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

*correspondence should be addressed to 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 signalling pathway activity. Delineating the interface between morphology and signalling 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 signalling sub-networks 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 signalling 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 NF-kB signalling by numerous kinase and transcriptional regulators. Further analysis of our network implicates a gene expression module enriched in the RAP1 signalling pathway as a mediator between the sensing of mechanical stimuli and regulation of NF-kB 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 sub-type cancer and predict patient prognosis (Wu et al., 2020). 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 (Baskaran et al., 2020; Krakhmal et al., 2015). 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 (Cooper and Giancotti, 2019). Cell morphology and tumour organisation have been found to be a factor in modulating the intra-cellular signalling state through pathways able to integrate mechanical stimuli from the extracellular environment (Miralles et al., 2003; Olson and Nordheim, 2010; Orsulic et al., 1999; Zheng et al., 2009). The discovery of mechanosensitive pathways in various tissues has revealed a complex interplay between cell morphology and signalling (Kumar et al., 2016). 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 (Wu et al., 2020), 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 (Roche, 2018).

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% (Siegel et al., 2019). During the development of breast cancer tumours, cells undergo progressive transcriptional and morphological changes that can ultimately lead towards EMT and subsequent metastasis (Feng et al., 2018; Lee et al., 2015; Wu et al., 2020). Breast cancer sub-types 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 (Fedele et al., 2017) with fewer cell-to-cell contacts (Dai et al., 2015). Luminal tumour subtypes on the other hand, are associated with good to intermediate outcomes for patients (Dai et al., 2015) and have a clear epithelial (or 'cobblestone') morphology with increased cell-cell contacts (Neve et al., 2006). 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.

Signalling and transcriptomic programs are known to be modulated by external physical cues in the contexts of embryonic development (Wozniak and Chen, 2009), stem-cell maintenance (Bergert et al., 2020; De Belly et al., 2020) and angiogenesis (Chatterjee, 2018). Numerous studies have flagged NF-kB as a focal point for mechano-transductive pathways in various contents (Cowell et al., 2009; Ishihara et al., 2019; Shrum et al., 2009; Tong and Tergaonkar, 2014), 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 NF-kB activation by combining high-throughput image analysis of breast cancer cell lines with network modelling (Sero et al., 2015). They found a relationship between cell shape, mechanical stimuli and cellular responses to NF-kB 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 (Sailem and Bakal, 2017), 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 NF-kB 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 (Vogel and Marcotte, 2012). 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 signalling 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 (Bougault et al., 2012; Horton et al., 2016; Robertson et al., 2015; Rolfe et al., 2014). Furthermore, how exactly extracellular mechanical cues are 'sensed' by the cell and passed on to NF-kB 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 signalling.

Here we have developed a powerful network-based approach to bridge the gap between widely available and cheap expression data, signalling events and large-scale biological phenotypes such as cell shape (Figure 1A). Our study aims to identify a data-derived cellular signalling network, specific to the regulation of cell shape beyond NF-kB, by considering functional co-expression modules and cell signalling processes rather than individual genes.

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) (Langfelder and Horvath, 2008) 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 (Sailem and Bakal, 2017) (Methods). These described the size, perimeter and texture of the cell and the nucleus (n = 75,653). Of 102 GEMs (Supplemental Figure S1A), 34 were significantly correlated ($P<0.05$; Student's *t*-test, Pearson's Correlation; Supplemental Table S1; Supplemental Figure S1B-C) with one of 8 cell shape features (Figure 1B). A full list of the genes within the identified modules is presented in Supplemental Table S2.

We used Enrichr and their suite of gene set libraries (Kuleshov et al., 2016) to functionally annotate and label some of the modules using enrichment of genes contained within them. We found that the 'RAP1 signalling' module is also enriched for terms such as VEGF signalling and hemostasis, while the 'Insulin signalling' 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 signalling (Supplemental Table S3). Modules that are most correlated with all features are the '*ARNT* KO' module, 'ARRDC3−AS1' module and the 'ECM organisation' module (Supplemental 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 signalling pathways that regulate them

To link these expression modules to the intra-cellular signalling network, we considered both the regulation of modules as transcriptional units as well as the signalling pathways that significantly

regulate the identified regulons. Specifically, we first found 17 transcription factor (TF) regulons, as defined in the database TRRUST v2 (Han et al., 2018), to be significantly enriched (P<0.1; Fisher's exact test) in our modules (Supplemental Table S4). 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 signalling activity in these cells. These TFs include the EMT antagonist FOXA1 (Song et al., 2010), and HOXB7 (Wu et al., 2006) and ZFP36 (Van Tubergen et al., 2013).

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 (Lachmann et al., 2010), DNA binding preferences from JASPAR (Stormo, 2013; Wasserman and Sandelin, 2004), TF protein-protein interactions and TFs from ENCODE ChIP-seq (Euskirchen et al., 2007) to get a more comprehensive picture of the pathways involved in regulation of cell morphology. Using the identified TFs (Supplemental Table S5) we then used Enrichr (Kuleshov et al., 2016) to perform a Reactome signalling pathway (Jassal et al., 2020) enrichment analysis. Results from this analysis showed that 6 modules shared pathways associated with downstream signalling and regulation of NOTCH (Figure 1C). To ensure that our approach is not biased to any particular pathway, we repeated our approach on 1,000 resampled GEMs, and created pathway-specific null distributions for each identified pathway. All pathways we identified from morphology-correlated modules had significantly lower p-values than randomised modules (FDR adjusted P<0.05). The only exceptions were one association with "signalling by NOTCH" and modules associated with "Signal Transduction", a spurious pathway containing the complete intra-cellular signalling system (Supplemental Table S6).

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 10 morphological features including area, ruffliness, protrusion area and neighbour frequency and performed differential expression analysis between the identified clusters (Figure 2A; Supplemental Figure S2A). 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 except for 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. We found that these groups are closely aligned with the expression of the cell adhesion protein, CDH2 (also known as N-cadherin, Figure 2A), the expression for which is closely associated with a migratory and metastatic phenotype (Shih and Yamada, 2012). Representative images of the morphologically clustered cell lines are shown along with the clustering heatmap in Figure 2A (complete dataset of images provided online; https://datadryad.org/stash/dataset/doi:10.5061/dryad.tc5g4).

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 Supplemental Table S7, 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 down-regulated genes include 'fatty acid and beta-oxidation' and 'ERBB network pathway'. In the genes up-regulated in the luminal-like cluster, we observed enrichment of terms such as 'hallmark-oxidative phosphorylation'. down-regulated 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 signalling pathway' (Supplemental Table S7).

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 signalling module is up-regulated in basal-like clusters and down-regulated 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 (Dai et al., 2015). 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-signalling by NF-kB) and ECM organisation module (enriched in focal adhesion proteins - see Supplemental Table S3).

To link the observed gene expression differences (Supplemental Figure S2B-C) to cell signalling we used the tool DOROTHEA (Garcia-Alonso et al., 2019) to calculate transcription factor activities, as their modulation is one of the main results of cell signalling 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 MSC and HOXA9. down-regulated regulomes include ZEB2, MYC, ESRRA and KLF5 (Supplemental Figure S2D).

Assembly of a data-driven cell shape regulatory network

To integrate our data-driven GEMs with signalling pathways, we used the Prize Collecting Steiner Forest (PCSF) algorithm (Akhmedov et al., 2016). 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 signalling. The network used for this process was extracted from the database OmniPath (Türei et al., 2016) to provide a map of the intracellular signalling network described as a signed and directed graph. We incorporated identified GEMs into the network by interlinking them as nodes with the relevant TFs and signalling pathways.

The resulting network of 691 nodes included 97.11% of the genes identified by our analysis. The new proteins that were included by the PCSF algorithm to maximise prize collection showed gene set enrichment of common terms (Pathways from Panther (Mi et al., 2013)) 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) ($P<0.001$; Supplemental Figure S3A).

Studying the network properties of our PCSF-derived regulatory network we find that the degree distribution is typical for a biological network (Supplemental Figures S3B-D). 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, TP53, SMAD3 - Supplemental Figure S3E) indicating that the PCSF solution was not achieved by the inclusion of new hub proteins that are not of interest to our analysis. Nevertheless, a small minority of high centrality nodes were not in the original prizes, implicating them as mediating the cross talk between pathways identified in Figure 1C. These include the proteins 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 (Subramanian et al., 2017). 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 (Hamilton et al., 2007). 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 (Moret et al., 2019) 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. We found 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 (Figure 3A; Supplemental Figure S4A). 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, MCF-7 and non-tumorigenic mammary cell line MCF-10A) with results with limited statistical significance (Supplemental Table S8). We additionally used a positive control where the control cells had been treated with TRAIL (TNF-related apoptosis-inducing ligand) to ensure that the observed morphological effects were not caused by apoptotic factors (Supplemental Figure S4B, Supplemental Table S8).

We found that there is greater variance in the effect size for kinase inhibitors targeting proteins contained within the predicted regulatory network than those outside. The individual effect on cell morphology for each drug is shown in Supplemental Figure S4A-B. 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 3B 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's 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's 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.

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 signalling that can be extracted from transcriptomics data (Szalai and Saez-Rodriguez, 2020), we applied network propagation to identify sub-networks and nodes of which differentially regulated transcription factors have an effect. The algorithm Random Walk with Restart (RWR (Tong et al., 2008)) was used to diffuse from activated and inactivated transcription factors in our network reflected by the normalised enrichment scores of transcription factors identified by DOROTHEA (Garcia-Alonso et al., 2019) (Figure 4A & B; Supplemental Table S9; Methods).

