morphology encodes metastatic potential". In: *Science Advances* 6.4 (Jan. 1, 2020). Publisher: American Association for the Advancement of Science Section: Research Article, eaaw6938. ISSN: 2375-2548. DOI: 10.1126/sciadv.aaw6938. URL: https://advances.sciencemag.org/content/6/4/eaaw6938 (visited on 11/30/2020).

[2] Janani P. Baskaran et al. "Cell shape, and not 2D migration, predicts extracellular matrix-driven 3D cell invasion in breast cancer". In: *APL Bioengineering* 4.2 (May 5, 2020). Publisher: American Institute of Physics, p. 026105. DOI: 10.1063/1.5143779. URL: https://aip.scitation.org/doi/10.1063/1.5143779 (visited on 02/03/2021).

[3] N. V. Krakhmal et al. "Cancer Invasion: Patterns and Mechanisms". In: *Acta Naturae* 7.2 (2015), pp. 17–28. ISSN: 2075-8251. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4463409/ (visited on 02/03/2021).

[4] Jonathan Cooper and Filippo G. Giancotti. "Integrin Signaling in Cancer: Mechanotransduction, Stemness, Epithelial Plasticity, and Therapeutic Resistance". In: *Cancer Cell* 35.3 (Mar. 2019), pp. 347–367. ISSN: 15356108. DOI: 10.1016/j.ccell.2019.01.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S1535610819300418 (visited on 02/03/2021).

[5] Francesc Miralles et al. "Actin dynamics control SRF activity by regulation of its coactivator MAL". In: *Cell* 113.3 (May 2, 2003), pp. 329–342. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(03)00278-2.

[6] Eric N. Olson and Alfred Nordheim. "Linking actin dynamics and gene transcription to drive cellular motile functions". In: *Nature Reviews. Molecular Cell Biology* 11.5 (May 2010), pp. 353–365. ISSN: 1471-0080. DOI: 10.1038/nrm2890.

[7] S. Orsulic et al. "E-cadherin binding prevents beta-catenin nuclear localization and beta-catenin/LEF-1-mediated transactivation". In: *Journal of Cell Science* 112.8 (Apr. 15, 1999). Publisher: The Company of Biologists Ltd Section: Journal Articles, pp. 1237–1245. ISSN: 0021-9533, 1477-9137. URL: https://jcs.biologists.org/content/112/8/1237 (visited on 01/20/2021).

[8] Bin Zheng et al. "Nuclear actin and actin-binding proteins in the regulation of transcription and gene expression". In: *The FEBS journal* 276.10 (May 2009), pp. 2669–2685. ISSN: 1742-4658. DOI: 10.1111/j.1742-4658.2009.06986.x.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[9] Sandeep Kumar et al. "Functional screening of mammalian mechanosensitive genes using Drosophila RNAi library– Smarcd3/Bap60 is a mechanosensitive pro-inflammatory gene". In: *Scientific Reports* 6.1 (Nov. 7, 2016). Number: 1 Publisher: Nature Publishing Group, p. 36461. ISSN: 2045-2322. DOI: 10.1038/srep36461. URL: https://www.nature.com/articles/srep36461 (visited on 11/17/2020).

[10] Joëlle Roche. "The Epithelial-to-Mesenchymal Transition in Cancer". In: *Cancers* 10.2 (Feb. 16, 2018). ISSN: 2072-6694. DOI: 10.3390/cancers10020052. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5836084/ (visited on 11/30/2020).

[11] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. "Cancer statistics, 2019". In: *CA: A Cancer Journal for Clinicians* 69.1 (2019). _eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21 pp. 7–34. ISSN: 1542-4863. DOI: https://doi.org/10.3322/caac.21551. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21551 (visited on 02/03/2021).

[12] Yixiao Feng et al. "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis". In: *Genes & Diseases* 5.2 (May 12, 2018), pp. 77–106. ISSN: 2352-4820. DOI: 10.1016/j.gendis.2018.05.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6147049/ (visited on 11/30/2020).

[13] Meng-Horng Lee et al. "Normal mammary epithelial cells promote carcinoma basement membrane invasion by inducing microtubule-rich protrusions". In: *Oncotarget* 6.32 (Oct. 20, 2015), pp. 32634–32645. ISSN: 1949-2553. DOI: 10.18632/oncotarget.4728.

[14] Monica Fedele, Laura Cerchia, and Gennaro Chiappetta. "The Epithelial-to-Mesenchymal Transition in Breast Cancer: Focus on Basal-Like Carcinomas". In: *Cancers* 9.10 (Sept. 30, 2017). ISSN: 2072-6694. DOI: 10.3390/cancers9100134. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5664073/ (visited on 11/30/2020).

[15] Xiaofeng Dai et al. "Breast cancer intrinsic subtype classification, clinical use and future trends". In: *American Journal of Cancer Research* 5.10 (Sept. 15, 2015), pp. 2929–2943. ISSN: 2156-6976. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4656721/ (visited on 11/30/2020).

[16] Richard M. Neve et al. "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes". In: *Cancer Cell* 10.6 (Dec. 1, 2006), pp. 515–527. ISSN: 1535-6108. DOI: 10.1016/j.ccr.2006.10.008. URL: http://www.sciencedirect.com/science/article/pii/S153561080600314X (visited on 01/20/2021).

[17] Michele A. Wozniak and Christopher S. Chen. "Mechanotransduction in development: a growing role for contractility". In: *Nature reviews. Molecular cell biology* 10.1 (Jan. 2009), pp. 34–43. ISSN: 1471-0072. DOI: 10.1038/nrm2592. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2952188/ (visited on 02/03/2021).

[18] Martin Bergert et al. "Cell Surface Mechanics Gate Embryonic Stem Cell Differentiation". In: *Cell Stem Cell* (Nov. 17, 2020). ISSN: 1934-5909. DOI: 10.1016/j.stem.2020.10.017. URL: http://www.sciencedirect.com/science/article/pii/S1934590920305336 (visited on 12/09/2020).

[19] Henry De Belly et al. "Membrane Tension Gates ERK-Mediated Regulation of Pluripotent Cell Fate". In: *Cell Stem Cell* (Nov. 19, 2020). ISSN: 1934-5909. DOI: 10.1016/j.stem.2020.10.018. URL: http://www.sciencedirect.com/science/article/pii/S1934590920305348 (visited on 02/03/2021).

[20] Shampa Chatterjee. "Endothelial Mechanotransduction, Redox Signaling and the Regulation of Vascular Inflammatory Pathways". In: *Frontiers in Physiology* 9 (2018). Publisher: Frontiers. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.00524. URL: https://www.frontiersin.org/articles/10.3389/fphys.2018.00524/full (visited on 02/03/2021).

[21] Catherine F. Cowell et al. "Loss of cell-cell contacts induces NF-kappaB via RhoA-mediated activation of protein kinase D1". In: *Journal of Cellular Biochemistry* 106.4 (Mar. 1, 2009), pp. 714–728. ISSN: 1097-4644. DOI: 10.1002/jcb.22067.

[22] Yasuhiro Ishihara et al. "Role of NF-kB RelB in Aryl Hydrocarbon Receptor-Mediated Ligand Specific Effects". In: *International Journal of Molecular Sciences* 20.11 (May 30, 2019). ISSN: 1422-0067. DOI: 10.3390/ijms20112652. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6600526/ (visited on 11/27/2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[23] Cynthia K. Shrum, Daniel Defrancisco, and Mollie K. Meffert. "Stimulated nuclear translocation of NF-kappaB and shuttling differentially depend on dynein and the dynactin complex". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.8 (Feb. 24, 2009), pp. 2647–2652. ISSN: 1091-6490. DOI: 10.1073/pnas.0806677106.

[24] Louis Tong and Vinay Tergaonkar. "Rho protein GTPases and their interactions with NFB: crossroads of inflammation and matrix biology". In: *Bioscience Reports* 34.3 (June 25, 2014). ISSN: 0144-8463. DOI: 10.1042/BSR20140021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4069681/ (visited on 11/29/2020).

