

# SCIENTIFIC REPORTS



OPEN

## Phenotype-driven identification of epithelial signalling clusters

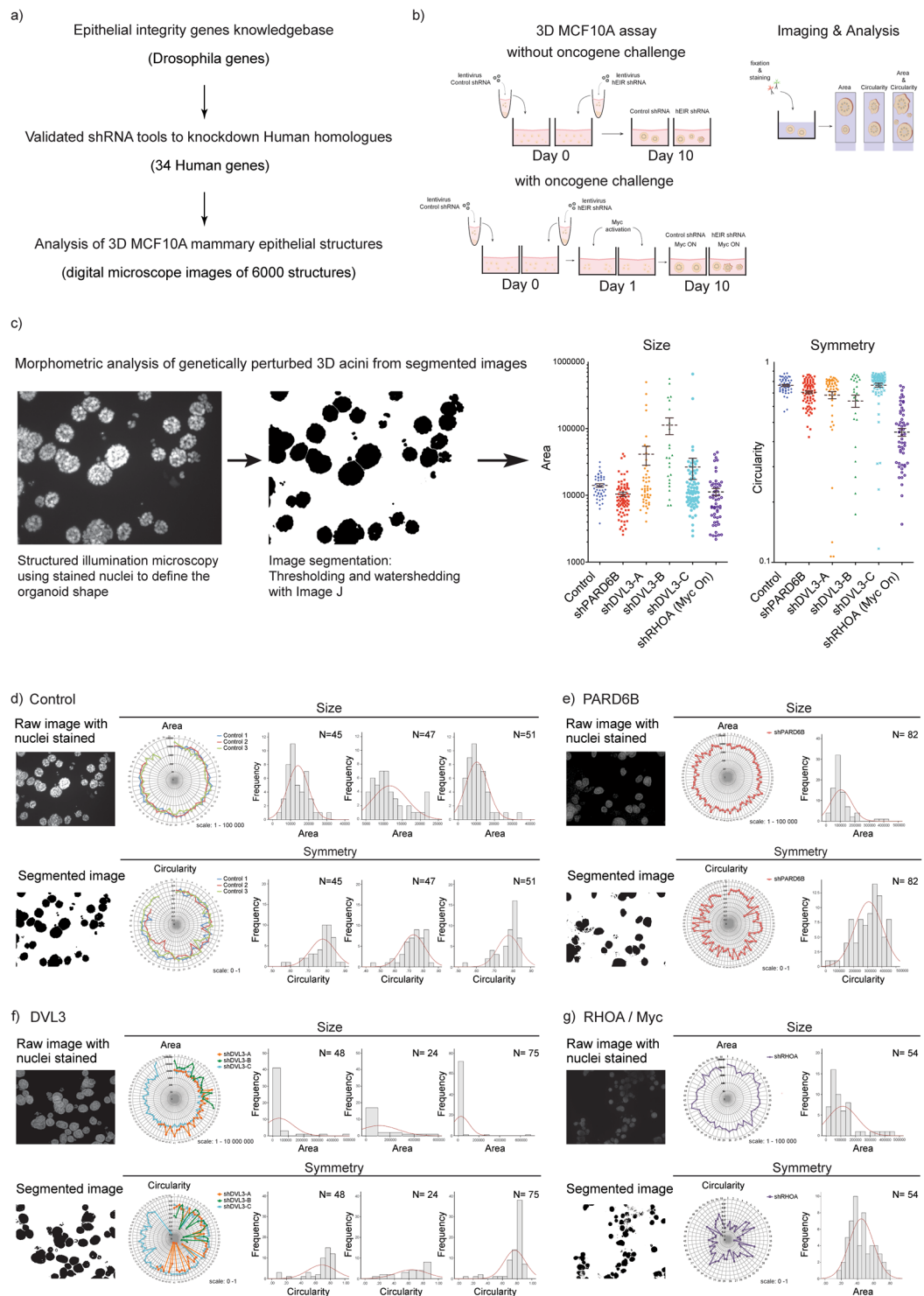
Elsa Marques<sup>1</sup>, Tomi Peltola<sup>2</sup>, Samuel Kaski<sup>2</sup> & Juha Klefström<sup>1</sup>

In metazoans, epithelial architecture provides a context that dynamically modulates most if not all epithelial cell responses to intrinsic and extrinsic signals, including growth or survival signalling and transforming oncogene action. Three-dimensional (3D) epithelial culture systems provide tractable models to interrogate the function of human genetic determinants in establishment of context-dependency. We performed an arrayed genetic shRNA screen in mammary epithelial 3D cultures to identify new determinants of epithelial architecture, finding that the key phenotype impacting shRNAs altered not only the data population average but even more noticeably the population distribution. The broad distributions were attributable to sporadic gene silencing actions by shRNA in unselected populations. We employed Maximum Mean Discrepancy concept to capture similar population distribution patterns and demonstrate here the feasibility of the test in identifying an impact of shRNA in populations of 3D structures. Integration of the clustered morphometric data with protein-protein interactions data enabled hypothesis generation of novel biological pathways underlying similar 3D phenotype alterations. The results present a new strategy for 3D phenotype-driven pathway analysis, which is expected to accelerate discovery of context-dependent gene functions in epithelial biology and tumorigenesis.

Epithelial cells form sheets that line the body cavities, surfaces of different organs and they also form specialized glandular structures. Characteristic for epithelial cells is that they jointly respond to external and internal stimuli to deliver epithelial-specific functions such as directional secretion. Such tissue-wide actions require well-developed communication systems across the layers of epithelial cells, to ensure that the whole tissue performs or responds in a unified manner<sup>1</sup>. The joint specialized signalling functions of the apicobasally polarized epithelial cells are partly coordinated via the same specialized cell-cell junctions that establish the structural cohesion; tight junctions, adhesion junctions, desmosomes and the cell-basement membrane (BM) connecting hemidesmosomes<sup>2,3</sup>. In cancer, loss of organized epithelial structure is the most important diagnostic criterion and a hallmark of progression<sup>4,5</sup>. Interestingly, emerging evidence suggests that loss of epithelial integrity is not just an epiphenomenon in cancer, but that loss of proper epithelial cell interactions with each other and the extracellular matrix actually contributes to perturbed cell growth, metabolism and proliferation programs as well as the invasiveness of cancerous cells<sup>6-9</sup>. While it is now clear that the structure and organization of epithelial tissue (*context*) fundamentally regulates individual cell functions within the epithelium and that loss of these control mechanisms contributes to neoplasia, little is still known how that secondary level signalling controls the biology of epithelial cells<sup>10</sup>. Three-dimensional (3D) culture models of BM-attached epithelial tissues, pioneered by Mina Bissell and collaborators, have greatly facilitated studies aiming to clarify the context-dependencies of epithelial cell functions in health and disease<sup>11,12</sup>. The 3D cultures are amenable for both genetic and pharmacological perturbation and the perturbed epithelial phenotypes and functions can be comprehensively visualized with confocal fluorescence microscopy or other imaging tools<sup>13-16</sup>.