The most relevant super-node in both luminal and basal diffusions was the gene expression module, RAP1 signalling, 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 signalling pathway. By performing RWR diffusions on each of the seed nodes separately (Supplemental Figure S5A-B) we can see that the source of this module's probability is mainly from the transcription factors JARID2 and RUNX2 in luminal-like cell lines, and JARID2 for basal-like. However, the transcription factors KLF5 and ESRRA in both morphological subtypes also contribute to the ranking of RAP1 signalling, via GSK3B and DVL1 (Figure 4C).

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 TGFB, IKBKB, which is an activator of NF-kB and YAP1. 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 and CREBBP.

The 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 (Holland et al., 2020) (Supplemental 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 NFKB1. 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 5A-B). The intersection of this analysis between in cell lines and the clinical data were the TFs: SP3, NFKB1, ZNF589, ZC3H8, HIF1A, STAT1, ZNF584, ZNF175 and KLF5.

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 (Schwartz et al., 2013) and a high propensity to metastasize (Reddy et al., 2020) (Figure 5C). 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 (Makki, 2015).

Content of RAP1 GEM and its network-neighbourhood shed light on signalling 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 (GO) terms enriched within RAP1 GEM (Figure 5D). This revealed that, as well as RAP1 signalling, the GEM is enriched in AGE-RAGE signalling pathway and HIF1 signalling pathway (consistent with HIF1A's activity correlating highly with RAP1 GEM in both cell line patient data). HIF1A is known to be regulated downstream of RAP1 (Li et al. 2021; Menon et al. 2012), although not explicitly in breast cancer.

NF-kB has been previously linked to the regulation of cell shape in breast cancer. To explore the interface of RAP1 GEM with NF-kB in terms of intra-cellular signalling, we identified a subnetwork of our network responsible for mediating 'information-flow' between those two nodes, using the algorithm maximum flow (Figure 5E). By studying the flow of information from RAP1 signalling, we can see that a LATS2/WWTR1/DVL1 (all of WNT signalling) 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 5B and Supplemental Figure S6A).

Discussion

We present a method that uses transcriptomics and phenotypic data to derive a concise sub-network describing the signalling involved in the regulation of cell shape. This analysis

recovered known processes like ‘adherens junction proteins’, ‘cadherin’ (Eslami Amirabadi et al., 2019) and ‘integrin’ (Filer and Buckley, 2013; Taherian et al., 2011) as well as pathways responsible for the regulation of cell shape in development, such as WNT (Kadzik et al., 2014; Wildwater et al., 2011), TGFB (Lee et al., 2013) and NOTCH (Kontomanolis et al., 2018). All of these pathways have previously been linked to the development of metastatic phenotypes in breast cancer cells (Imamura et al., 2012; Kontomanolis et al., 2018; Yin et al., 2018). Moreover, individual transcription factors identified include the known promoters of metastasis, SOX2 (Liu et al. 2018, 2; Li et al. 2019, 2), HOXA9 (Ko and Naora 2014) and ESRRA (Berman et al. 2017). Also, among these transcription factors were known regulators of cell shape and EMT, including KLF5 (Chen et al. 2015), ZEB2 (DaSilva-Arnold et al. 2019) and MYC (Cowling and Cole 2007; Lourenco et al. 2019).

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 signalling, 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 up-regulated in basal-like cell lines while down-regulated 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 signalling network, being at the network confluence of multiple activated transcription factors. 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, as well as those used to identify it in the network propagation analysis. In this way, our methodology uses cell line data for network construction and validation, but through our network approach we focus on more general effects which can be tested and successfully validated in a wider breast cancer clinical context. Hence, we believe these results to be relevant in more general breast cancer applications, but are also reflecting the inherent context-specificity that exists in biology.

The name-sake of our identified module, 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 (Boettner and Van Aelst, 2009; Zhang et al., 2017). Specifically, RAP1 has been shown to modulate and activate NF-kB activity in response to TNFA stimulation in mesenchymal stem cells (Zhang et al., 2015) and modulate migration and adhesion (Sawant et al., 2018). RAP1 is able to regulate IKKs (IκB kinases) in a spatio-temporal manner (Ohba et al., 2003), and is crucial for IKBK to be able to phosphorylate NF-kB subunit RELA to make it competent (Teo et al., 2010). Here, we used our network-centric methodology to highlight a transcriptomic module, characterised in part by RAP1 signalling and that this is a key node in our phenotype-specific signalling network. 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 characteristic of a more robust breast cancer niche.

The RAP1 signalling GEM identified in the network analysis represents a subset of the transcriptome observable among our analysed cell lines. While it is enriched in RAP1 signalling, 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 signalling pathways in order to modulate transcription factor activity and consequently cell shape. Other pathways enriched in the expression module include HIF1 signalling pathway, which is known to be activated by RAP1 in melanoma (Lee et al., 2015a), but this has not been shown in breast cancer. HIF1 (Hypoxia-

inducible factor 1) is also of special relevance in tumorigenesis because hypoxia is one of the key stimuli that a cancer cell is able to process in order to determine its fate and maintain the cancer stem cell niche (Plaks et al., 2015). AGE-RAGE signalling was also enriched in our module of interest. AGE-RAGE signalling pathway has recently been shown to overlap with RAP1 signalling pathway in cardiac fibroblasts to alter the expression of NF-kB (Burr and Stewart, 2021), although this crosstalk has also not been illustrated in breast cancer. Here, we observe genes of RAP1 signalling and AGE-RAGE functioning as a cohesive unit while also being correlated with NF-kB activity. The combined enrichment of AGE-RAGE, HIF1 and RAP1 signalling is of particular interest because it implies a novel interaction between these three processes, common to all our cell lines, in a manner that has not been previously described in breast cancer.

We also observe that our gene expression module of interest is significantly correlated with NF-kB in both clinical samples and cell culture. Other authors have flagged the direct effect of RAP1 on the cytoskeleton and NF-kB (Mun and Jeon, 2012; Zhang et al., 2015), but here we go further, using our unbiased systems approach to link RAP1 signalling 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 hypothesise 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 (Roy Choudhury et al., 2019).

Previous studies have established a connection between the NF-kB signalling pathway and regulation of cell shape in breast cancer (Sailem and Bakal, 2017; Sero et al., 2015). Our findings also illustrate the significance of this pathway in the regulation of cell morphology, with multiple NF-kB 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 up-regulated 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, indicating a strong association with cell shape. By studying terms enriched in this module from the Enrichr library, we find both ‘TNF-alpha signalling via NF-kB’ to be enriched as well as genes down-regulated during AHR nuclear translocator (*ARNT*) shRNA KO. Signalling by TNFA is able to activate NF-kB, a transcription factor known to control the expression of many EMT related genes (Pires et al., 2017) which has shown to be more sensitive to TNFA 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 (Sero et al., 2015). Here we observe also that an *ARNT* KO/TNFA module is up-regulated 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) (Huang et al., 2015). Previous studies have also shown its ability to modulate NF-kB signalling with the activated form possibly interfering with the action of activated RELA (Øvrevik et al., 2014). Our findings that the upregulation of a gene expression module that is associated with *ARNT* knockdown further gives credence to NF-kB being positively regulated in mesenchymal-like cell morphologies. Furthermore, the results of our network propagation yielded activators and transcriptional coactivators of NF-kB (IKBKB (Teo et al., 2010), NR0B2 (Zou et al., 2015) and CREBBP (Bhatt and Ghosh, 2014)). These findings indicate that NF-kB is modulated by both phosphorylation (through stimulation by TNFA), spatial-temporal location (through RAP1) and transcriptional co-activation (through NR0B2 and CREBBP) in breast cancer in a shape-dependent manner.

Aside from the biological findings of this study, we illustrate an approach for network analysis of a specific course-grained phenotype through expression; a notoriously poor (if cheap and widely

available) proxy for gauging intracellular signalling (Piran et al., 2020). In contrast to existing methods that use gene expression as a direct proxy for signalling (Ben-Hamo et al., 2014; Guan et al., 2012; Padi and Quackenbush, 2018; Soul et al., 2015), our approach infers transcription factor activities from the expression data and uses these as an anchor to infer upstream signalling 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 signalling activities of the cell (Szalai and Saez-Rodriguez, 2020). Such an approach has been previously used, for example by the tool CARNIVAL (Liu et al., 2019). However, this and other available tools neglect the propensity for the transcriptome to be regulated in a highly context specific and modular structure (Kitano, 2002; Sharma and Petsalaki, 2019). 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 signalling 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 unbiased latent modular structures 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 cell; the transcriptome and intracellular signalling. Both can be easily combined with complex, multivariate phenotypic data which here has revealed a clearer picture of how signalling 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 signalling. 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 signalling and gene regulation in the context of breast cancer.

Methods

WGCNA analysis

Using Weighted correlation network analysis, we performed co-expression module identification using the R package WGCNA (Langfelder and Horvath, 2008). 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 and with the alignment performed to the NCBI Human Reference Genome GRCh37 (Papatheodorou et al., 2020). We collated 13 breast cancer and 1 non-tumorigenic cell line for which imaging data was available (BT474, CAMA1, T47D, ZR75.1, SKBR3, MCF-7, HCC1143, HCC1954, HCC70, hs578T, JIMT1, MCF10A, MDAMB157 and MDAMB231 (Sero et al., 2015)). We acquired representative images of each cell line from Sero et al., 2015 (https://datadryad.org/stash/dataset/doi:10.5061/dryad.tc5g4).

Cell imaging segmentation was performed using Acapella software (PerkinElmer) with an automated spinning disk confocal microscope. The presented images (Figure 2) are taken from the above link, stained with DAPI (blue), Alexa 488 (green) and DHE (red). Using Ensembl-BioMart, we filtered genes to only include protein-coding genes (Kinsella et al., 2011) and genes whose FPKM was greater than 1, leaving a total of 15,304 genes.