[25] Julia E Sero et al. "Cell shape and the microenvironment regulate nuclear translocation of NF-B in breast epithelial and tumor cells". In: *Molecular Systems Biology* 11.3 (Mar. 3, 2015). ISSN: 1744-4292. DOI: 10.15252/msb.20145644. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380925/ (visited on 11/05/2020).

[26] Heba Z. Sailem and Chris Bakal. "Identification of clinically predictive metagenes that encode components of a network coupling cell shape to transcription by image-omics". In: *Genome Research* 27.2 (Feb. 2017), pp. 196–207. ISSN: 1088-9051. DOI: 10.1101/gr.202028.115. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5287226/ (visited on 11/03/2020).

[27] Christine Vogel and Edward M. Marcotte. "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". In: *Nature Reviews. Genetics* 13.4 (Mar. 13, 2012), pp. 227–232. ISSN: 1471-0064. DOI: 10.1038/nrg3185.

[28] Carole Bougault et al. "Dynamic Compression of Chondrocyte-Agarose Constructs Reveals New Candidate Mechanosensitive Genes". In: *PLOS ONE* 7.5 (May 17, 2012). Publisher: Public Library of Science, e36964. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0036964. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0036964 (visited on 02/04/2021).

[29] Edward R. Horton et al. "Mechanosensitivity of integrin adhesion complexes: role of the consensus adhesome". In: *Experimental Cell Research* 343.1 (Apr. 10, 2016), pp. 7–13. ISSN: 0014-4827. DOI: 10.1016/j.yexcr.2015.10.025. URL: http://www.sciencedirect.com/science/article/pii/S0014482715301300 (visited on 02/04/2021).

[30] Joseph Robertson et al. "Defining the phospho-adhesome through the phosphoproteomic analysis of integrin signalling". In: *Nature Communications* 6.1 (Feb. 13, 2015). Number: 1 Publisher: Nature Publishing Group, p. 6265. ISSN: 2041-1723. DOI: 10.1038/ncomms7265. URL: https://www.nature.com/articles/ncomms7265 (visited on 02/04/2021).

[31] Rebecca A. Rolfe et al. "Identification of mechanosensitive genes during skeletal development: alteration of genes associated with cytoskeletal rearrangement and cell signalling pathways". In: *BMC Genomics* 15.1 (Jan. 20, 2014), p. 48. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-48. URL: https://doi.org/10.1186/1471-2164-15-48 (visited on 02/04/2021).

[32] Aravind Subramanian et al. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles". In: *Cell* 171.6 (Nov. 30, 2017), 1437–1452.e17. ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.10.049.

[33] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1 (Dec. 29, 2008), p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559. URL: https://doi.org/10.1186/1471-2105-9-559 (visited on 11/03/2020).

[34] Maxim V. Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". In: *Nucleic Acids Research* 44 (W1 2016), W90–97. ISSN: 1362-4962. DOI: 10.1093/nar/gkw377.

[35] Heonjong Han et al. "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions". In: *Nucleic Acids Research* 46 (D1 2018), pp. D380–D386. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1013.

[36] Yan Song, M. Kay Washington, and Howard C. Crawford. "Loss of FOXA1/2 Is Essential for the Epithelial-to-Mesenchymal Transition in Pancreatic Cancer". In: *Cancer Research* 70.5 (Mar. 1, 2010). Publisher: American Association for Cancer Research Section: Tumor and Stem Cell Biology, pp. 2115–2125. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-09-2979. URL: https://cancerres.aacrjournals.org/content/70/5/2115 (visited on 01/21/2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[37] Xinyan Wu et al. "HOXB7, a Homeodomain Protein, Is Overexpressed in Breast Cancer and Confers Epithelial-Mesenchymal Transition". In: *Cancer Research* 66.19 (Oct. 1, 2006). Publisher: American Association for Cancer Research Section: Cell, Tumor, and Stem Cell Biology, pp. 9527–9534. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-05-4470. URL: https://cancerres.aacrjournals.org/content/66/19/9527 (visited on 01/21/2021).

[38] Elizabeth A. Van Tubergen et al. "Inactivation or loss of TTP promotes invasion in head and neck cancer via transcript stabilization and secretion of MMP9, MMP2, and IL-6". In: *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 19.5 (Mar. 1, 2013), pp. 1169–1179. ISSN: 1557-3265. DOI: 10.1158/1078-0432.CCR-12-2927.

[39] Alexander Lachmann et al. "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments". In: *Bioinformatics (Oxford, England)* 26.19 (Oct. 1, 2010), pp. 2438–2444. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq466.

[40] Gary D. Stormo. "Modeling the specificity of protein-DNA interactions". In: *Quantitative Biology (Beijing, China)* 1.2 (June 2013), pp. 115–130. ISSN: 2095-4689. DOI: 10.1007/s40484-013-0012-4.

[41] Wyeth W. Wasserman and Albin Sandelin. "Applied bioinformatics for the identification of regulatory elements". In: *Nature Reviews. Genetics* 5.4 (Apr. 2004), pp. 276–287. ISSN: 1471-0056. DOI: 10.1038/nrg1315.

[42] Ghia M. Euskirchen et al. "Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies". In: *Genome Research* 17.6 (June 2007), pp. 898–909. ISSN: 1088-9051. DOI: 10.1101/gr.5583007.

[43] Bijay Jassal et al. "The reactome pathway knowledgebase". In: *Nucleic Acids Research* 48 (D1 2020), pp. D498–D503. ISSN: 1362-4962. DOI: 10.1093/nar/gkz1031.

[44] Wenting Shih and Soichiro Yamada. "N-cadherin-mediated cell–cell adhesion promotes cell migration in a three-dimensional matrix". In: *Journal of Cell Science* 125.15 (Aug. 1, 2012), pp. 3661–3670. ISSN: 0021-9533. DOI: 10.1242/jcs.103861. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3445327/ (visited on 11/06/2020).

[45] Luz Garcia-Alonso et al. "Benchmark and integration of resources for the estimation of human transcription factor activities". In: *Genome Research* 29.8 (Jan. 8, 2019). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1363–1375. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.240663.118. URL: http://genome.cshlp.org/content/29/8/1363 (visited on 11/09/2020).

[46] Murodzhon Akhmedov, Ivo Kwee, and Roberto Montemanni. "A divide and conquer matheuristic algorithm for the Prize-collecting Steiner Tree Problem". In: *Computers & Operations Research* 70 (June 1, 2016), pp. 18–25. ISSN: 0305-0548. DOI: 10.1016/j.cor.2015.12.015. URL: http://www.sciencedirect.com/science/article/pii/S0305054815003019 (visited on 11/24/2020).

[47] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. "OmniPath: guidelines and gateway for literature-curated signaling pathway resources". In: *Nature Methods* 13.12 (Dec. 2016). Number: 12 Publisher: Nature Publishing Group, pp. 966–967. ISSN: 1548-7105. DOI: 10.1038/nmeth.4077. URL: https://www.nature.com/articles/nmeth.4077 (visited on 11/24/2020).

[48] Huaiyu Mi, Anushya Muruganujan, and Paul D. Thomas. "PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees". In: *Nucleic Acids Research* 41 (Database issue Jan. 2013), pp. D377–386. ISSN: 1362-4962. DOI: 10.1093/nar/gks1118.

[49] Nicholas A Hamilton et al. "Fast automated cell phenotype image classification". In: *BMC Bioinformatics* 8 (Mar. 30, 2007), p. 110. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1847687/ (visited on 11/25/2020).

[50] Nienke Moret et al. "Cheminformatics Tools for Analyzing and Designing Optimized Small-Molecule Collections and Libraries". In: *Cell Chemical Biology* 26.5 (May 16, 2019), 765–777.e3. ISSN: 2451-9456. DOI: 10.1016/j.chembiol.2019.02.018. URL: http://www.sciencedirect.com/science/article/pii/S245194561930073X (visited on 12/08/2020).