However, while genetic perturbation of such 3D structures that develop from single cells is relatively straightforward by shRNA or gene editing tools, the statistical population level analysis of perturbed 3D phenotypes can be challenging. The low throughput of phenotype analysis can present a major bottleneck in genetic screens, which produce large image datasets. Currently, there are no standard procedures for systematic analysis of genetically perturbed populations of epithelial 3D structures.

<sup>1</sup>Cancer Cell Circuitry Laboratory, Research Programs Unit/Translational Cancer Biology & Medicum, University of Helsinki, P.O. Box 63 (street address: Haartmaninkatu 8), 00014 University of Helsinki, Helsinki, Finland. <sup>2</sup>Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, PO BOX 15400, FI-00076, Aalto, Finland. Correspondence and requests for materials should be addressed to J.K. (email: [Juha.Klefstrom@helsinki.fi](mailto:Juha.Klefstrom@helsinki.fi))



**Figure 1.** Impact of hEIR gene silencing on the phenotype of 3D MCF10A acinar structures at population level. **(a)** Outline of the shRNA screen in 3D culture. Human homologs of known Drosophila epithelial integrity regulating genes were identified and an shRNA knockdown library was designed and validated. The screen produced a databank of over 6000 images of the structures. The shRNA-induced morphological alterations were analysed in this study. **(b)** Experimental set up for analysis of shRNA perturbed 3D structures with and without oncogene challenge. The validated shRNAs were introduced in MCF10A cells expressing the conditionally active form of MYC (MycER). The shRNA expressing MCF10A acinar structures were allowed to form in Matrigel with or without active MYC for 10 days, then fixed and immunostained for morphometric analyses. In these assays, MYC activation enhanced the growth of the 3D structures and prevented the maturation-associated cell cycle exit. **(c)** Morphometric quantification of shRNA-induced phenotypic changes. Structured illumination microscopy was used to outline the shapes of the acinar structures. The shapes were captured in digital images, the original grayscale images were converted to binary images and watershed algorithm was

occasionally used for segmentation. The area (proxy for size) and circularity (proxy for symmetry) values were obtained with ImageJ software. The graphs on the right show the average and distribution of the area and circularity values corresponding to example 3D acinar populations. Each dot represents a single 3D structure. (d–g) Distributions of the size and symmetry parameters of 3D acinar populations engineered to express indicated shRNAs. In each figure, the left panels show digital microscope-acquired images of the structures (as in c) and the corresponding segmented images. The radar graphs show distribution of area and circularity values obtained for the 3D acini of each population. The data distribution histograms are overlaid with the normality curve in red. (d) Data analysis of three independent control populations, (e) PARD6B silenced population, (f) data analysis of populations exposed to three different shRNA species targeting DVL3. (g) Data analysis of population exposed to active MYC together with RHOA silencing.

Here we present a new statistical framework for systematic analysis of the shRNA mediated phenotypic effects in arrayed populations of MCF10A 3D acinar structures. The statistical framework is based on univariate non-parametric Wilcoxon Rank-Sum statistical test and multivariate maximum mean discrepancy (MMD) analysis of the population distributions<sup>17,18</sup>. The MMD test metrics was employed to describe the population distances in form of heatmap and network visualization. The statistical significance defined phenotype-linked MMD networks were then integrated with a STRING database of protein-protein interactions. These integrated phenotype-proteome networks exposed protein-protein interactions known to be critical for the formation of epithelial morphology as well as new hypothetical protein networks. We also provide evidence suggesting that MYC oncoprotein-dependent changes in epithelial integrity can be interrogated through phenotype associated protein networks. In broader scope, the Wilcoxon-MMD statistical framework offers a robust data analysis platform suitable for analysis of any genetic or pharmacological screen with 3D phenotype change as a read-out.

## Results

**Impact of a loss-of gene function to 3D epithelial structures in three-dimensional culture.** The morphological data examined here was derived from a recent shRNA mediated gene knockdown screen performed in a well-established mammary epithelial MCF10A 3D culture model of the mammary acinar structure. The screen was originally devised to identify genes and pathways, which play a role in regulation of acinar morphogenesis and epithelial maturation associated cell cycle exit<sup>14</sup>. The arrayed shRNA screen targeted Drosophila-informed human candidate genes with validated shRNAs, and identified distinct sets of genes important for establishment of proper acinar architecture and appropriately timed cell cycle restriction<sup>14</sup>. The shRNA effects were examined in both non-transformed and MYC oncoprotein transformed acinar structures. Altogether 34 human genes (collectively called as hEIR genes for human epithelial integrity related) could be efficiently silenced with the shRNA reagents generated in the study (Fig. 1a). The functionally validated shRNAs were lentivirally introduced into MCF10A mammary epithelial cells in an arrayed format and the cells were then directly transferred to Matrigel™, which facilitates the formation of 3D structures. The MCF10A cells used in the study contained a conditionally active MYC oncogene, which enabled analysis of the hEIR gene function under an oncogene challenge (Fig. 1b). The phenotypes developed in these cultures, with or without active MYC, were captured using structured illumination microscopy followed by digital image processing (Fig. 1c). The whole experiment produced over 6000 digital images, exposing a variety of gene knockdown induced morphological alterations<sup>14</sup>.

To identify suitable statistical tests for interrogation of the morphometric data, we first determined whether the 3D acinar populations were normally distributed. Examples showing area (surrogate for size) and circularity (surrogate for symmetry) value distributions within indicated acinar populations are shown in Fig. 1d–g. The silencing of indicated genes resulted in prominent phenotype alterations already observable during visual inspection of the images. Loss of PARD6B resulted in abnormally small yet symmetric acinar structures (Fig. 1e), loss of DVL3 resulted in abnormally large and asymmetric structures (Fig. 1f) whereas loss of RHOA combined with MYC (RHOA/Myc) caused widespread apoptosis leading to appearance of highly irregular structures (Fig. 1g). Shapiro-Wilk normality test of area distributions indicated that only a minority of the populations (6/55 without MYC and 4/55 with MYC) were normally distributed. Furthermore, also only a minority of circularity distributions were Gaussian (12/55 without MYC and 17/55 with MYC).