We created a signed, weighted adjacency matrix using $\log_2$ transformed gene expression values and a soft threshold power (beta) 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). Jack knife cross-validation was used to assess the robustness of the identified modules to the removal of different cell lines from the analysis (Supplemental Figure S1C) and all showed a high degree of conservation between resampled runs.

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

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 (Sero et al., 2015). We correlated these features with module eigengenes using Pearson's correlation and we tested these values for significance by calculating Student asymptotic p-values for given correlations. Multiple hypothesis testing was performed using a permutation based procedure whereby we recalculated the correlation matrix 1,000 times with resampled data. We then generated null distributions for each ranked correlation statistic in our matrix, and compared them to our real data of the same rank. We include in the Supplemental Table S1 confidence intervals of our permutation-based multiple-correction procedure. For the modules that correlated with morphological features (Pearson Correlation Coefficient 0.5 and Student $P<0.05$), we identified enriched signalling pathways using the R package Enrichr (Kuleshov et al., 2016), and the signalling database Reactome (Jassal et al., 2020). Reactome was used in preference to other pathway databases, because of the more consistent inclusion of TFs within the annotated pathways. Using the database TRRUST v2 (Accessed : 01/07/18)(Han et al., 2018), we identified TF regulons that significantly overlap (Fisher's exact 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 (Supplemental Table S3). 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 signalling databases of KEGG, Reactome, Panther and Wikipathways (Accessed : 01/04/20)(Kanehisa et al., 2016; Martens et al., 2021; Mi et al., 2013). 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 Supplemental Figure S2A). We performed differential expression analysis using the R package DESeq2 (Love et al., 2014). 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 $\log_2$ fold changes with the cluster of interest as the numerator and the remaining cell lines acting as a control. Using the R package FGSEA (Korotkevich et al.,

2021), we performed gene set enrichment analysis of the differentially regulated proteins using the complete pathways gene set (Release 01 April 2020) from MSigDB (Liberzon et al., 2015) and the WGCNA gene expression modules identified in previous analysis. We calculated transcription factor regulon enrichment using the software DOROTHEA (Accessed: 01/04/20)(Holland et al., 2020)

Network Generation

Using a Prize Collecting Steiner Forest (PCSF) algorithm, we generated a cell-shape regulatory network implemented through the R package PCSF (Akhmedov et al., 2016). For the prize-carrying nodes to be collected by the PCSF algorithm, we used the transcription factors significantly regulating the WGCNA modules using TRRUST v2 ($p<0.1$), the differentially activated transcription factors identified by DOROTHEA ($p<0.1$), and the signalling 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 (Lachmann et al., 2010)**,** DNA binding preferences from JASPAR (Stormo, 2013; Wasserman and Sandelin, 2004), TF protein-protein interactions and TFs from ENCODE ChIP-seq (Euskirchen et al., 2007). 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 (Luna et al., 2016). This was tested for bias to specific pathways by generating pathway-specific null distributions from 1,000 resampled GEMs. Distributions of p-values for each Reactome pathway were generated, where failed tests (because of no TF enrichment) were given a p-value of 1. Results of this were corrected for multiple-hypothesis testing using FDR correction.

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 (Accessed : 06/05/20) (Türei et al., 2021), extracted using the R package OmnipathR (Türei et al., 2016). We set each prize for significant TFs or signalling 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 v2 (Han et al., 2018)) 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 signalling, 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 signalling proteins and thus account for feedback between cell shape signalling 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 (Mi et al., 2013) and GSE package Enrichr (Kuleshov et al., 2016).

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 (Resnik, 1995) between the biological process GO terms of the two interacting pairs, with the sign of the interaction being inherited from Omnipath (Türei et al., 2016). 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 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. We generated a graphical representation of the network edges and TFs responsible for the ranking of RAP1 signalling by plotting all the shortest paths between RAP1 and the TFs that caused it to have a non-zero affinity score when each TF was considered in isolation.

Breast cancer cell morphology following kinase inhibitor treatment

We used single cell, small molecule kinase inhibition data from Harvard Medical School (HMS) Library of Integrated Network-based Cellular Signatures (LINCS) Center (Stathias et al., 2020), which is funded by NIH grants U54 HG006097 and U54 HL127365 (available from: https://lincs.hms.harvard.edu/mills-unpubl-2015/, Accessed: 01/08/20). 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-throughput image analysis (Hamilton et al., 2007; Haralick et al., 1973). 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/, Accessed: 01/08/20). 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 10µm drug treatment relative to the mean of the control set (untreated and DMSO treated cells). Spearman's rank correlation was calculated between the drug target's network centrality and the absolute log fold change of the morphological variable. We also used the Kolmogorov-Smirnov statistic to assess significance between cell morphology after treatment with drugs targeting kinases inside versus

outside our predicted network. This was also repeated on other breast cancer cell lines and using a TRAIL (apoptosis inducing) control (Supplemental table S8).

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:

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

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 (Brin and Page, 1998). This centrality measure has been shown to be effective in prioritizing proteins by relative importance in signalling or protein-protein interaction networks (Iván and Grolmusz, 2011). We used this statistic as a measure of a kinase inhibitor's influence on cell shape.

Analysis of BRCA cell line and TCGA sample RNA-seq data

For the cell lines, we used RNA-seq data from Expression Atlas (in FPKM - E-MTAB-2770 and E-MTAB-2706) (Papatheodorou et al., 2020). This was analysed using DESeq2 (Love et al., 2014) as per the methodology in the subsection entitled “Clustering and differential expression”. Both TF and module activity was calculated using the algorithm VIPER (Alvarez et al., 2016). 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, Accessed: 01/04/21). For computational efficiency, we use Gamma-Poisson models to predict differentially expressed genes from our samples using the package glmGamPoi (Ahlmann-Eltze and Huber, 2020). 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 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 (Resnik, 1995) as the maximum 'carrying capacity' of an edge in the network. To visualise the optimised solution (as implemented by the R package igraph (Csardi and Nepusz, 2006)) we selected only those edges in the 99th percentile of the flow-carrying edges in the network. Visualisation was performed using the software, Cytoscape (Shannon et al., 2003). Maximum flow was performed with the R package igraph (Csardi and Nepusz, 2006).

Quantification and Statistical Analysis

Statistical tests were performed in base R (R Core Team 2021) unless otherwise mentioned in the methods and p-value cut-offs are shown in parentheses after reporting an effect as significant. Weighted Pearson's 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's exact test was used to test significance of overlap between TRRUST regulons and identified gene expression modules (Supplemental Table S4 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 (Chen et al., 2013)) calculates a score combining the Fisher's 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 τ-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 (Love et al., 2014) 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 ($P<0.1$). Figures were presented using ggplot (Wickham 2009).

Software Availability

The complete R scripts and data used for this methodology are available on Gitlab (https://gitlab.ebi.ac.uk/petsalakilab/phenotype_networks) and as a Supplemental Code file.

Competing Interests Statement

The authors declare no competing interests.

Acknowledgments

We would like to thank the European Molecular Biology Laboratory (EMBL), and EMBL-EBI for funding the project. CGB was funded by the EMBL International PhD Programme. 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**.**

Figure Legends

Figure 1. Overview of workflow and resultant gene expression modules and pathways. 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 RNA-seq 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 Supplemental Table S3) and the y axis shows the signalling pathways found to be significantly ($P<0.01$) enriched in the TFs that regulate the given module (as defined by Supplemental Table S5). 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 $-\log_{10}(P)$ and the colour indicates a log 10 transformation of the odds ratio.

Figure 2. Clustering breast cancer cell lines into groups of similar morphology. 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 ($-\log_{10}(P)$) 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. Colours indicate labeling with DAPI (blue), Alexa 488 (green) and DHE (red).

Figure 3. Effect of drug perturbation of derived network on breast cancer cell line morphology. 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's correlation coefficients are shown above the inset plots.

Figure 4. Network propagation of active transcription factors within cell shape network. A, B 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 Supplemental Table S9. **C.** Sub-networks illustrating the paths between activated transcription factors (in basal-like and luminal-like) and the 'RAP1 signalling' gene expression module. Transcription factors are shown as diamond-shaped nodes, with their colour representing their activity. The 'RAP1 signalling' gene expression module is shown as a grey rectangle. Signalling proteins are shown as black nodes.

Figure 5. Expression of RAP1 gene expression module in further breast cancer cell lines, and in clinical samples. 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, NF-kB1, 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 $-\log_2$(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. **D.** 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 ($-\log_2(P)$). **E.** Sub-network showing the top flow-carrying edges (99th percentile) calculated using the maximum-flow algorithm between RAP1 gene expression module and NFKB1.

References

Ahlmann-Eltze C, Huber W. 2020. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**: 5701–5702.

Akhmedov M, Kwee I, Montemanni R. 2016. A divide and conquer matheuristic algorithm for the Prize-collecting Steiner Tree Problem. *Comput Oper Res* **70**: 18–25.

Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A. 2016. Network-based inference of protein activity helps functionalize the genetic landscape of cancer. *Nat Genet* **48**: 838–847.

Baskaran JP, Weldy A, Guarin J, Munoz G, Shpilker PH, Kotlik M, Subbiah N, Wishart A, Peng Y, Miller MA, et al. 2020. Cell shape, and not 2D migration, predicts extracellular matrix-driven 3D cell invasion in breast cancer. *APL Bioeng* **4**: 026105.

Ben-Hamo R, Gidoni M, Efroni S. 2014. PhenoNet: identification of key networks associated with disease phenotype. *Bioinformatics* **30**: 2399–2405.

Bergert M, Lembo S, Sharma S, Russo L, Milovanović D, Gretarsson KH, Börmel M, Neveu PA, Hackett JA, Petsalaki E, et al. 2020. Cell Surface Mechanics Gate Embryonic Stem Cell Differentiation. *Cell Stem Cell*. http://www.sciencedirect.com/science/article/pii/S1934590920305336 (Accessed December 9, 2020).