[51] Bence Szalai and Julio Saez-Rodriguez. "Why do pathway methods work better than they should?" In: *bioRxiv* (July 30, 2020). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.07.30.228296. DOI: 10.1101/2020.07.30.228296. URL: https://www.biorxiv.org/content/10.1101/2020.07.30.228296v1 (visited on 12/09/2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[52] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. "Random walk with restart: fast solutions and applications". In: *Knowledge and Information Systems* 14.3 (Mar. 1, 2008), pp. 327–346. ISSN: 0219-3116. DOI: 10.1007/s10115-007-0094-2. URL: https://doi.org/10.1007/s10115-007-0094-2 (visited on 11/25/2020).

[53] Christian H. Holland et al. "Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data". In: *Genome Biology* 21.1 (Feb. 12, 2020), p. 36. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1949-z. URL: https://doi.org/10.1186/s13059-020-1949-z (visited on 11/24/2020).

[54] Theresa L Schwartz et al. "Metaplastic breast cancer: histologic characteristics, prognostic factors and systemic treatment strategies". In: *Experimental Hematology & Oncology* 2 (Nov. 14, 2013), p. 31. ISSN: 2162-3619. DOI: 10.1186/2162-3619-2-31. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3832232/ (visited on 06/15/2021).

[55] Tejaswini P. Reddy et al. "A comprehensive overview of metaplastic breast cancer: clinical features and molecular aberrations". In: *Breast Cancer Research* 22.1 (Nov. 4, 2020), p. 121. ISSN: 1465-542X. DOI: 10.1186/s13058-020-01353-z. URL: https://doi.org/10.1186/s13058-020-01353-z (visited on 06/15/2021).

[56] Jaafar Makki. "Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance". In: *Clinical Medicine Insights. Pathology* 8 (Dec. 21, 2015), pp. 23–31. ISSN: 1179-5557. DOI: 10.4137/CPath.S31563. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4689326/ (visited on 06/23/2021).

[57] Jae-Wook Lee et al. "Hypoxia/reoxygenation-experienced cancer cell migration and metastasis are regulated by Rap1- and Rac1-GTPase activation via the expression of thymosin beta-4". In: *Oncotarget* 6.12 (Mar. 23, 2015), pp. 9820–9833. ISSN: 1949-2553. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496400/ (visited on 06/21/2021).

[58] H. Eslami Amirabadi et al. "Characterizing the invasion of different breast cancer cell lines with distinct E-cadherin status in 3D using a microfluidic system". In: *Biomedical Microdevices* 21.4 (Nov. 23, 2019), p. 101. ISSN: 1572-8781. DOI: 10.1007/s10544-019-0450-5. URL: https://doi.org/10.1007/s10544-019-0450-5 (visited on 11/26/2020).

[59] Andrew Filer and Christopher D. Buckley. "15 - Fibroblasts and Fibroblast-like Synoviocytes". In: *Kelley's Textbook of Rheumatology (Ninth Edition)*. Ed. by Gary S. Firestein et al. Philadelphia: W.B. Saunders, Jan. 1, 2013, pp. 215–231. ISBN: 978-1-4377-1738-9. DOI: 10.1016/B978-1-4377-1738-9.00015-3. URL: http://www.sciencedirect.com/science/article/pii/B9781437717389000153 (visited on 11/26/2020).

[60] Aliakbar Taherian et al. "Differences in integrin expression and signaling within human breast cancer cells". In: *BMC Cancer* 11.1 (July 13, 2011), p. 293. ISSN: 1471-2407. DOI: 10.1186/1471-2407-11-293. URL: https://doi.org/10.1186/1471-2407-11-293 (visited on 11/26/2020).

[61] Rachel S. Kadzik et al. "Wnt ligand/Frizzled 2 receptor signaling regulates tube shape and branchpoint formation in the lung through control of epithelial cell shape". In: *Proceedings of the National Academy of Sciences* 111.34 (Aug. 26, 2014). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 12444–12449. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1406639111. URL: https://www.pnas.org/content/111/34/12444 (visited on 01/07/2021).

[62] Marjolein Wildwater et al. "Cell shape and Wnt signaling redundantly control the division axis of C. elegans epithelial stem cells". In: *Development* 138.20 (Oct. 15, 2011). Publisher: Oxford University Press for The Company of Biologists Limited Section: DEVELOPMENT AND STEM CELLS, pp. 4375–4385. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.066431. URL: https://dev.biologists.org/content/138/20/4375 (visited on 01/07/2021).

[63] J. Lee, J.-H. Choi, and C.-K. Joo. "TGF- 1 regulates cell fate during epithelial–mesenchymal transition by upregulating survivin". In: *Cell Death & Disease* 4.7 (July 2013). Number: 7 Publisher: Nature Publishing Group, e714–e714. ISSN: 2041-4889. DOI: 10.1038/cddis.2013.244. URL: https://www.nature.com/articles/cddis2013244 (visited on 01/07/2021).

[64] Emmanuel N. Kontomanolis et al. "The Notch Pathway in Breast Cancer Progression". In: *The Scientific World Journal* 2018 (July 8, 2018). ISSN: 2356-6140. DOI: 10.1155/2018/2415489. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6077551/ (visited on 01/07/2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[65] Takeshi Imamura, Atsuhiko Hikita, and Yasumichi Inoue. "The roles of TGF- signaling in carcinogenesis and breast cancer metastasis". In: *Breast Cancer (Tokyo, Japan)* 19.2 (Apr. 2012), pp. 118–124. ISSN: 1880-4233. DOI: 10.1007/s12282-011-0321-2.

[66] Ping Yin et al. "Wnt signaling in human and mouse breast cancer: Focusing on Wnt ligands, receptors and antagonists". In: *Cancer Science* 109.11 (Nov. 2018), pp. 3368–3375. ISSN: 1347-9032. DOI: 10.1111/cas.13771. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6215866/ (visited on 01/07/2021).

[67] Benjamin Boettner and Linda Van Aelst. "Control of cell adhesion dynamics by Rap1 signaling". In: *Current opinion in cell biology* 21.5 (Oct. 2009), pp. 684–693. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2009.06.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841981/ (visited on 07/07/2021).

[68] Yi-Lei Zhang et al. "Roles of Rap1 signaling in tumor cell migration and invasion". In: *Cancer Biology & Medicine* 14.1 (Feb. 2017), pp. 90–99. ISSN: 2095-3941. DOI: 10.20892/j.issn.2095-3941.2016.0086. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5365179/ (visited on 11/27/2020).

[69] Ketki Sawant et al. "Rap1 GTPase promotes coordinated collective cell migration in vivo". In: *Molecular Biology of the Cell* 29.22 (Nov. 1, 2018), pp. 2656–2673. ISSN: 1059-1524. DOI: 10.1091/mbc.E17-12-0752. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249841/ (visited on 07/07/2021).

[70] Yusuke Ohba, Kazuo Kurokawa, and Michiyuki Matsuda. "Mechanism of the spatio-temporal regulation of Ras and Rap1". In: *The EMBO Journal* 22.4 (Feb. 17, 2003), pp. 859–869. ISSN: 0261-4189. DOI: 10.1093/emboj/cdg087. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC145447/ (visited on 11/28/2020).

[71] Hsiangling Teo et al. "Telomere-independent Rap1 is an IKK adaptor and regulates NF-kappaB-dependent gene expression". In: *Nature Cell Biology* 12.8 (Aug. 2010), pp. 758–767. ISSN: 1476-4679. DOI: 10.1038/ncb2080.

[72] Stephanie D. Burr and James A. Stewart. "Rap1a Overlaps the AGE/RAGE Signaling Cascade to Alter Expression of -SMA, p-NF-B, and p-PKC- in Cardiac Fibroblasts Isolated from Type 2 Diabetic Mice". In: *Cells* 10.3 (Mar. 4, 2021), p. 557. ISSN: 2073-4409. DOI: 10.3390/cells10030557. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8000763/ (visited on 06/21/2021).

[73] Michael Ku Yu et al. "Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems". In: *Cell Systems* 2.2 (Feb. 24, 2016), pp. 77–88. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.02.003. URL: http://www.sciencedirect.com/science/article/pii/S2405471216300333 (visited on 04/07/2017).