As a result of these population structure analyses, the statistical power of the phenotype differences was evaluated with non-parametric Wilcoxon Rank-Sum statistical test. To keep the amount of experimental work manageable, the original shRNA screen was performed as seven independent experimental sets over a time period of several years. Each experimental set included an internal shRNA control and between 4 to 12 shRNA constructs. The two-sample Wilcoxon Rank-Sum test, which is a non-parametric counterpart of the commonly used t-test, was performed for each experimental set separately. The statistical significance of the size and symmetry difference between each gene-silenced 3D acinar population and the experimental set-specific control was examined (Fig. 2a; green: significant size difference and orange: significant symmetry difference). Benjamini-Hochberg procedure was used for multiple testing correction, defining the false-discovery rate (FDR) < 0.05<sup>19,20</sup>. The gene-silenced 3D acinar populations grown under MYC oncogene challenge were analysed similarly and the results are shown in Fig. 2b. The analysis shows that 9 shRNA constructs, corresponding to 9 targeted genes, altered the acinar size and 25 shRNA constructs (21 genes) caused a symmetry defect (Fig. 2c). Under oncogenic MYC challenge, 14 shRNA constructs (11 genes) altered the acinar size and 10 shRNA (9 genes) led to symmetry defect (Fig. 2d). There was little overlap between the genes identified as hits in the two separate screens; with or without MYC (Figs 1b, 2c,d). These results could be explained by the fact that the examined structures were in completely different proliferative state at the time of analysis; without MYC the 3D structures were quiescent

a) Acinar phenotype

Set 1					Set 2					Set 3														
Ctrl vs	TMO3	PAR6B	CDC42	CDC42	Ctrl vs	LATS1	STK3	NUMB-A	NUMB-B	SCRIB	NUMB-A	NUMB-B	Ctrl vs	PAR6B	PAR6B	CTNNB1	MAPK10	IKKγ	IKKγ	EZR-1B	DAG-1-2	GPM2-1	STAU2-2	
Size																								
Symmetry																								

Set 4							Set 5													
Ctrl vs	FZD1	MOI	MOI	DVL3-C	DVL3-A	DVL3-B	Ctrl vs	GSK3b	CRB3-3	FAT1-3B	FUZ-1B	GSK3b	YAP1-1B	RHOA-2	RHOA-3	CDH1-2	CDH1-3	CDH1-4	CRB3-2	
Size																				
Symmetry																				

Set 6						Set 7						
Ctrl vs	CTNNB1	LLGL2	MPP5-A	MPP5-B	MPP5-3A	Ctrl vs	TSC2	TSC2-3B	RDX-1B	MPP5-3B	PARD6A	MAPK11
Size												
Symmetry												

b) Acinar phenotype with oncogene challenge

Set 1					Set 2					Set 3														
Ctrl vs	TMO3	PAR6B	CDC42	CDC42	Ctrl vs	LATS1	STK3	NUMB-A	NUMB-B	SCRIB	NUMB-A	NUMB-B	Ctrl vs	PAR6B	PAR6B	CTNNB1	MAPK10	IKKγ	IKKγ	EZR-1B	DAG-1-2	GPM2-1	STAU2-2	
Size																								
Symmetry																								

Set 4							Set 5													
Ctrl vs	FZD1	MOI	MOI	DVL3-C	DVL3-A	DVL3-B	Ctrl vs	GSK3b	CRB3-3	FAT1-3B	FUZ-1B	GSK3b	YAP1-1B	RHOA-2	RHOA-3	CDH1-2	CDH1-3	CDH1-4	CRB3-2	
Size																				
Symmetry																				

Set 6						Set 7						
Ctrl vs	CTNNB1	LLGL2	MPP5-A	MPP5-B	MPP5-3A	Ctrl vs	TSC2	TSC2-3B	RDX-1B	MPP5-3B	PARD6A	MAPK11
Size												
Symmetry												

c)

Acinar phenotype		
<b>Reduced growth</b> NF2-2	<b>Loss of symmetry</b> CDC42-A STK3 NUMB-B CTNNB1-A	<b>Reduced growth with loss of symmetry</b> PARD6B
<b>Overgrowth</b> MOESIN * MOBKL1A-B CRB3-3	IKKγ-2B IKKγ-1B EZRIN-1B DAG1-2 GPM2-1 STAU-2 * MOBKL1A-A * DVL3-A * MPP5-B	<b>Reduced growth with improved symmetry</b> LLGL2
	<b>Improved symmetry</b> RHOA-2 TSC2 TSC2-3B RADIXIN-1B * MPP5-3B PARD6A MAPK11-3B	<b>Overgrowth with loss of symmetry</b> PARD6G * DVL3-B GSK3b
		<b>Overgrowth with improved symmetry</b> -

d)

Acinar phenotype with oncogene challenge		
<b>Reduced growth</b> YAP-1B CADH1-3 CTNNB1-B LLGL2 * MPP5-3A	<b>Loss of symmetry</b> NUMB-B	<b>Reduced growth with loss of symmetry</b> RHOA-2
<b>Overgrowth</b> EZRIN-1B CRB3-2 * MPP5-A	<b>Improved symmetry</b> MOESIN * MOBKL1A-1B TSC2	<b>Reduced growth with improved symmetry</b> CTNNB1-A * MOBKL1A-A NF2-2 FUZ-1B
		<b>Overgrowth with loss of symmetry</b> * MPP5-B
		<b>Overgrowth with improved symmetry</b> -