Berman AY, Manna S, Schwartz NS, Katz YE, Sun Y, Behrmann CA, Yu JJ, Plas DR, Alayev A, Holz MK. 2017. ERRα regulates the growth of triple-negative breast cancer cells via S6K1-dependent mechanism. *Signal Transduct Target Ther* **2**: 1–9.

Bhatt D, Ghosh S. 2014. Regulation of the NF-κB-Mediated Transcription of Inflammatory Genes. *Front Immunol* **5**: 71.

Boettner B, Van Aelst L. 2009. Control of cell adhesion dynamics by Rap1 signaling. *Curr Opin Cell Biol* **21**: 684–693.

Bougault C, Aubert-Foucher E, Paumier A, Perrier-Groult E, Huot L, Hot D, Duterque-Coquillaud M, Mallein-Gerin F. 2012. Dynamic Compression of Chondrocyte-Agarose Constructs Reveals New Candidate Mechanosensitive Genes. *PLOS ONE* **7**: e36964.

Brin S, Page L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput Netw ISDN Syst* **30**: 107–117.

Burr SD, Stewart JA. 2021. Rap1a Overlaps the AGE/RAGE Signaling Cascade to Alter Expression of α-SMA, p-NF-κB, and p-PKC-ζ in Cardiac Fibroblasts Isolated from Type 2 Diabetic Mice. *Cells* **10**: 557.

Chatterjee S. 2018. Endothelial Mechanotransduction, Redox Signaling and the Regulation of Vascular Inflammatory Pathways. *Front Physiol* **9**. https://www.frontiersin.org/articles/10.3389/fphys.2018.00524/full (Accessed February 3,

2021).

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**: 128.

Chen W-C, Lin H-H, Tang M-J. 2015. Matrix-Stiffness–Regulated Inverse Expression of Krüppel-Like Factor 5 and Krüppel-Like Factor 4 in the Pathogenesis of Renal Fibrosis. *Am J Pathol* **185**: 2468–2481.

Cooper J, Giancotti FG. 2019. Integrin Signaling in Cancer: Mechanotransduction, Stemness, Epithelial Plasticity, and Therapeutic Resistance. *Cancer Cell* **35**: 347–367.

Cowell CF, Yan IK, Eiseler T, Leightner AC, Döppler H, Storz P. 2009. Loss of cell-cell contacts induces NF-kappaB via RhoA-mediated activation of protein kinase D1. *J Cell Biochem* **106**: 714–728.

Cowling VH, Cole MD. 2007. E-cadherin repression contributes to c-Myc-induced epithelial cell transformation. *Oncogene* **26**: 3582–3586.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal* **Complex Systems**: 1695.

Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, Shi B. 2015. Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res* **5**: 2929–2943.

DaSilva-Arnold SC, Kuo C-Y, Davra V, Remache Y, Kim PCW, Fisher JP, Zamudio S, Al-Khan A, Birge RB, Illsley NP. 2019. ZEB2, a master regulator of the epithelial-mesenchymal transition, mediates trophoblast differentiation. *Mol Hum Reprod* **25**: 61–75.

De Belly H, Stubb A, Yanagida A, Labouesse C, Jones PH, Paluch EK, Chalut KJ. 2020. Membrane Tension Gates ERK-Mediated Regulation of Pluripotent Cell Fate. *Cell Stem Cell*. http://www.sciencedirect.com/science/article/pii/S1934590920305348 (Accessed February 3, 2021).

Eslami Amirabadi H, Tuerlings M, Hollestelle A, SahebAli S, Luttge R, van Donkelaar CC, Martens JWM, den Toonder JMJ. 2019. Characterizing the invasion of different breast cancer cell lines with distinct E-cadherin status in 3D using a microfluidic system. *Biomed Microdevices* **21**: 101.

Euskirchen GM, Rozowsky JS, Wei C-L, Lee WH, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, et al. 2007. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* **17**: 898–909.

Fedele M, Cerchia L, Chiappetta G. 2017. The Epithelial-to-Mesenchymal Transition in Breast Cancer: Focus on Basal-Like Carcinomas. *Cancers* **9**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5664073/ (Accessed November 30, 2020).

Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, Ji X, Liu W, Huang B, Luo W, et al. 2018. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes Dis* **5**: 77–106.

Filer A, Buckley CD. 2013. 15 - Fibroblasts and Fibroblast-like Synoviocytes. In *Kelley's Textbook of Rheumatology (Ninth Edition)* (eds. G.S. Firestein, R.C. Budd, S.E. Gabriel, I.B. McInnes, and J.R. O'Dell), pp. 215–231, W.B. Saunders, Philadelphia

http://www.sciencedirect.com/science/article/pii/B9781437717389000153 (Accessed November 26, 2020).

Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. 2019. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**: 1363–1375.

Guan Y, Gorenshteyn D, Burmeister M, Wong AK, Schimenti JC, Handel MA, Bult CJ, Hibbs MA, Troyanskaya OG. 2012. Tissue-Specific Functional Networks for Prioritizing Phenotype and Disease Genes. *PLOS Comput Biol* **8**: e1002694.

Hamilton NA, Pantelic RS, Hanson K, Teasdale RD. 2007. Fast automated cell phenotype image classification. *BMC Bioinformatics* **8**: 110.

Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, et al. 2018. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* **46**: D380–D386.

Haralick RM, Shanmugam K, Dinstein I. 1973. Textural Features for Image Classification. *IEEE Trans Syst Man Cybern* **SMC-3**: 610–621.

Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, Joughin BA, Stegle O, Lauffenburger DA, Heyn H, et al. 2020. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol* **21**: 36.

Horton ER, Astudillo P, Humphries MJ, Humphries JD. 2016. Mechanosensitivity of integrin adhesion complexes: role of the consensus adhesome. *Exp Cell Res* **343**: 7–13.

Huang C-R, Lee C-T, Chang K-Y, Chang W-C, Liu Y-W, Lee J-C, Chen B-K. 2015. Down-regulation of ARNT promotes cancer metastasis by activating the fibronectin/integrin β1/FAK axis. *Oncotarget* **6**: 11530–11546.

Imamura T, Hikita A, Inoue Y. 2012. The roles of TGF-β signaling in carcinogenesis and breast cancer metastasis. *Breast Cancer Tokyo Jpn* **19**: 118–124.

Ishihara Y, Kado SY, Hoeper C, Harel S, Vogel CFA. 2019. Role of NF-kB RelB in Aryl Hydrocarbon Receptor-Mediated Ligand Specific Effects. *Int J Mol Sci* **20**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6600526/ (Accessed November 27, 2020).

Iván G, Grolmusz V. 2011. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics* **27**: 405–407.

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, et al. 2020. The reactome pathway knowledgebase. *Nucleic Acids Res* **48**: D498–D503.

Kadzik RS, Cohen ED, Morley MP, Stewart KM, Lu MM, Morrisey EE. 2014. Wnt ligand/Frizzled 2 receptor signaling regulates tube shape and branch-point formation in the lung through control of epithelial cell shape. *Proc Natl Acad Sci* **111**: 12444–12449.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**: D457–D462.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database J Biol Databases Curation* **2011**: bar030.

Kitano H. 2002. Systems Biology: A Brief Overview. *Science* **295**: 1662–1664.

Ko SY, Naora H. 2014. HOXA9 promotes homotypic and heterotypic cell interactions that facilitate ovarian cancer dissemination via its induction of P-cadherin. *Mol Cancer* **13**: 170.

Kontomanolis EN, Kalagasidou S, Pouliliou S, Anthoulaki X, Georgiou N, Papamanolis V, Fasoulakis ZN. 2018. The Notch Pathway in Breast Cancer Progression. *Sci World J* **2018**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6077551/ (Accessed January 7, 2021).

Korotkevich G, Sukhov V, Sergushichev A. 2019. Fast gene set enrichment analysis. *bioRxiv* 060012.

Krakhmal NV, Zavyalova MV, Denisov EV, Vtorushin SV, Perelmuter VM. 2015. Cancer Invasion: Patterns and Mechanisms. *Acta Naturae* **7**: 17–28.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**: W90-97.

Kumar S, Jang I, Kim CW, Kang D-W, Lee WJ, Jo H. 2016. Functional screening of mammalian mechanosensitive genes using Drosophila RNAi library– Smarcd3/Bap60 is a mechanosensitive pro-inflammatory gene. *Sci Rep* **6**: 36461.

Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. 2010. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinforma Oxf Engl* **26**: 2438–2444.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Lee J, Choi J-H, Joo C-K. 2013. TGF- β 1 regulates cell fate during epithelial–mesenchymal transition by upregulating survivin. *Cell Death Dis* **4**: e714–e714.

Lee J-W, Ryu Y-K, Ji Y-H, Kang JH, Moon E-Y. 2015a. Hypoxia/reoxygenation-experienced cancer cell migration and metastasis are regulated by Rap1- and Rac1-GTPase activation via the expression of thymosin beta-4. *Oncotarget* **6**: 9820–9833.

Lee M-H, Wu P-H, Gilkes D, Aifuwa I, Wirtz D. 2015b. Normal mammary epithelial cells promote carcinoma basement membrane invasion by inducing microtubule-rich protrusions. *Oncotarget* **6**: 32634–32645.

Li H, Liang J, Wang J, Han J, Li S, Huang K, Liu C. 2021. Mex3a promotes oncogenesis through the RAP1/MAPK signaling pathway in colorectal cancer and is inhibited by hsa-miR-6887-3p. *Cancer Commun* **41**: 472–491.