[74] Y. Zhang et al. "Rap1-mediated nuclear factor-kappaB (NF- B) activity regulates the paracrine capacity of mesenchymal stem cells in heart repair following infarction". In: *Cell Death Discovery* 1.1 (Aug. 24, 2015). Number: 1 Publisher: Nature Publishing Group, pp. 1–11. ISSN: 2058-7716. DOI: 10.1038/cddiscovery.2015.7. URL: https://www.nature.com/articles/cddiscovery20157 (visited on 11/28/2020).

[75] Ankit Roy Choudhury et al. "Mechanobiology of Cancer Stem Cells and Their Niche". In: *Cancer Microenvironment* 12.1 (Apr. 19, 2019), pp. 17–27. ISSN: 1875-2292. DOI: 10.1007/s12307-019-00222-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6529500/ (visited on 02/03/2021).

[76] Bruno R. B. Pires et al. "NF-kappaB Is Involved in the Regulation of EMT Genes in Breast Cancer Cells". In: *PLoS ONE* 12.1 (Jan. 20, 2017). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0169622. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5249109/ (visited on 01/21/2021).

[77] Chi-Ruei Huang et al. "Down-regulation of ARNT promotes cancer metastasis by activating the fibronectin/integrin 1/FAK axis". In: *Oncotarget* 6.13 (Mar. 20, 2015), pp. 11530–11546. ISSN: 1949-2553. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4484474/ (visited on 01/28/2021).

[78] Johan Øvrevik et al. "AhR and Arnt differentially regulate NF-B signaling and chemokine responses in human bronchial epithelial cells". In: *Cell Communication and Signaling : CCS* 12 (July 24, 2014), p. 48. ISSN: 1478-811X. DOI: 10.1186/s12964-014-0048-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222560/ (visited on 11/27/2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[79] An Zou et al. "New Insights into Orphan Nuclear Receptor SHP in Liver Cancer". In: *Nuclear receptor research* 2 (2015), p. 101162. ISSN: 2314-5706. DOI: 10.11131/2015/101162. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4618403/ (visited on 06/25/2021).

[80] Li Hua Wang et al. "Inhibition of adhesive interaction between multiple myeloma and bone marrow stromal cells by PPAR cross talk with NF-B and C/EBP". In: *Blood* 110.13 (Dec. 15, 2007), pp. 4373–4384. ISSN: 0006-4971. DOI: 10.1182/blood-2006-07-038026. URL: https://doi.org/10.1182/blood-2006-07-038026 (visited on 06/25/2021).

[81] Mehran Piran et al. "Can We Assume the Gene Expression Profile as a Proxy for Signaling Network Activity?" In: *Biomolecules* 10.6 (June 2020). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 850. DOI: 10.3390/biom10060850. URL: https://www.mdpi.com/2218-273X/10/6/850 (visited on 12/01/2020).

[82] Rotem Ben-Hamo, Moriah Gidoni, and Sol Efroni. "PhenoNet: identification of key networks associated with disease phenotype". In: *Bioinformatics* 30.17 (Sept. 1, 2014), pp. 2399–2405. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu199. URL: https://doi.org/10.1093/bioinformatics/btu199 (visited on 12/09/2020).

[83] Yuanfang Guan et al. "Tissue-Specific Functional Networks for Prioritizing Phenotype and Disease Genes". In: *PLOS Computational Biology* 8.9 (Sept. 27, 2012). Publisher: Public Library of Science, e1002694. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002694. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002694 (visited on 12/10/2020).

[84] Megha Padi and John Quackenbush. "Detecting phenotype-driven transitions in regulatory network structure". In: *npj Systems Biology and Applications* 4.1 (Apr. 19, 2018). Number: 1 Publisher: Nature Publishing Group, pp. 1–12. ISSN: 2056-7189. DOI: 10.1038/s41540-018-0052-5. URL: https://www.nature.com/articles/s41540-018-0052-5 (visited on 12/10/2020).

[85] Jamie Soul et al. "PhenomeExpress: A refined network analysis of expression datasets by inclusion of known disease phenotypes". In: *Scientific Reports* 5.1 (Jan. 29, 2015). Number: 1 Publisher: Nature Publishing Group, p. 8117. ISSN: 2045-2322. DOI: 10.1038/srep08117. URL: https://www.nature.com/articles/srep08117 (visited on 12/09/2020).

[86] Bence Szalai and Julio Saez-Rodriguez. "Why do pathway methods work better than they should?" In: *FEBS Letters* 594.24 (Dec. 3, 2020). _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.14011, pp. 4189–4200. ISSN: 1873-3468. DOI: https://doi.org/10.1002/1873-3468.14011. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.14011 (visited on 12/08/2020).

[87] Anika Liu et al. "From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL". In: *npj Systems Biology and Applications* 5.1 (Nov. 11, 2019). Number: 1 Publisher: Nature Publishing Group, pp. 1–10. ISSN: 2056-7189. DOI: 10.1038/s41540-019-0118-z. URL: https://www.nature.com/articles/s41540-019-0118-z (visited on 01/07/2021).

[88] Hiroaki Kitano. "Systems Biology: A Brief Overview". In: *Science* 295.5560 (Mar. 1, 2002). Publisher: American Association for the Advancement of Science Section: Special Reviews, pp. 1662–1664. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1069492. URL: https://science.sciencemag.org/content/295/5560/1662 (visited on 12/09/2020).

[89] Sumana Sharma and Evangelia Petsalaki. "Large-scale datasets uncovering cell signalling networks in cancer: context matters". In: *Current Opinion in Genetics & Development*. Cancer Genomics 54 (Feb. 1, 2019), pp. 118–124. ISSN: 0959-437X. DOI: 10.1016/j.gde.2019.05.001. URL: http://www.sciencedirect.com/science/article/pii/S0959437X18301278 (visited on 01/07/2021).

[90] Irene Papatheodorou et al. "Expression Atlas update: from tissues to single cells". In: *Nucleic Acids Research* 48 (D1 Jan. 8, 2020). Publisher: Oxford Academic, pp. D77–D83. ISSN: 0305-1048. DOI: 10.1093/nar/gkz947. URL: https://academic.oup.com/nar/article/48/D1/D77/5609521 (visited on 11/24/2020).

[91] Rhoda J. Kinsella et al. "Ensembl BioMarts: a hub for data retrieval across taxonomic space". In: *Database: The Journal of Biological Databases and Curation* 2011 (2011), bar030. ISSN: 1758-0463. DOI: 10.1093/database/bar030.

[92] Minoru Kanehisa et al. "KEGG as a reference resource for gene and protein annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D457–D462. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1070. URL: https://doi.org/10.1093/nar/gkv1070 (visited on 01/22/2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[93] Marvin Martens et al. "WikiPathways: connecting communities". In: *Nucleic Acids Research* 49 (D1 Jan. 8, 2021), pp. D613–D621. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1024. URL: https://doi.org/10.1093/nar/gkaa1024 (visited on 01/22/2021).

[94] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Dec. 5, 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8 (visited on 11/06/2020).

[95] Gennady Korotkevich, Vladimir Sukhov, and Alexey Sergushichev. "Fast gene set enrichment analysis". In: *bioRxiv* (Oct. 22, 2019). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 060012. DOI: 10.1101/060012. URL: https://www.biorxiv.org/content/10.1101/060012v2 (visited on 11/24/2020).

[96] Arthur Liberzon et al. "The Molecular Signatures Database (MSigDB) hallmark gene set collection". In: *Cell systems* 1.6 (Dec. 23, 2015), pp. 417–425. ISSN: 2405-4712. DOI: 10.1016/j.cels.2015.12.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707969/ (visited on 11/24/2020).

[97] Augustin Luna et al. "PaxtoolsR: pathway analysis in R using Pathway Commons". In: *Bioinformatics (Oxford, England)* 32.8 (Apr. 15, 2016), pp. 1262–1264. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv733.