\* see supplemental table 1 - phenotypes requiring special attention

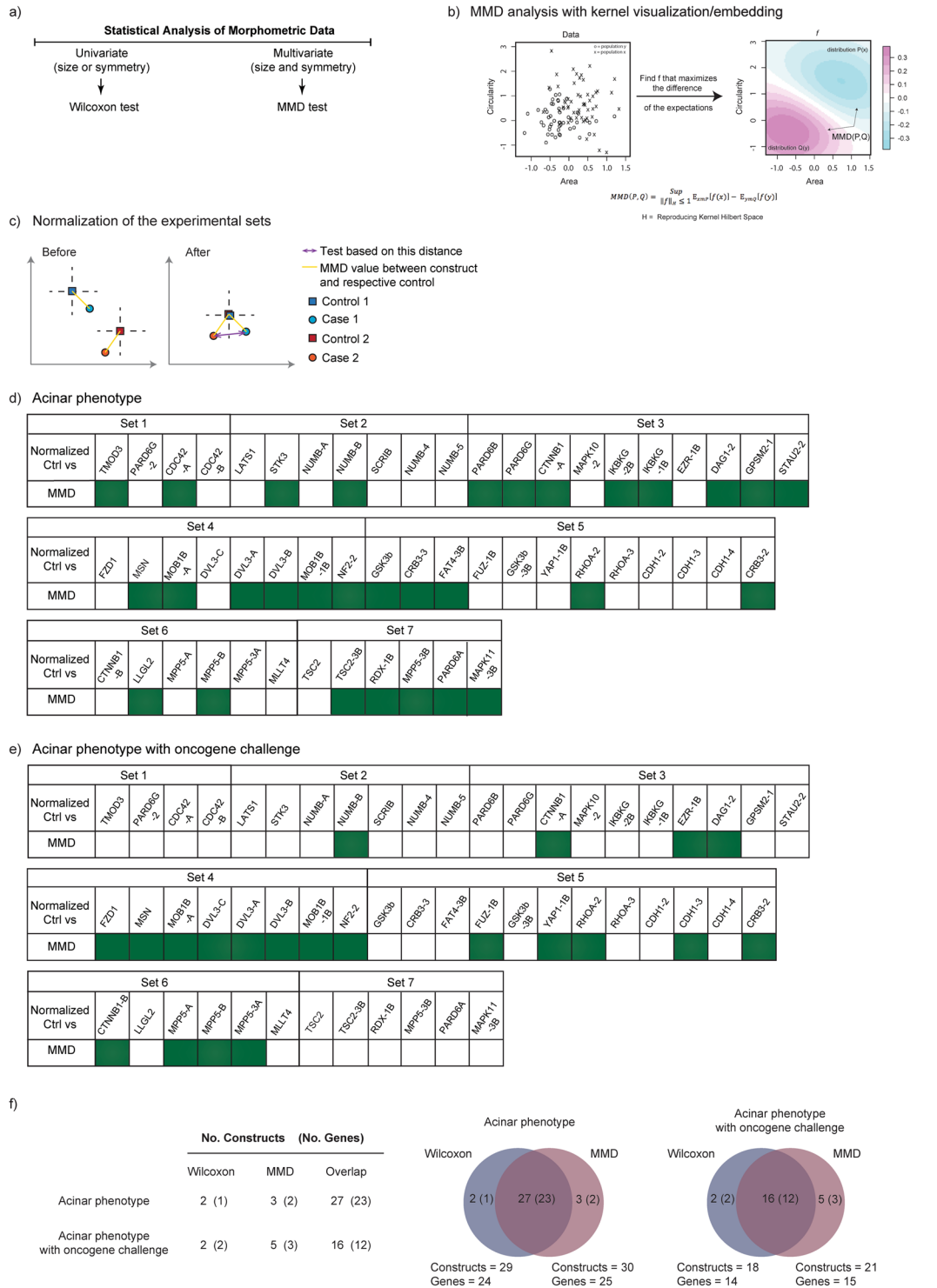
**Figure 2.** Defining phenotypically altered populations by univariate Wilcoxon Rank-Sum statistical test. **(a)** The whole shRNA screen was performed in seven independent experimental sets, with independent controls for each set. The highlighted cells denote shRNAs, which led to development of structures with significantly different size (green) and/or symmetry (yellow) as compared to control. **(b)** As in A, but the data represents acinar populations exposed to concomitant MYC activity and shRNA mediated gene silencing. **(c,d)** Categorization of the shRNAs based on their impact on the 3D morphology. List of shRNAs that showed significant impact on the morphology **(a,b)** and the subcategories that indicate the direction of the phenotypic change. \* highlights the phenotypes that require attention (Supplemental Table 1). Nonparametric Wilcoxon Rank Sum test; FDR < 0.05.

whereas chronic MYC activity prevented the cell cycle exit<sup>14</sup>. Surprisingly, we observed that the knockdown of numerous genes improved the symmetry of acinar structures (Fig. 2c,d). This could happen, since the control 3D MCF10A acinar structures show certain degree of heterogeneity and the average is not a perfect symmetry. Altogether, the analysis demonstrates that the shRNA mediated silencing of hEIR genes leads to morphologically altered 3D phenotypes, which are sufficiently prevalent in unselected acinar populations for statistical analysis of inter-population differences.