Li Z-R, Jiang Y, Hu J-Z, Chen Y, Liu Q-Z. 2019. SOX2 knockdown inhibits the migration and invasion of basal cell carcinoma cells by targeting the SRPK1-mediated PI3K/AKT signaling pathway. *Oncol Lett* **17**: 1617–1625.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**: 417–425.

Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. 2019. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *Npj Syst Biol Appl* **5**: 1–10.

Liu P, Tang H, Song C, Wang J, Chen B, Huang X, Pei X, Liu L. 2018. SOX2 Promotes Cell Proliferation and Metastasis in Triple Negative Breast Cancer. *Front Pharmacol* **9**: 942.

Lourenco C, Kalkat M, Houlahan KE, De Melo J, Longo J, Done SJ, Boutros PC, Penn LZ. 2019. Modelling the MYC-driven normal-to-tumour switch in breast cancer. *Dis Model Mech* **12**: dmm038083.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Luna A, Babur Ö, Aksoy BA, Demir E, Sander C. 2016. PaxtoolsR: pathway analysis in R using Pathway Commons. *Bioinforma Oxf Engl* **32**: 1262–1264.

Makki J. 2015. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clin Med Insights Pathol* **8**: 23–31.

Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, A. Miller R, Digles D, Lopes EN, Ehrhart F, et al. 2021. WikiPathways: connecting communities. *Nucleic Acids Res* **49**: D613–D621.

Menon J, Doebele RC, Gomes S, Bevilacqua E, Reindl KM, Rosner MR. 2012. A Novel Interplay between Rap1 and PKA Regulates Induction of Angiogenesis in Prostate Cancer. *PLOS ONE* **7**: e49893.

Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **41**: D377-386.

Miralles F, Posern G, Zaromytidou A-I, Treisman R. 2003. Actin dynamics control SRF activity by regulation of its coactivator MAL. *Cell* **113**: 329–342.

Moret N, Clark NA, Hafner M, Wang Y, Lounkine E, Medvedovic M, Wang J, Gray N, Jenkins J, Sorger PK. 2019. Cheminformatics Tools for Analyzing and Designing Optimized Small-Molecule Collections and Libraries. *Cell Chem Biol* **26**: 765-777.e3.

Mun H, Jeon TJ. 2012. Regulation of actin cytoskeleton by Rap1 binding to RacGEF1. *Mol Cells* **34**: 71–76.

Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe J-P, Tong F, et al. 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**: 515–527.

Ohba Y, Kurokawa K, Matsuda M. 2003. Mechanism of the spatio-temporal regulation of Ras and Rap1. *EMBO J* **22**: 859–869.

Olson EN, Nordheim A. 2010. Linking actin dynamics and gene transcription to drive cellular motile functions. *Nat Rev Mol Cell Biol* **11**: 353–365.

Orsulic S, Huber O, Aberle H, Arnold S, Kemler R. 1999. E-cadherin binding prevents beta-catenin nuclear localization and beta-catenin/LEF-1-mediated transactivation. *J Cell Sci* **112**: 1237–1245.

Øvrevik J, Låg M, Lecureur V, Gilot D, Lagadic-Gossmann D, Refsnes M, Schwarze PE, Skuland T, Becher R, Holme JA. 2014. AhR and Arnt differentially regulate NF-κB signaling and chemokine responses in human bronchial epithelial cells. *Cell Commun Signal CCS* **12**: 48.

Padi M, Quackenbush J. 2018. Detecting phenotype-driven transitions in regulatory

network structure. *Npj Syst Biol Appl* **4**: 1–12.

Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, Fonseca NA, Füllgrabe A, Green M, Huang N, et al. 2020. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res* **48**: D77–D83.

Piran M, Karbalaei R, Piran M, Aldahdooh J, Mirzaie M, Ansari-Pour N, Tang J, Jafari M. 2020. Can We Assume the Gene Expression Profile as a Proxy for Signaling Network Activity? *Biomolecules* **10**: 850.

Pires BRB, Mencalha AL, Ferreira GM, de Souza WF, Morgado-Díaz JA, Maia AM, Corrêa S, Abdelhay ESFW. 2017. NF-kappaB Is Involved in the Regulation of EMT Genes in Breast Cancer Cells. *PLoS ONE* **12**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5249109/ (Accessed January 21, 2021).

Plaks V, Kong N, Werb Z. 2015. The Cancer Stem Cell Niche: How Essential is the Niche in Regulating Stemness of Tumor Cells? *Cell Stem Cell* **16**: 225–238.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Reddy TP, Rosato RR, Li X, Moulder S, Piwnica-Worms H, Chang JC. 2020. A comprehensive overview of metaplastic breast cancer: clinical features and molecular aberrations. *Breast Cancer Res* **22**: 121.

Resnik P. 1995. 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*, pp. 448–453, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Robertson J, Jacquemet G, Byron A, Jones MC, Warwood S, Selley JN, Knight D, Humphries JD, Humphries MJ. 2015. Defining the phospho-adhesome through the phosphoproteomic analysis of integrin signalling. *Nat Commun* **6**: 6265.

Roche J. 2018. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers* **10**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5836084/ (Accessed November 30, 2020).

Rolfe RA, Nowlan NC, Kenny EM, Cormican P, Morris DW, Prendergast PJ, Kelly D, Murphy P. 2014. Identification of mechanosensitive genes during skeletal development: alteration of genes associated with cytoskeletal rearrangement and cell signalling pathways. *BMC Genomics* **15**: 48.

Roy Choudhury A, Gupta S, Chaturvedi PK, Kumar N, Pandey D. 2019. Mechanobiology of Cancer Stem Cells and Their Niche. *Cancer Microenviron* **12**: 17–27.

Sailem HZ, Bakal C. 2017. Identification of clinically predictive metagenes that encode components of a network coupling cell shape to transcription by image-omics. *Genome Res* **27**: 196–207.

Sawant K, Chen Y, Kotian N, Preuss KM, McDonald JA. 2018. Rap1 GTPase promotes coordinated collective cell migration in vivo. *Mol Biol Cell* **29**: 2656–2673.

Schwartz TL, Mogal H, Papageorgiou C, Veerapong J, Hsueh EC. 2013. Metaplastic breast cancer: histologic characteristics, prognostic factors and systemic treatment strategies. *Exp Hematol Oncol* **2**: 31.

Sero JE, Sailem HZ, Ardy RC, Almuttaqi H, Zhang T, Bakal C. 2015. Cell shape and the microenvironment regulate nuclear translocation of NF-κB in breast epithelial and tumor

cells. *Mol Syst Biol* **11**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380925/ (Accessed November 5, 2020).

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**: 2498–2504.

Sharma S, Petsalaki E. 2019. Large-scale datasets uncovering cell signalling networks in cancer: context matters. *Curr Opin Genet Dev* **54**: 118–124.

Shih W, Yamada S. 2012. N-cadherin-mediated cell–cell adhesion promotes cell migration in a three-dimensional matrix. *J Cell Sci* **125**: 3661–3670.

Shrum CK, Defrancisco D, Meffert MK. 2009. Stimulated nuclear translocation of NF-kappaB and shuttling differentially depend on dynein and the dynactin complex. *Proc Natl Acad Sci U S A* **106**: 2647–2652.

Siegel RL, Miller KD, Jemal A. 2019. Cancer statistics, 2019. *CA Cancer J Clin* **69**: 7–34.

Song Y, Washington MK, Crawford HC. 2010. Loss of FOXA1/2 Is Essential for the Epithelial-to-Mesenchymal Transition in Pancreatic Cancer. *Cancer Res* **70**: 2115–2125.

Soul J, Hardingham TE, Boot-Handford RP, Schwartz J-M. 2015. PhenomeExpress: A refined network analysis of expression datasets by inclusion of known disease phenotypes. *Sci Rep* **5**: 8117.

Stathias V, Turner J, Koleti A, Vidovic D, Cooper D, Fazel-Najafabadi M, Pilarczyk M, Terryn R, Chung C, Umeano A, et al. 2020. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res* **48**: D431–D439.

Stormo GD. 2013. Modeling the specificity of protein-DNA interactions. *Quant Biol Beijing China* **1**: 115–130.

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. 2017. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**: 1437-1452.e17.

Szalai B, Saez-Rodriguez J. 2020. Why do pathway methods work better than they should? *FEBS Lett* **594**: 4189–4200.

Taherian A, Li X, Liu Y, Haas TA. 2011. Differences in integrin expression and signaling within human breast cancer cells. *BMC Cancer* **11**: 293.

Teo H, Ghosh S, Luesch H, Ghosh A, Wong ET, Malik N, Orth A, de Jesus P, Perry AS, Oliver JD, et al. 2010. Telomere-independent Rap1 is an IKK adaptor and regulates NF-kappaB-dependent gene expression. *Nat Cell Biol* **12**: 758–767.

Tong H, Faloutsos C, Pan J-Y. 2008. Random walk with restart: fast solutions and applications. *Knowl Inf Syst* **14**: 327–346.

Tong L, Tergaonkar V. 2014. Rho protein GTPases and their interactions with NFκB: crossroads of inflammation and matrix biology. *Biosci Rep* **34**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4069681/ (Accessed November 29, 2020).

Türei D, Korcsmáros T, Saez-Rodriguez J. 2016. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* **13**: 966–967.

Van Tubergen EA, Banerjee R, Liu M, Vander Broek R, Light E, Kuo S, Feinberg SE, Willis AL, Wolf G, Carey T, et al. 2013. Inactivation or loss of TTP promotes invasion in

head and neck cancer via transcript stabilization and secretion of MMP9, MMP2, and IL-6. *Clin Cancer Res Off J Am Assoc Cancer Res* **19**: 1169–1179.

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**: 227–232.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**: 276–287.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York http://ggplot2.org.