[98] Dénes Türei et al. "Integrated intra- and intercellular signaling knowledge for multicellular omics analysis". In: *bioRxiv* (Aug. 17, 2020). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.08.03.221242. DOI: 10.1101/2020.08.03.221242. URL: https://www.biorxiv.org/content/10.1101/2020.08.03.221242v2 (visited on 12/08/2020).

[99] Philip Resnik. "Using information content to evaluate semantic similarity in a taxonomy". In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 20, 1995, pp. 448–453. ISBN: 978-1-55860-363-9. (Visited on 11/25/2020).

[100] Vasileios Stathias et al. "LINCS Data Portal 2.0: next generation access point for perturbation-response signatures". In: *Nucleic Acids Research* 48 (D1 Jan. 8, 2020), pp. D431–D439. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1023. URL: https://doi.org/10.1093/nar/gkz1023 (visited on 01/25/2021).

[101] R. M. Haralick, K. Shanmugam, and I. Dinstein. "Textural Features for Image Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (Nov. 1973). Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, pp. 610–621. ISSN: 2168-2909. DOI: 10.1109/TSMC.1973.4309314.

[102] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine". In: *Computer Networks and ISDN Systems*. Proceedings of the Seventh International World Wide Web Conference 30.1 (Apr. 1, 1998), pp. 107–117. ISSN: 0169-7552. DOI: 10.1016/S0169-7552(98)00110-X. URL: http://www.sciencedirect.com/science/article/pii/S016975529800110X (visited on 11/25/2020).

[103] Gábor Iván and Vince Grolmusz. "When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks". In: *Bioinformatics* 27.3 (Feb. 1, 2011). Publisher: Oxford Academic, pp. 405–407. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq680. URL: https://academic.oup.com/bioinformatics/article/27/3/405/321946 (visited on 11/16/2020).

[104] Mariano J. Alvarez et al. "Network-based inference of protein activity helps functionalize the genetic landscape of cancer". In: *Nature genetics* 48.8 (Aug. 2016), pp. 838–847. ISSN: 1061-4036. DOI: 10.1038/ng.3593. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5040167/ (visited on 07/07/2021).

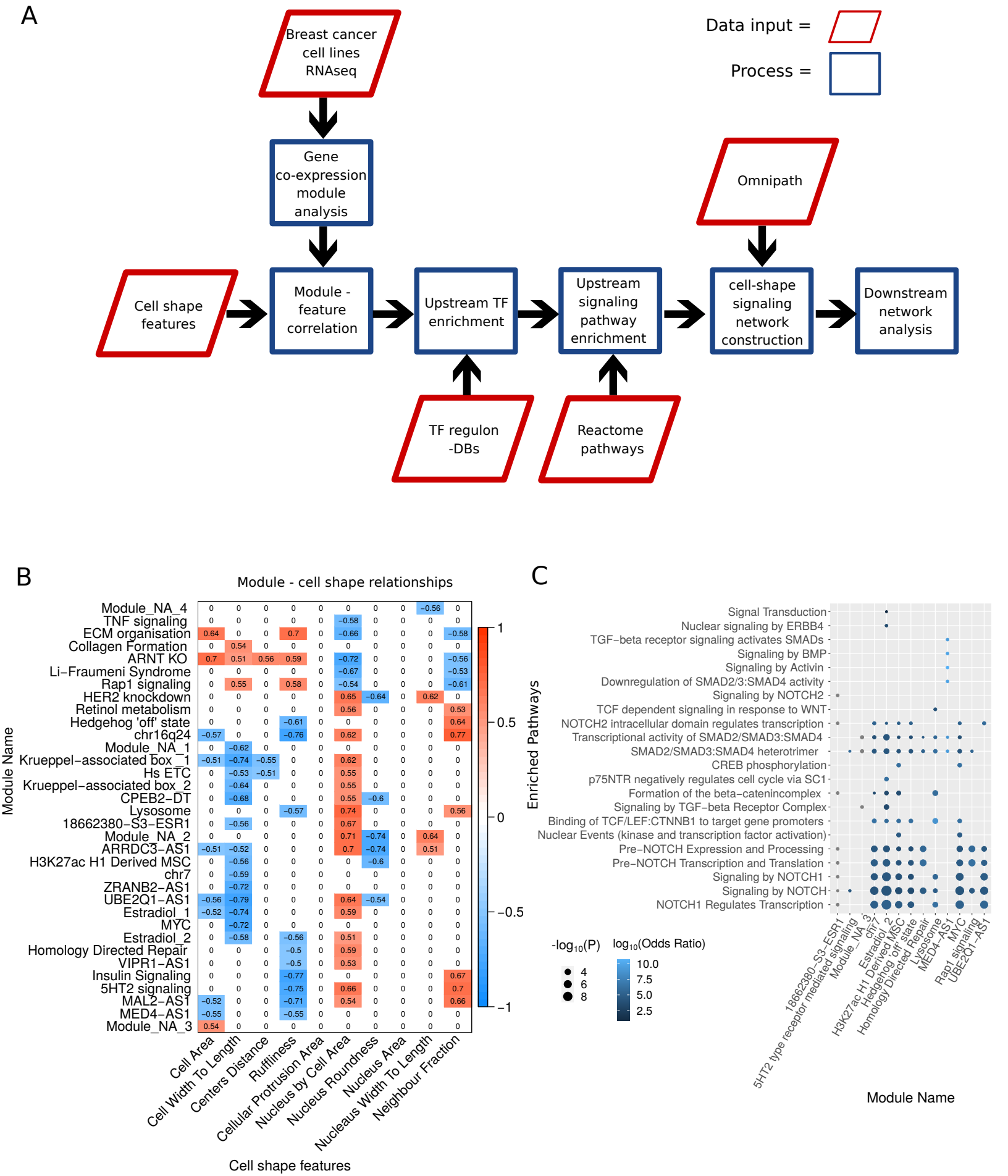
[105] Constantin Ahlmann-Eltze and Wolfgang Huber. "glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data". In: *Bioinformatics* 36.24 (Dec. 15, 2020), pp. 5701–5702. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa1009. URL: https://doi.org/10.1093/bioinformatics/btaa1009 (visited on 06/09/2021).

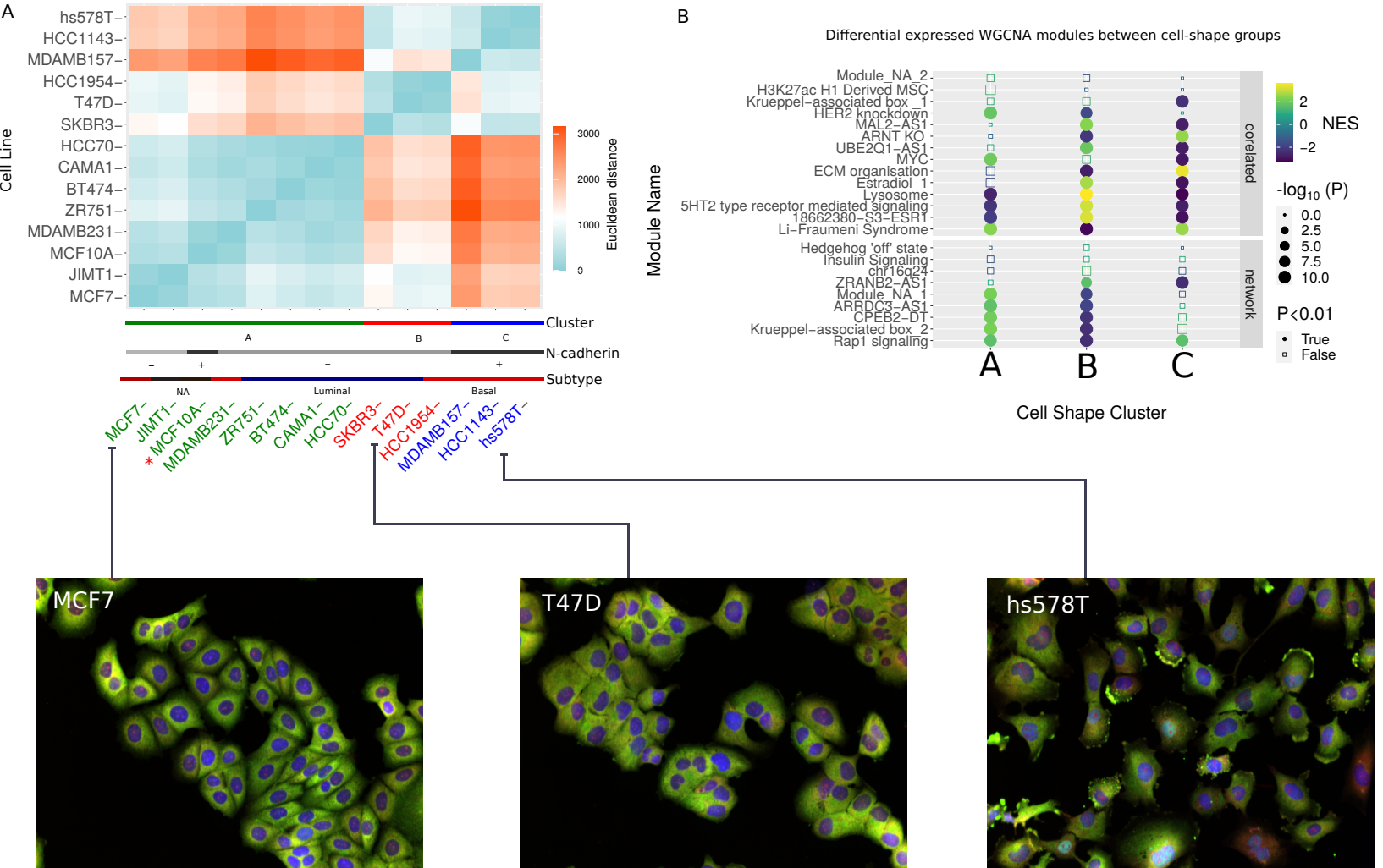
[106] Gabor Csardi and Tamas Nepusz. "The igraph software package for complex network research". In: *InterJournal, Complex Systems* 1695.5 (2006), pp. 1–9. URL: http://igraph.sf.net.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; this version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