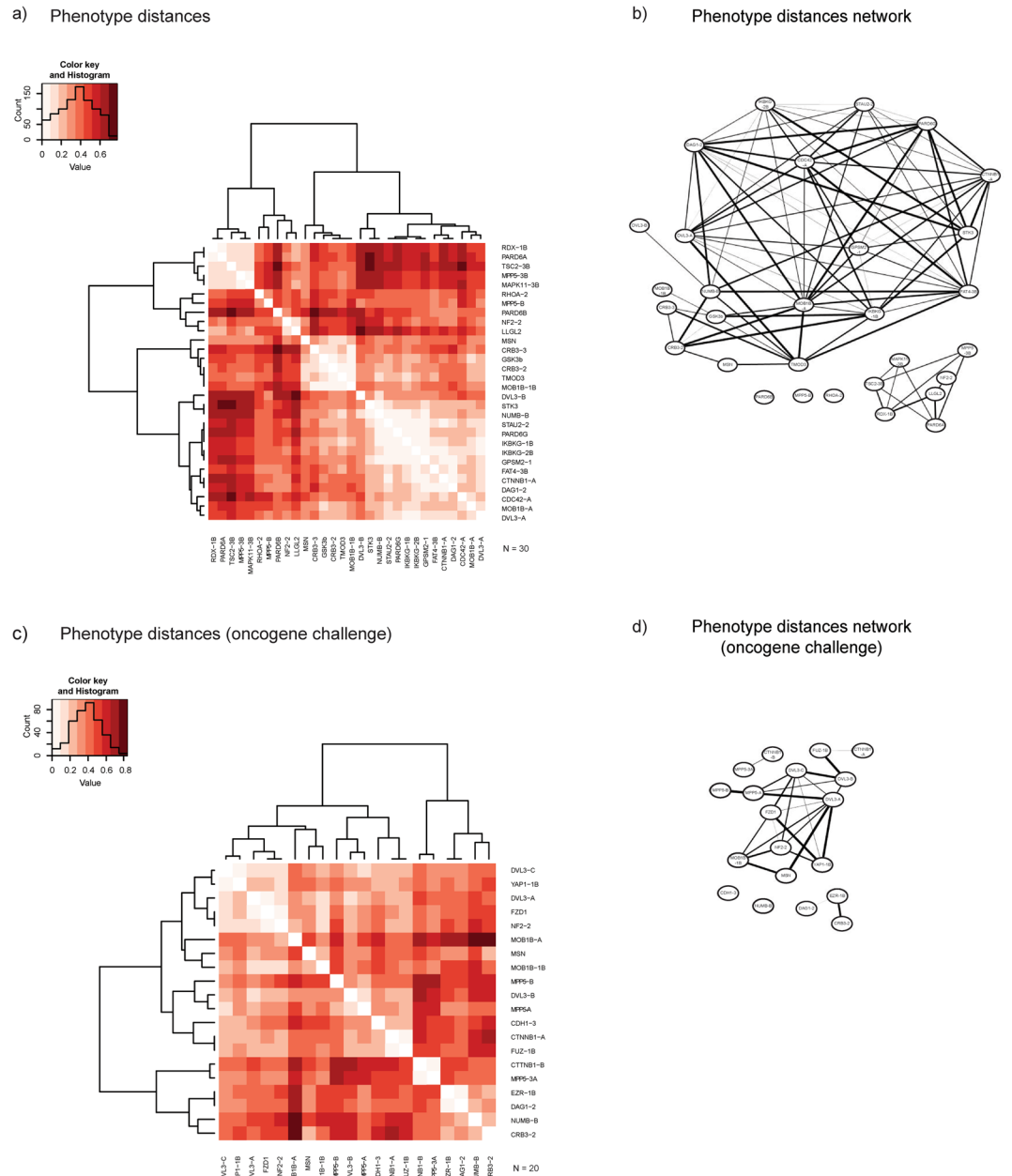
**Maximum mean discrepancy of the population distributions.** The raw data examination indicated that each shRNA with discernible effect on the acinar phenotype also caused a broad distribution of the phenotype defining (size or symmetry) values within the population (Fig. 1c–g). These findings together with the notion that only a minority of the shRNA silenced populations were normally distributed were consistent with a marked overall intra-population heterogeneity. The high degree of population heterogeneity was somewhat expected finding since shRNA mediated gene silencing is not 100% penetrant. Since a change in population distribution was a consistent shRNA's impact defining feature, we applied the maximum mean discrepancy (MMD) test to scrutinize the data population distributions in multivariate setting (Fig. 3a)<sup>21</sup>. The non-parametric MMD analysis (see Materials and Methods for mathematical details) is a kernel based test designed to analyse and compare probability distributions, which has recently found several applications in the analysis of multivariate data produced by molecular biology and omics approaches<sup>17,21</sup>. The MMD analysis is defined in terms of embedding the data probability distributions into an (implicit) infinite-dimensional feature space. Representation of the distributions in a reproducing kernel Hilbert space preserves their statistical features and allows computing distances and testing differences between them. In more practical terms, multivariate observations from two data populations are empirically compared by computing their similarities based on the chosen kernel function. For each pair of our populations, the MMD value was obtained using all the individual data-points in the populations and considering their size and symmetry values as multivariate observations (Fig. 3b). Using the MMD test, we first compared each gene-silenced population against its set-specific control with the null hypothesis that distributions of 3D acinar structures between two populations are equal against the alternative hypothesis that the distributions are different. The Benjamini-Hochberg procedure was used for multiple testing correction, defining significance at false-discovery rate (FDR) < 0.05<sup>19,20</sup>. To compare the 3D acinar populations across all the independent experimental sets, the control samples of each set were used for normalization. In particular, we centred the empirical embedding of the gene-silenced data populations in each experimental set by overlapping the embedding of their set-specific control, preserving the relative distances of the constructs in each independent experimental set (Fig. 3c). Altogether, 30 shRNA constructs (25 genes) were identified that resulted in significantly different population distributions from control (Fig. 3d, highlighted in green). In the presence of MYC oncogene challenge, 21 shRNA constructs (15 genes) resulted in significantly different population shape (Fig. 3e). We next examined whether the results from the two different statistical analysis (Wilcoxon Rank-Sum and MMD test) are consistent in identifying significant differences between the 3D acinar populations. The Venn diagram representations show an almost perfect overlap between the gene hits identified by the two different statistical analyses performed, thus validating the MMD approach (Fig. 3f).

**Deciphering gene function via population distribution patterns.** Importantly, the MMD test provided a distance metric to evaluate the relationships between different population distributions. We used both heatmap and network representation to analyse and visualize the MMD defined distances. The distance relationships of the 30 populations showing significant distances from their respective control (Benjamini-Hochberg: FDR < 0.05) are presented as heatmap in Fig. 4a and as a MMD network in Fig. 4b (network edges pruned by MMD < 0.1). Note that in the heatmaps (Fig. 4a and c) the dark red colour indicates longer distance (high MMD values) and the pale red closeness (low MMD values). In the network (Fig. 4b), each node represents shRNA silenced genes and the edges/interactions represent the MMD values (thinner line width means shorter distance). These analyses of population distribution distances exposed a number of short distance separated genes, which interestingly encode proteins of discrete signalling modules. These genes encoded, well known mediators of Wnt signalling pathway (Dishevelled/DVL; beta-catenin/CTNNB1; GSK3beta), Hippo pathway (Serine/Threonine kinase 3/STK3 (MST2); MOB kinases activator 1B/MOB1B) as well as proteins implicated in cell polarity regulation (Membrane palmitoylated protein 5/MPP5; PAR-6 family cell polarity regulators/PARD6beta and gamma; Lethal Giant Larvae Homolog 2/LLGL2; Crumbs 3/Crb3; CDC42). The same analysis and visualization methods were applied to the data populations representing simultaneous hEIR gene knockdown and MYC activation (oncogene challenge) (Fig. 4c,d). The analyses of shRNA perturbed populations grown under oncogene challenge also indicated a prominent role for Wnt and Hippo signalling pathways in growth regulation, showing that loss of these signalling modules contributes to the oncogenic MYC function in transformation of the 3D structures.