Wildwater M, Sander N, Vreede G de, Heuvel S van den. 2011. Cell shape and Wnt signaling redundantly control the division axis of C. elegans epithelial stem cells. *Development* **138**: 4375–4385.

Wozniak MA, Chen CS. 2009. Mechanotransduction in development: a growing role for contractility. *Nat Rev Mol Cell Biol* **10**: 34–43.

Wu P-H, Gilkes DM, Phillip JM, Narkar A, Cheng TW-T, Marchand J, Lee M-H, Li R, Wirtz D. 2020. Single-cell morphology encodes metastatic potential. *Sci Adv* **6**: eaaw6938.

Wu X, Chen H, Parker B, Rubin E, Zhu T, Lee JS, Argani P, Sukumar S. 2006. HOXB7, a Homeodomain Protein, Is Overexpressed in Breast Cancer and Confers Epithelial-Mesenchymal Transition. *Cancer Res* **66**: 9527–9534.

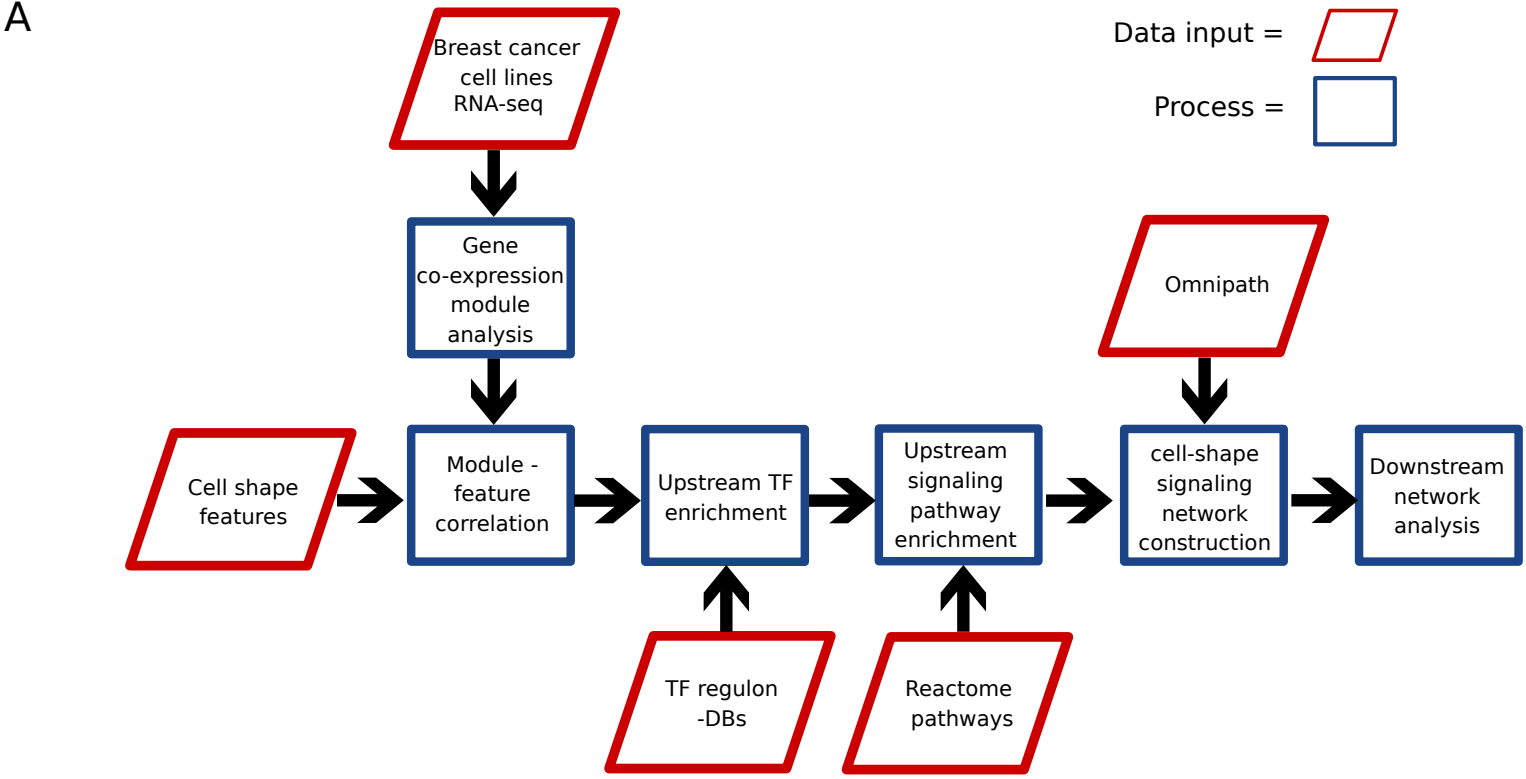
Yin P, Wang W, Zhang Z, Bai Y, Gao J, Zhao C. 2018. Wnt signaling in human and mouse breast cancer: Focusing on Wnt ligands, receptors and antagonists. *Cancer Sci* **109**: 3368–3375.

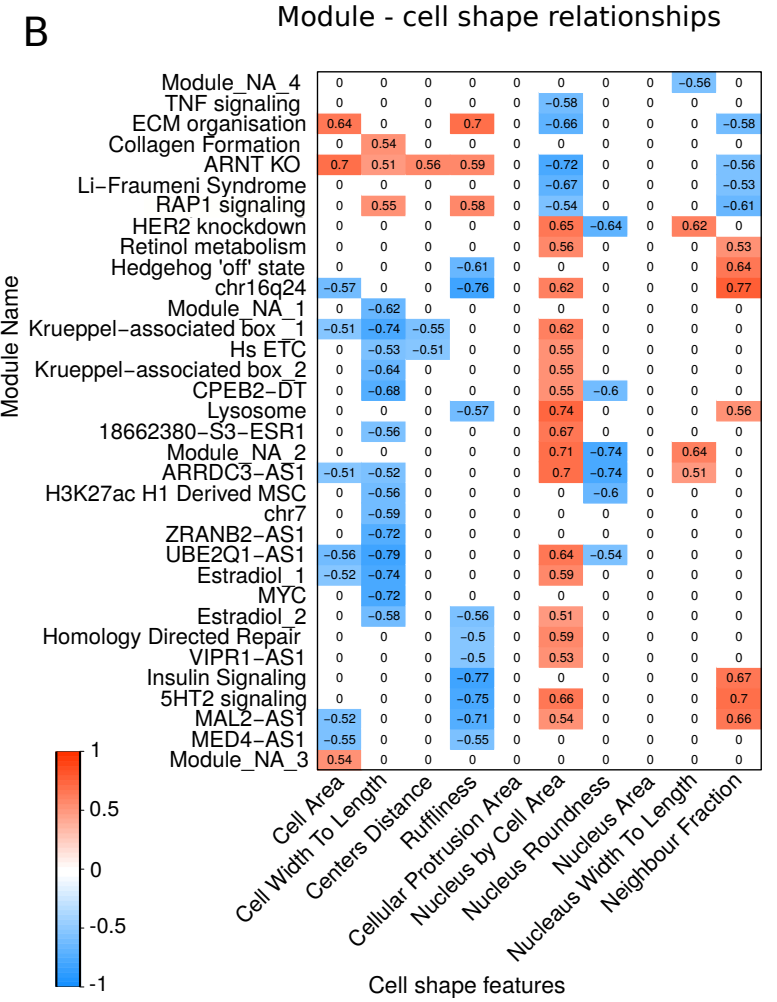
Zhang Y, Chiu S, Liang X, Gao F, Zhang Z, Liao S, Liang Y, Chai Y-H, Low DJH, Tse H-F, et al. 2015. Rap1-mediated nuclear factor-kappaB (NF- κ B) activity regulates the paracrine capacity of mesenchymal stem cells in heart repair following infarction. *Cell Death Discov* **1**: 1–11.

Zhang Y-L, Wang R-C, Cheng K, Ring BZ, Su L. 2017. Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer Biol Med* **14**: 90–99.

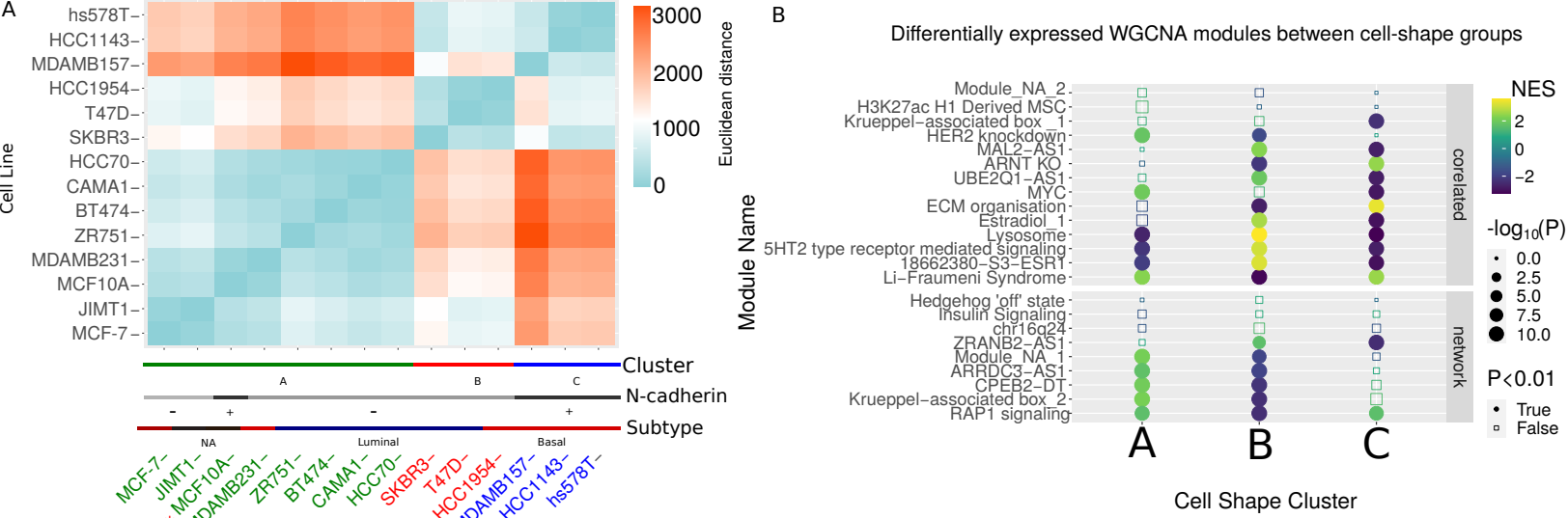
Zheng B, Han M, Bernier M, Wen J. 2009. Nuclear actin and actin-binding proteins in the regulation of transcription and gene expression. *FEBS J* **276**: 2669–2685.