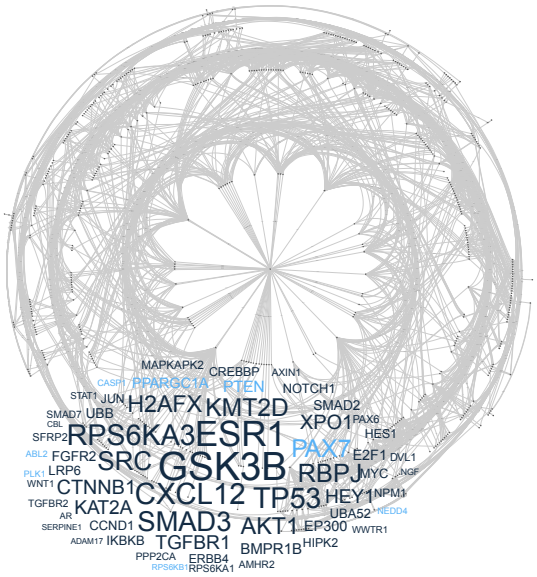
[107] Paul Shannon et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/ (visited on 06/17/2021).

[108] Edward Y. Chen et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". In: *BMC bioinformatics* 14 (Apr. 15, 2013), p. 128. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-128.


A
Data input =
Process =
Breast cancer cell lines RNAseq
Gene co-expression module analysis
Omnipath
Cell shape features
Module - feature correlation
Upstream TF enrichment
Upstream signaling pathway enrichment
cell-shape signaling network construction
Downstream network analysis
TF regulon -DBs
Reactome pathways
B
Module - cell shape relationships
Module Name
Cell shape features
Module_NA_4
TNF signaling
ECM organisation
Collagen Formation
ARNT KO
Li–Fraumeni Syndrome
Rap1 signaling
HER2 knockdown
Retinol metabolism
Hedgehog 'off' state
chr16q24
Module_NA_1
Krueppel–associated box_1
Hs ETC
Krueppel–associated box_2
CPEB2–DT
Lysosome
18662380–S3–ESR1
Module_NA_2
ARRDC3–AS1
H3K27ac H1 Derived MSC
chr7
ZRANB2–AS1
UBE2Q1–AS1
Estradiol_1
MYC
Estradiol_2
Homology Directed Repair
VIPR1–AS1
Insulin Signaling
5HT2 signaling
MAL2–AS1
MED4–AS1
Module_NA_3
Cell Area
Cell Width To Length
Centers Distance
Ruffliness
Cellular Protrusion Area
Nucleus by Cell Area
Nucleus Roundness
Nucleus Area
Nucleaus Width To Length
Neighbour Fraction
1
0.5
0
−0.5
−1
C
Enriched Pathways
Signal Transduction
Nuclear signaling by ERBB4
TGF–beta receptor signaling activates SMADs
Signaling by BMP
Signaling by Activin
Downregulation of SMAD2/3:SMAD4 activity
Signaling by NOTCH2
TCF dependent signaling in response to WNT
NOTCH2 intracellular domain regulates transcription
Transcriptional activity of SMAD2/SMAD3:SMAD4
SMAD2/SMAD3:SMAD4 heterotrimer
CREB phosphorylation
p75NTR negatively regulates cell cycle via SC1
Formation of the beta–catenincomplex
Signaling by TGF–beta Receptor Complex
Binding of TCF/LEF:CTNNB1 to target gene promoters
Nuclear Events (kinase and transcription factor activation)
Pre–NOTCH Expression and Processing
Pre–NOTCH Transcription and Translation
Signaling by NOTCH1
Signaling by NOTCH
NOTCH1 Regulates Transcription
-log10(P)
4
6
8
log10(Odds Ratio)
10.0
7.5
5.0
2.5
5HT2 type receptor mediated signaling
18662380–S3–ESR1
Module_NA_3
chr7
Estradiol_2
H3K27ac H1 Derived MSC
Hedgehog 'off' state
Homology Directed Repair
Lysosome
MED4–AS1
MYC
Rap1 signaling
UBE2Q1–AS1
Module Name