**Integrating phenotype-associated genes with proteomic data to find novel biochemical pathways.** The MMD analysis of population distribution distances and the results from Wilcoxon test, which provided information about the direction of the morphological change, together offered data to further explore the relationship between the genes found in the present study and different biological pathways. It is well-established that the cell polarity and epithelial integrity are dynamic processes critically controlled by evolutionary conserved protein complexes, for example the PAR complex, formed of aPKC-PARD3-PARD6-CDC42 proteins, and the SCRIB complex composed of SCRIB-LLGL2-DLG<sup>2,22–24</sup>. To layer proteomic information on top of the current genetic and phenotypic data, we determined the presence of hEIR gene hits in the STRING database<sup>25</sup>. This well-known proteomic database collects information of the known and predicted protein protein interactions covering global interactome studies, experiments and published literature. The query showed that hEIR genes are



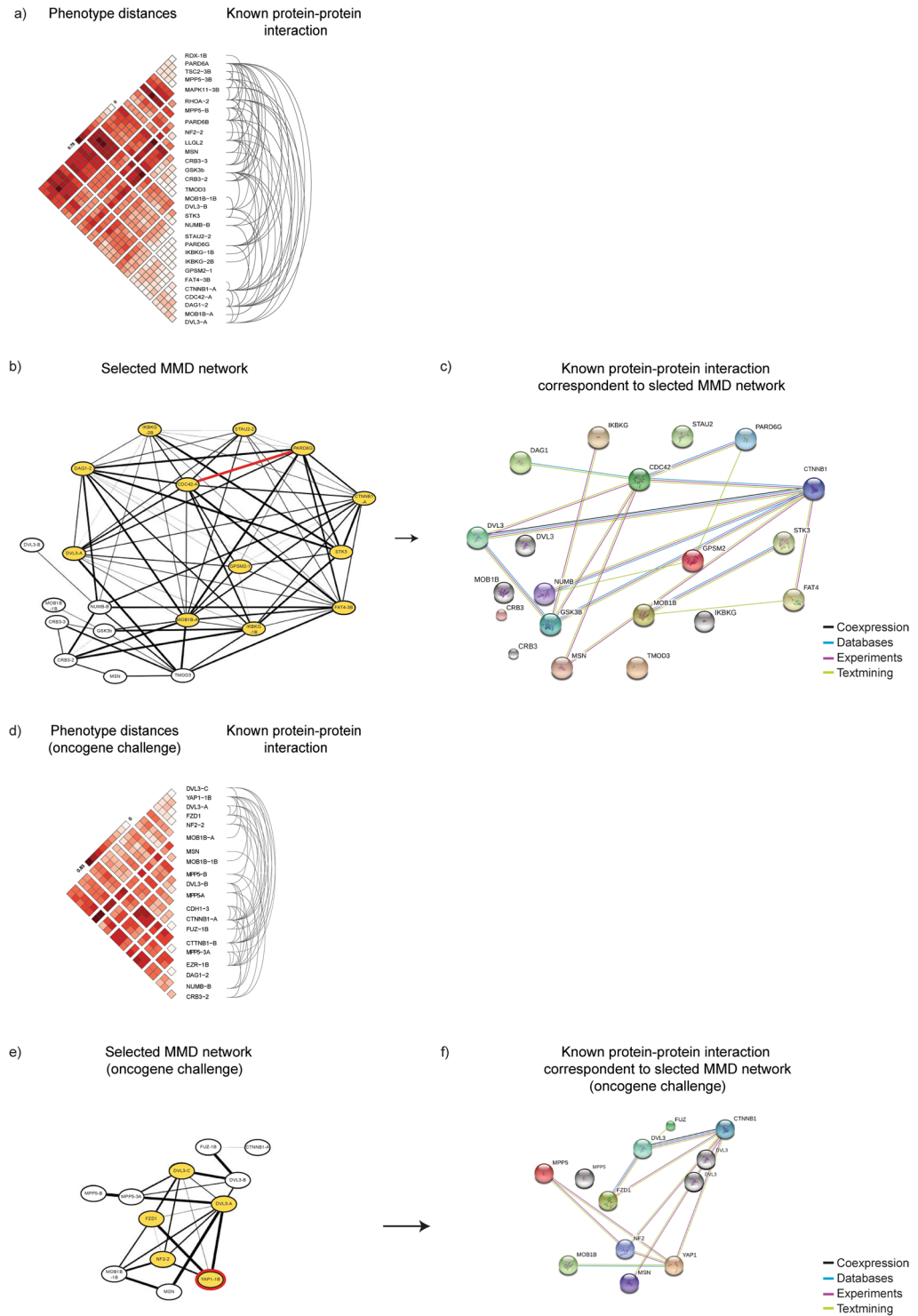
**Figure 3.** Maximum Mean Discrepancy (MMD) analysis of the data population distributions. **(a)** Diagram depicts the two statistical approaches used in this study to analyse the morphometric data. **(b)** Schematic representation of Maximum Mean Discrepancy analysis. The data are first plotted in a multivariate way, next a function that maximizes the distance between populations is applied and finally, the witness function illustrating the difference in the two populations is plotted. **(c)** Schematic representation of the normalization procedure of independent experimental sets. The populations are mean-centred using their respective control populations. The normalization procedure preserved the relative distances between the controls and the populations exposed to shRNA mediated gene silencing. **(d,e)** Summary tables showing the shRNAs that with or without oncogene challenge resulted in populations exhibiting significantly different distribution (MMD) in comparison to normalized control (FDR < 0.05). **(f)** Table and Venn Diagram illustrate concordance between the results from Wilcoxon Rank-Sum test and MMD analysis.



**Figure 4.** Heatmap and network representation of the MMD distances. **(a)** Heatmap representation of the populations significantly different from the control ( $n = 30$ ). In heatmap, the colour gradient from light to dark red reflects an increase in the populations distance (see inset). The data was hierarchically clustered to better visualize the phenotype relationships in the data matrix. **(b)** Network representation of the 3D populations based on the experimental-set-normalized pair-wise MMD values. The network was constructed with Cytoscape software<sup>48</sup>; each node depicts the shRNA resulting in the phenotypic alteration and the edge (interaction) lines thickness is inversely related to the closeness of the populations (thicker lines mean longer distance). The edges were pruned by implementing the threshold of  $MMD \leq 0.1$ . **(c)** Heatmap representation of the MYC challenged populations significantly different from the control ( $n = 20$ ). **(d)** Network representation of the MYC challenged 3D populations based on the experimental-set-normalized pair-wise MMD values.