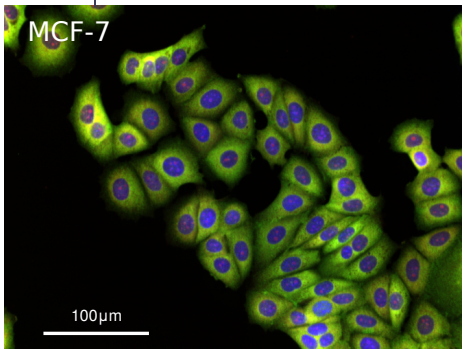
Zou A, Lehn S, Magee N, Zhang Y. 2015. New Insights into Orphan Nuclear Receptor SHP in Liver Cancer. *Nucl Recept Res* **2**: 101162.

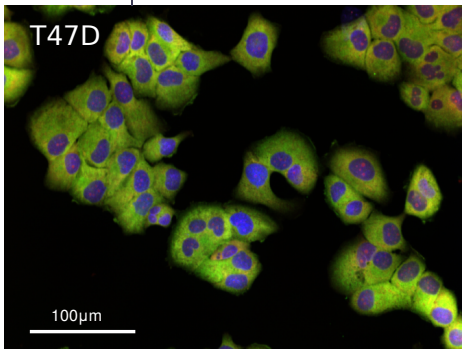

A
Data input =
Process =
Breast cancer cell lines RNA-seq
Gene co-expression module analysis
Omnipath
Cell shape features
Module - feature correlation
Upstream TF enrichment
Upstream signaling pathway enrichment
cell-shape signaling network construction
Downstream network analysis
TF regulon -DBs
Reactome pathways

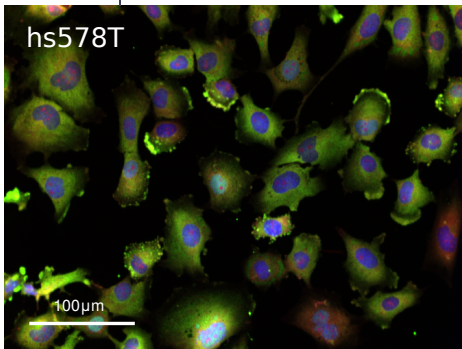

B
Module - cell shape relationships
Module Name
Module_NA_4
TNF signaling
ECM organisation
Collagen Formation
ARNT KO
Li−Fraumeni Syndrome
RAP1 signaling
HER2 knockdown
Retinol metabolism
Hedgehog 'off' state
chr16q24
Module_NA_1
Krueppel−associated box_1
Hs ETC
Krueppel−associated box_2
CPEB2−DT
Lysosome
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
1
0.5
0
-0.5
-1
Cell Area
Cell Width To Length
Centers Distance
Ruffliness
Cellular Protrusion Area
Nucleus by Cell Area
Nucleus Roundness
Nucleus Area
Nucleus Width To Length
Neighbour Fraction
Cell shape features

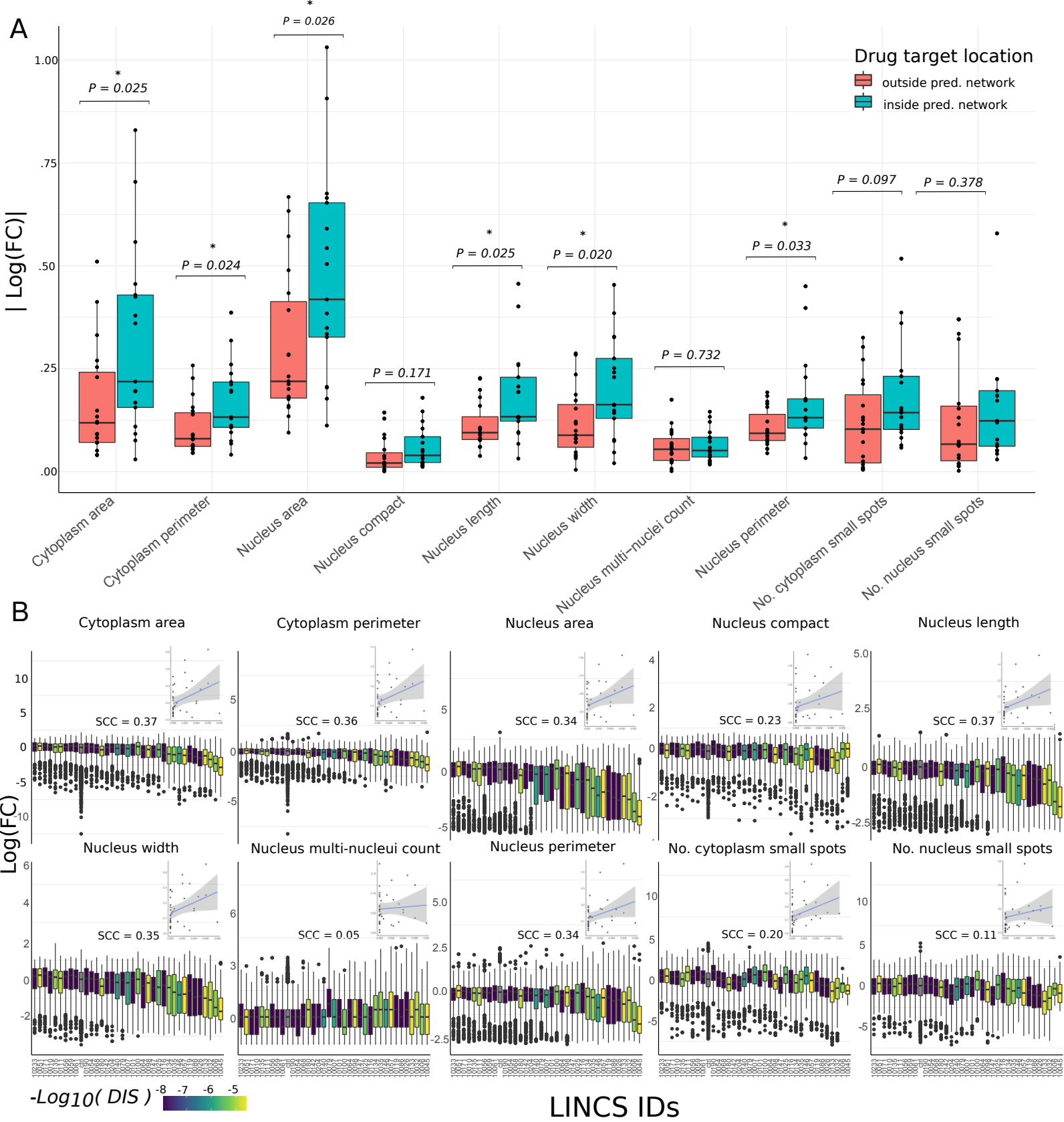

C
Enriched Pathways
Signal Transduction
Nuclear signaling by ERBB4
TGF-β receptor signaling activates SMADs
Signaling by BMP
Signaling by Activin
Downregulation of SMAD2/3:4 activity
Signaling by NOTCH2
TCF dependent signaling in response to WNT
NOTCH2 regulates transcription
Transcriptional activity of SMAD2/3:4
SMAD2/3:4 heterotrimer
CREB phosphorylation
p75NTR negatively regulates cell cycle via SC1
Formation of the β-catenin complex
Signaling by TGF-β Receptor Complex
Binding of TCF/LEF:CTNNB1 to gene promoters
Nuclear Events(kinase and TF activation)
Pre-NOTCH Expression and Processing
Pre-NOTCH Transcription and Translation
Signaling by NOTCH1
Signaling by NOTCH
NOTCH1 RegulatesTranscription
-log10(P)
4
6
8
log10(Odds Ratio)
10.0
7.5
5.0
2.5
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
5HT2 type receptor mediated signaling
Module Name