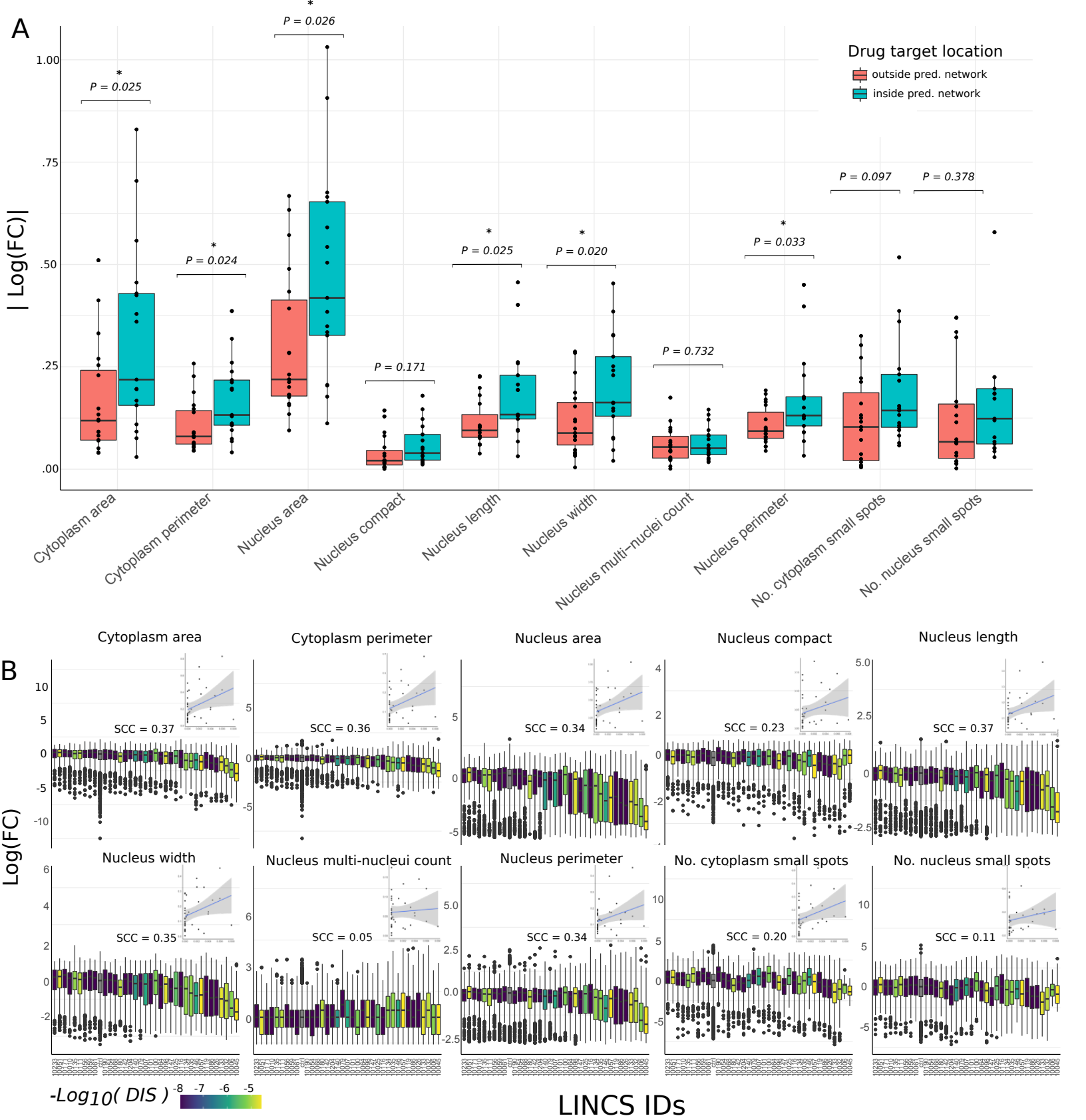

A
Cell Line
hs578T
HCC1143
MDAMB157
HCC1954
T47D
SKBR3
HCC70
CAMA1
BT474
ZR751
MDAMB231
MCF10A
JIMT1
MCF7
3000
2000
1000
0
Euclidean distance
Cluster
A
B
C
N-cadherin
-
+
-
+
Subtype
NA
Luminal
Basal
B
Differential expressed WGCNA modules between cell-shape groups
Module Name
Module_NA_2
H3K27ac H1 Derived MSC
Krueppel-associated box_1
HER2 knockdown
MAL2-AS1
ARNT KO
UBE2Q1-AS1
MYC
ECM organisation
Estradiol_1
Lysosome
5HT2 type receptor mediated signaling
18662380-S3-ESR1
Li-Fraumeni Syndrome
Hedgehog 'off' state
Insulin Signaling
chr16q24
ZRANB2-AS1
Module_NA_1
ARRDC3-AS1
CPEB2-DT
Krueppel-associated box_2
Rap1 signaling
corelated
network
A
B
C
Cell Shape Cluster
2
0
-2
NES
-log10 (P)
0.0
2.5
5.0
7.5
10.0
P<0.01
True
False
MCF7
JIMT1
*MCF10A
MDAMB231
ZR751
BT474
CAMA1
HCC70
SKBR3
T47D
HCC1954
MDAMB157
HCC1143
hs578T
MCF7
T47D
hs578T

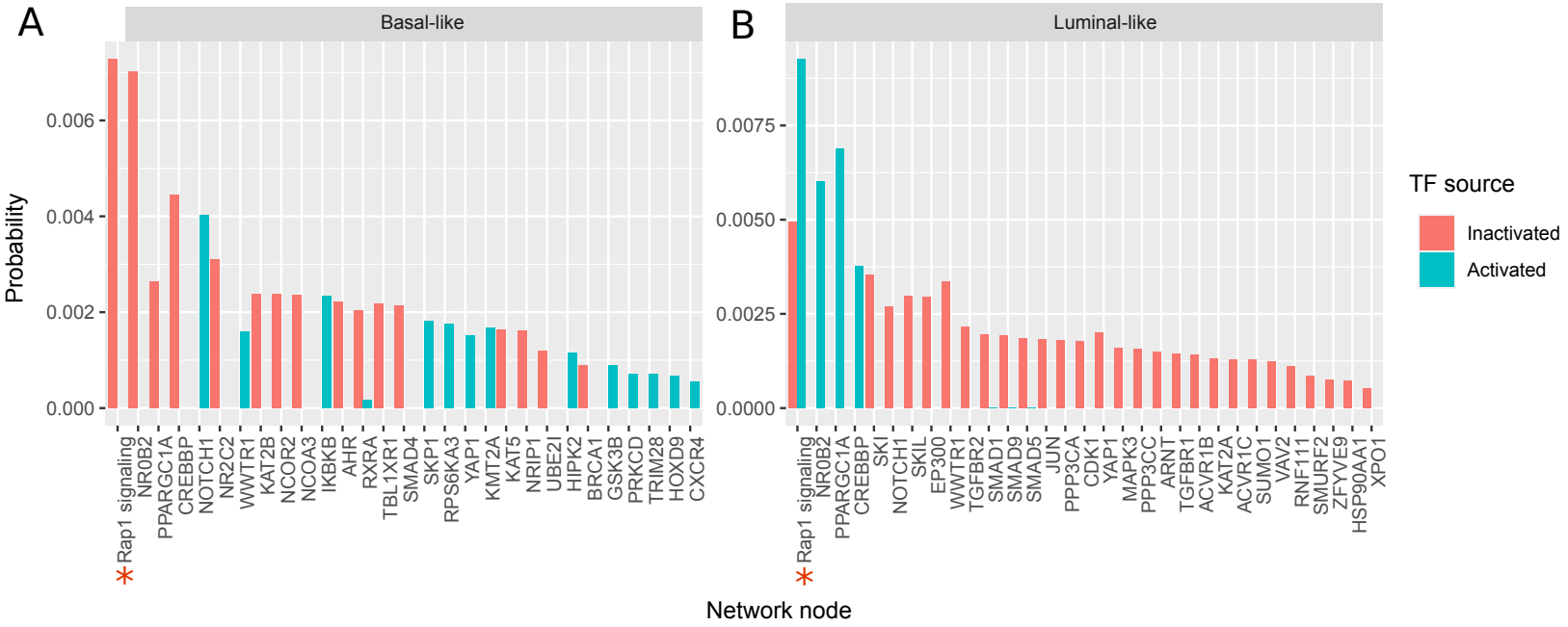
A


MAPKAPK2
CREBBP
AXIN1
CASP8
PPARGC1A
PTEN
NOTCH1
STAT1
JUN
H2AFX
KMT2D
SMAD2
SMAD7
UBB
XPO1
PAX6
CBL
SFRP2
RPS6KA3
ESR1
HES1
PAX7
ABL2
FGFR2
SRC
GSK3B
E2F1
DVL1
PLK1
LRP6
RBPJ
MYC
NGF
WNT1
CTNNB1
CXCL12
TP53
HEY1
NPM1
TGFBR2
KAT2A
UBA52
NEDD4
AR
SERPINE1
CCND1
SMAD3
AKT1
EP300
WWTR1
ADAM17
IKBKB
TGFBR1
BMPR1B
HIPK2
PPP2CA
ERBB4
AMHR2
RPS6KB1
RPS6KA1