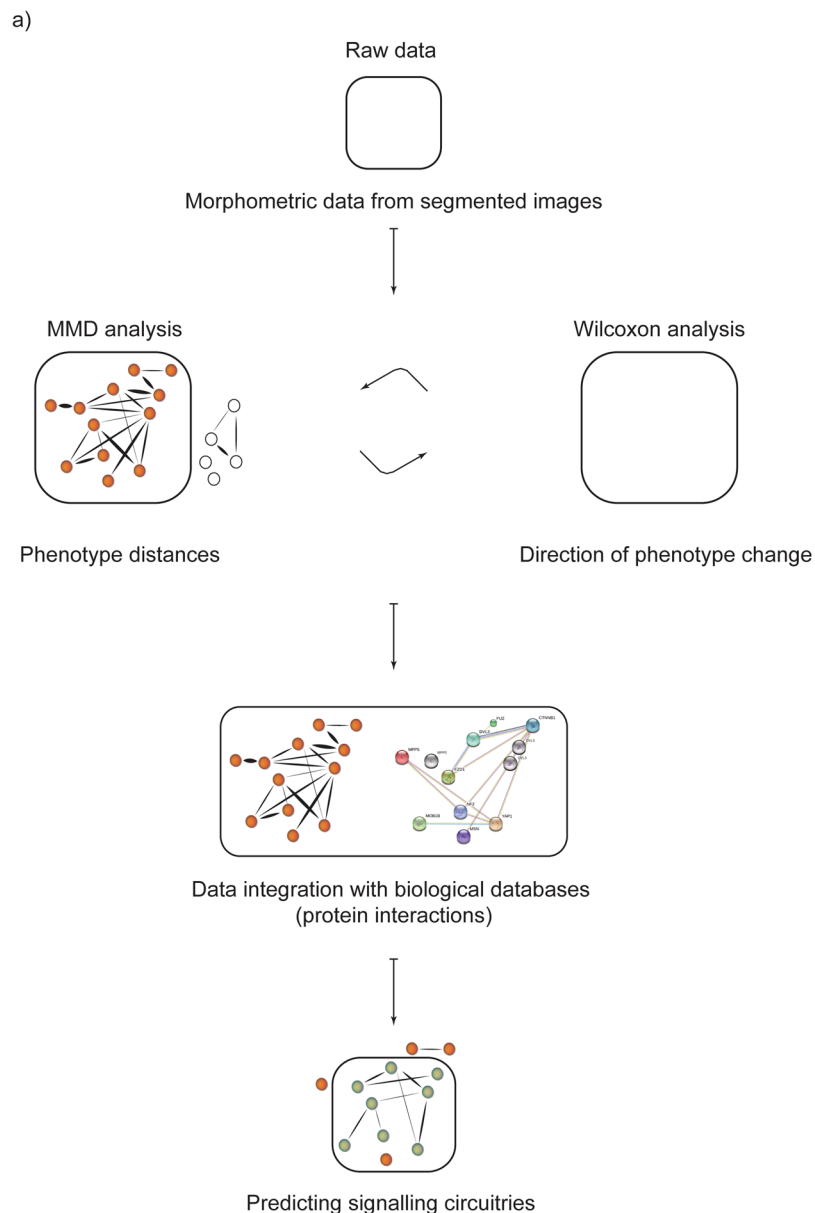
well represented in the STRING database and identified large number of protein interactions among the protein products of the hEIR genes (Fig. 5a).

To integrate data from our 3D phenotype screen with proteomic data, we first explored the network representation for “anchor interactions”, which means a subjectively chosen interaction of particular interest (starting point). In the present study, we chose to focus on PARD6G-CDC42 (red line) since the physical interaction between PAR6 and the small GTPase CDC42 form a well-established and critical component of the PAR complex<sup>26</sup>. The PAR complex has a general and conserved role in the regulation of asymmetric cellular processes<sup>27</sup>, which include establishment of the apical domain of the epithelial cells and orchestration of directional cell



**Figure 5.** Integrated analysis of MMD-based phenotypic distances together with proteomic data predicts novel biological pathways. **(a)** Epithelial phenotype critical genes contribute to the human protein interactome. The MMD-informed genes were queried against the genes present in STRING database of known protein-protein interactions. The figure demonstrates that the 3D phenotype altering genes identified in this study have multiple interactions with each other (arcs on the right side represent interactions described in the STRING database). **(b)** Anchor interaction analysis. Anchor denotes any interaction of primary interest. The cell polarity critical PAR6-CDC42 interaction is depicted as an anchor (red) and the first neighbours of the anchor are highlighted in yellow. **(c)** STRING network representation of protein interactions corresponding to the MMD clustered genes in **(b)**. **(d–f)** The figures as in **a–c** but here describing populations challenged with MYC oncogene. Instead of anchor interactions, a single anchor protein YAP-1 was selected.





**Figure 6.** Summary of the integrated 3D phenotypic and proteomic analysis workflow. Statistical analyses and data integration steps have been described in the text.

migration and asymmetric cell division<sup>28,29</sup>. As a second step, we highlighted the first neighbours (yellow nodes) of both PARD6G and CDC42 to visualize the 3D acinar populations exhibiting similar distances from normalized controls as PARD6G or CDC42 (Fig. 5b). As a third and last step, we compared the phenotype network with a STRING network of protein-protein interactions (compare Fig. 5b,c). In search of potentially interesting interactions supported by functional (present study) and biochemical (STRING) data, we noticed that PARD6G-CDC42 interaction shows closeness to Dishevelled Segment Polarity Protein 3/DVL3 and beta-catenin/CTNNB1. Both DVL3 and CTNNB1 are key mediators of the WNT signalling pathway, which plays a critical role in development and carcinogenesis<sup>30,31</sup>. On the other hand, the data suggests phenotypic closeness between PARD6G and MOB1B but also indicate that it is currently not known if the proteins directly interact in the cells.