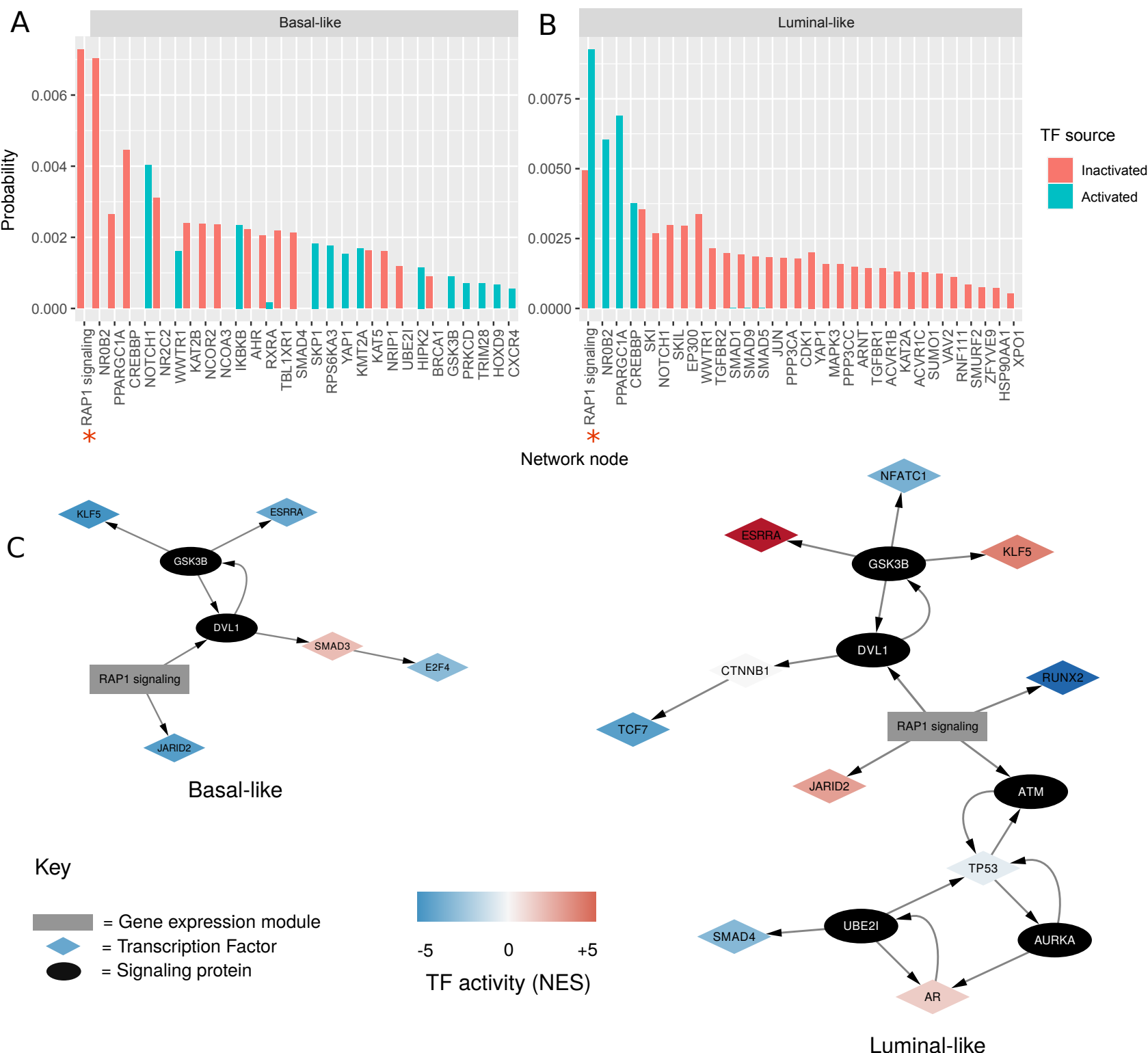

A
hs578T
HCC1143
MDAMB157
HCC1954
T47D
SKBR3
HCC70
CAMA1
BT474
ZR751
MDAMB231
MCF10A
JIMT1
MCF-7
Cell Line
3000
2000
1000
0
Euclidean distance
Cluster
A
B
C
N-cadherin
-
+
-
+
Subtype
NA
Luminal
Basal
B
Differentially expressed WGCNA modules between cell-shape groups
Module_NA_2
H3K27ac H1 Derived MSC
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
Module Name
A
B
C
Cell Shape Cluster
NES
2
0
-2
-log10(P)
0.0
2.5
5.0
7.5
10.0
P<0.01
True
False
MCF-7
JIMT1
*MCF10A
MDAMB231
ZR751
BT474
CAMA1
HCC70
SKBR3
T47D
HCC1954
MDAMB157
HCC1143
hs578T

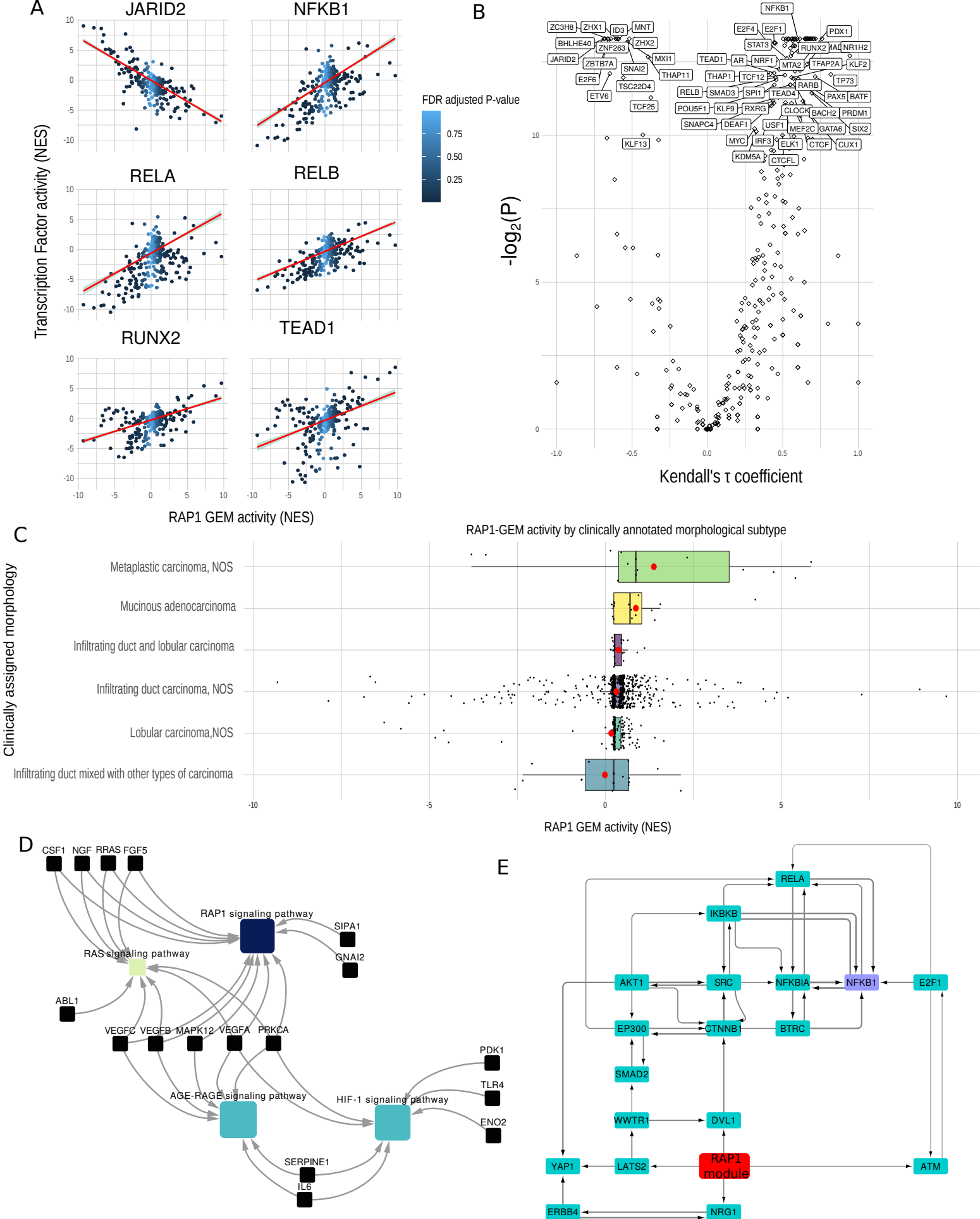

MCF-7
100µm


T47D
100µm


hs578T
100µm


A
| Log(FC)|
Drug target location
outside pred. network
inside pred. network
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)
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
C
KLF5
ESRRA
GSK3B
DVL1
SMAD3
E2F4
RAP1 signaling
JARID2
Basal-like
NFATC1
ESRRA
KLF5
GSK3B
DVL1
CTNNB1
TCF7
RUNX2
RAP1 signaling
JARID2
ATM
TP53
SMAD4
UBE2I
AURKA
AR
Luminal-like
Key
= Gene expression module
= Transcription Factor
= Signaling protein
-5
0
+5
TF activity (NES)


A
JARID2
NFKB1
RELA
RELB
RUNX2
TEAD1
Transcription Factor activity (NES)
RAP1 GEM activity (NES)
FDR adjusted P-value
0.75
0.50
0.25
B
-log2(P)
Kendall's τ coefficient
C
RAP1-GEM activity by clinically annotated morphological subtype
Clinically assigned morphology
Metaplastic carcinoma, NOS
Mucinous adenocarcinoma
Infiltrating duct and lobular carcinoma
Infiltrating duct carcinoma, NOS
Lobular carcinoma,NOS
Infiltrating duct mixed with other types of carcinoma
RAP1 GEM activity (NES)
D
CSF1
NGF
RRAS
FGF5
RAP1 signaling pathway
SIPA1
GNAI2
RAS signaling pathway
ABL1
VEGFC
VEGFB
MAPK12
VEGFA
PRKCA
PDK1
TLR4
ENO2
AGE-RAGE signaling pathway
HIF-1 signaling pathway
SERPINE1
IL6
E
RELA
IKBKB
AKT1
SRC
NFKBIA
NFKB1
E2F1
EP300
CTNNB1
BTRC
SMAD2
WWTR1
DVL1
YAP1
LATS2
RAP1 module
ATM
ERBB4
NRG1

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

Charlie George Barker, Eirini Petsalaki, Girolamo Giudice, et al.

Genome Res. published online February 23, 2022
Access the most recent version at doi:10.1101/gr.276059.121