B

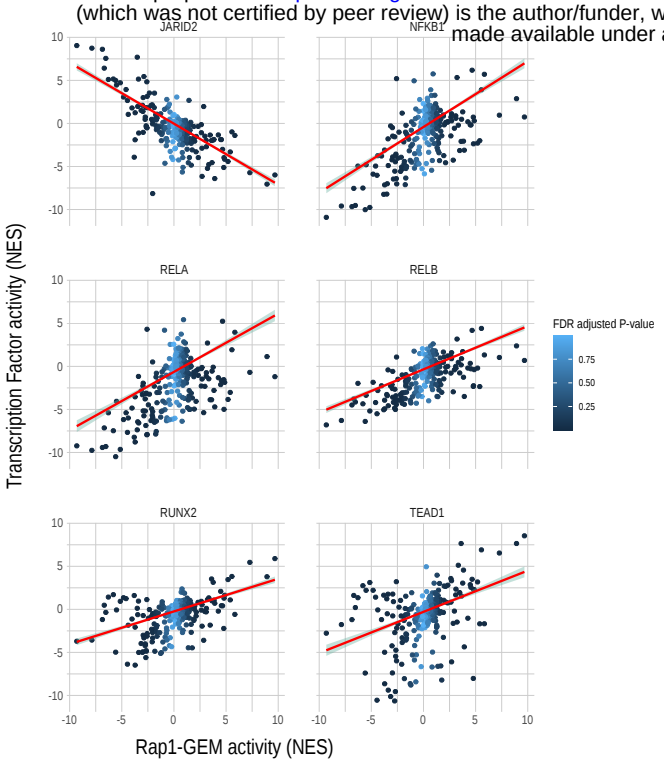
Enriched Term
Notch signaling pathway
TGF−beta signaling pathway
Alzheimer disease−presenilin pathway
Wnt signaling pathway
B cell activation
Interferon−gamma signaling pathway
CCKR signaling map ST
Insulin/IGF pathway−mitogen activated protein kinase kinase/MAP kinase cascade
Inflammation mediated by chemokine and cytokine signaling pathway
VEGF signaling pathway
Alzheimer disease−amyloid secretase pathway
PI3 kinase pathway
FAS signaling pathway
Angiogenesis
Insulin/IGF pathway−protein kinase B signaling cascade
EGF receptor signaling pathway
Ras Pathway
Apoptosis signaling pathway
p53 pathway feedback loops 2
Endothelin signaling pathway
T cell activation
Parkinson disease
Histamine H1 receptor mediated signaling pathway
Oxidative stress response
p53 pathway
PDGF signaling pathway
Ubiquitin proteasome pathway
Oxytocin receptor mediated signaling pathway
Hypoxia response via HIF activation
FGF signaling pathway
Interleukin signaling pathway
p38 MAPK pathway
p53 pathway by glucose deprivation
Hedgehog signaling pathway
5HT2 type receptor mediated signaling pathway
Cadherin signaling pathway
Thyrotropin−releasing hormone receptor signaling pathway
Angiotensin II−stimulated signaling through G proteins and beta−arrestin
Transcription regulation by bZIP transcription factor
Toll receptor signaling pathway
Integrin signalling pathway
0
1000
2000
3000
Combined score (P value + Z-score)
Term source
PCSF
SEED
-log10(P)
10
20
30
40
PCSF terms
Seed terms
14
16
4


A
Drug target location
outside pred. network
inside pred. network
| Log(FC)|
1.00
.75
.50
.25
.00
P = 0.025
P = 0.024
P = 0.026
P = 0.171
P = 0.025
P = 0.020
P = 0.732
P = 0.033
P = 0.097
P = 0.378
Cytoplasm area
Cytoplasm perimeter
Nucleus area
Nucleus compact
Nucleus length
Nucleus width
Nucleus multi−nuclei count
Nucleus perimeter
No. cytoplasm small spots
No. nucleus small spots
B
Log(FC)
SCC = 0.37
SCC = 0.36
SCC = 0.34
SCC = 0.23
SCC = 0.37
SCC = 0.35
SCC = 0.05
SCC = 0.34
SCC = 0.20
SCC = 0.11
-Log10( DIS )
-8 -7 -6 -5
LINCS IDs

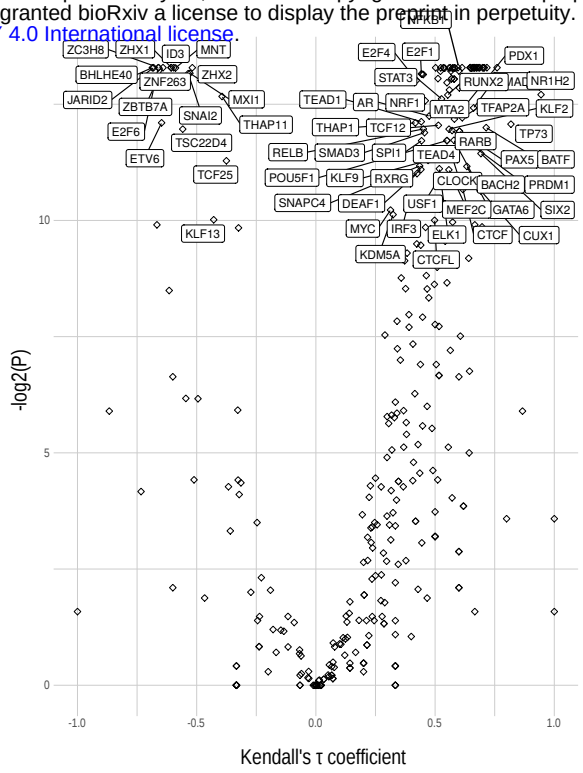

A
Basal-like
Probability
0.006
0.004
0.002
0.000
Rap1 signaling
NR0B2
PPARGC1A
CREBBP
NOTCH1
NR2C2
WWTR1
KAT2B
NCOR2
NCOA3
IKBKB
AHR
RXRA
TBL1XR1
SMAD4
SKP1
RPS6KA3
YAP1
KMT2A
KAT5
NRIP1
UBE2I
HIPK2
BRCA1
GSK3B
PRKCD
TRIM28
HOXD9
CXCR4
B
Luminal-like
0.0075
0.0050
0.0025
0.0000
Rap1 signaling
NR0B2
PPARGC1A
CREBBP
SKI
NOTCH1
SKIL
EP300
WWTR1
TGFBR2
SMAD1
SMAD9
SMAD5
JUN
PPP3CA
CDK1
YAP1
MAPK3
PPP3CC
ARNT
TGFBR1
ACVR1B
KAT2A
ACVR1C
SUMO1
VAV2
RNF111
SMURF2
ZFYVE9
HSP90AA1
XPO1
Network node
TF source
Inactivated
Activated

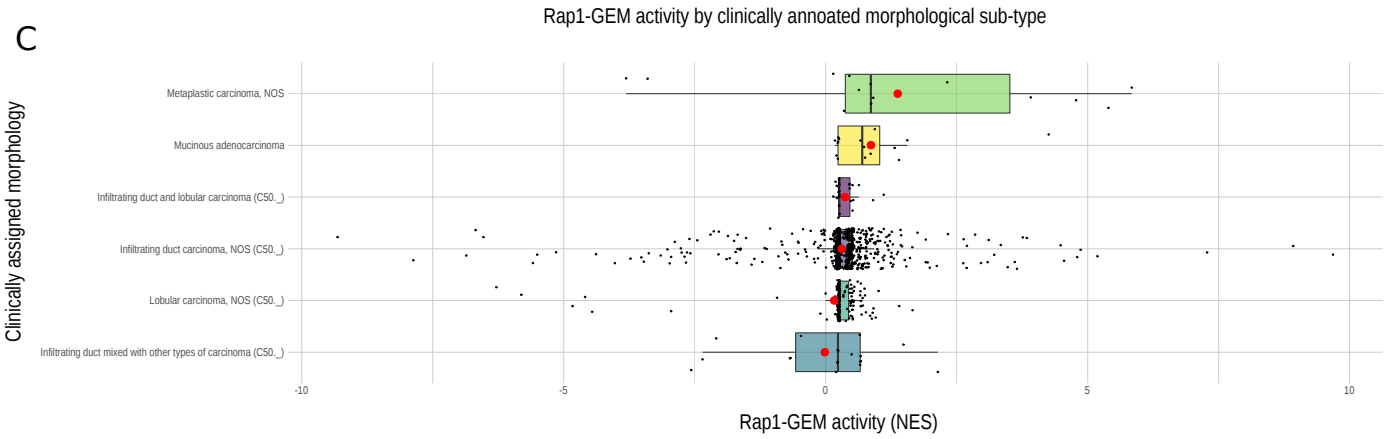
bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.430597; This version posted July 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A




B




C