We performed similar workflow of network analysis for the data populations representing the MYC challenged 3D structures (Fig. 5d-f). Exploration of the interactions of this smaller network did not reveal any obvious physical anchor interaction of interest. However, we considered YAP1 as an interesting anchor protein, since recent evidence indicates that the oncogenic MYC suppresses YAP/TAZ transcriptional co-activators<sup>32</sup>. Importantly, this MYC (up)-YAP (down) signalling axis limits the stem cell features of breast cancer cells, offering new avenues for therapeutic intervention. Our study shows that the YAP phenotype (reduced growth) is only exposed in the presence of MYC challenge, thus providing further evidence for a close signalling crosstalk between YAP and MYC. The parallel exploration of MMD and STRING networks highlighted an already known indirect interaction between YAP and NF2<sup>33,34</sup>. Neurofilament 2 (NF2) gene encodes a protein called Merlin,

which regulates at multiple points the Hippo signalling pathway<sup>35</sup>. The Hippo pathway is a conserved kinase pathway, which uses YAP/TAZ to control cell proliferation and tissue growth. Furthermore, dysregulation of NF2 and Hippo pathways have significant impact on the development of human cancer<sup>35–37</sup>. Our data predicts that in the presence of active MYC, loss of either NF2 or YAP activity may reduce the growth potential of the transformed structures and, speculatively, result in desired therapeutic effects on tumour cells with high MYC activity. These observations highlight the potential of MMD based phenotypic analyses of genetically altered 3D structures in search of testable hypothesis relevant for understanding epithelial biology and cancer. A summary of the analysis workflow is presented in Fig. 6.

## Discussion

Reverse genetics collectively refers to approaches aiming to interrogate the gene functions via analysis of the phenotypes caused by specifically engineered or mutated gene sequences<sup>38–40</sup>. The reverse genetic screens are most informative when the analysed phenotypes are easy to discern, quantitatively measured and highly penetrant. These conditions are often met in screens using single cells, such as yeast cells or simple genetically tractable organisms such as *C. elegans*, *Drosophila* or Zebrafish<sup>41,42</sup>. In mammalian systems, the complexity of tissue structures and long life-span of the organisms have long restricted genetic screens to simple 2D petri-dish based cell cultures, which are highly reductionist cell systems with respect to the highly organized 3D tissue structures of epithelial cells *in vivo*. However, the recent emerge of mammalian epithelial 3D cultures, which preserve or mimic tissue cohesion, cell-cell and cell-matrix interactions as well as many physical variables of the microenvironment, have offered new possibilities to broaden our understanding of the context-dependent regulation of individual cell functions<sup>43</sup>. Nonetheless, reverse genetic screens in 3D culture present many unanticipated technical hurdles not only for the experimental design but also for data analysis.

In the present study, we introduce a novel statistical framework for the analysis of 3D culture phenotype data. We accessed the data originating from an shRNA screen in MCF10A 3D culture, which employed 53 validated shRNAs designed to target 34 genes. Over 6000 digital images of MCF10A structures were analyzed. The multivariate statistical framework axiomatically differs from the previous analysis by Marques *et al.*<sup>14</sup>, which used averages and fold-changes separately for the phenotypic traits (size and symmetry) as the statistics to investigate the data. In the present study, following the investigation of the shape of the statistical distributions of the traits, and their evident non-normality, we based the analysis on non-parametric statistics, the Wilcoxon rank-sum test and the maximum mean discrepancy (MMD) statistic. MMD is a state-of-the-art statistical method for two-sample testing based on an elegant mathematical theory<sup>21</sup> and is widely applicable to different types of data (including structured data, such as strings and graphs), providing extensibility<sup>18,21</sup>. MMD considers the distribution of all the measured traits simultaneously to quantify the differences in the populations, and, it provides a metric space for the population distances. The test has previously been used in biological analysis, for example, to integrate biomarker data<sup>18</sup>, to minimize measurements variations when data are acquired in different experimental replicates and/or batches<sup>44</sup>, and to identify similar gene expression patterns in different experimental conditions<sup>17</sup>. In the current study, the results from MMD analysis agreed well with the traditional univariate Wilcoxon Rank-Sum test, which is a non-parametric equivalent of the paired t-test. These statistical approaches, when used in integrated manner, enabled us to identify the most significant phenotypic changes, organize them according to the direction of the change (small, large, loss of symmetry, improved symmetry), cluster closely related distributions and superimpose these clusters with global data on biomolecular interactions. This workflow of analyses, which is summarized in Fig. 6, predicted novel cell signaling pathways around PARD6-CDC42 cell polarity regulating complex. Furthermore, the integrated analysis suggested Hippo pathway effectors being essential for MYC-dependent deregulation of the 3D structures growth.

It is important to note, that even though the present study employed pre-validated shRNA species, there are number of remaining error sources to be considered when interpreting the data. For example, RNA interference mechanism inherently produces gene knockdown with non-congruous cell-to-cell efficacy rather than complete gene knockout<sup>45</sup>. Furthermore, the 3D structures develop from one cell and occasionally through fusions of neighbouring structures<sup>14</sup>. Thus, only a certain fraction of the shRNA-transduced individual 3D acini will exhibit the maximal phenotypic change. In addition, many phenotypic effects were observed only with a single shRNA species and the primary screen data represent a single experiment rather than biological replicates. Due to these fallibilities, which apply for any shRNA screen in 3D culture, the phenotype distance networks and related biochemical data, STRING protein-protein interactions provided here as an example, should be interpreted with caution. Nonetheless, the Wilcoxon-MMD statistical framework is expected to have many practical applications. It provides a set of tools for hypothesis generation on the basis of the primary screening data and, we believe that there is an increasing need for such tools especially, in the analysis of large datasets. Completion of a screen always requires further validation of the hits, but especially in the case of phenotype based screens, it is not always evident what to prioritize in hit picking. The hits are often prioritized on a subjective basis (interesting gene or pathway) or according to the strength of the phenotypic change. However, these approaches are prone to confirmation bias. The statistical framework presented in this study, which integrates Wilcoxon rank-sum and MMD tests, serves as a versatile solution for hypothesis-guided hit selection. In the presented solution, one can choose the first gene hit or gene interaction on a subjective basis and, subsequently, select neighborhood genes on the grounds of putative phenotype closeness. The hypothesis-guided hit picking does not exclude other approaches, but generates opportunity to identify novel phenotype impacting genetic interactions around the original gene of interest. The presented statistical system for 3D culture data analysis is generalizable and can be applied to deal with the large datasets from high-throughput analysis of miniaturized 3D cultures. However, it is stressed again, that the primary screen derived hypothetical interactions will have to be validated with appropriate laboratory analyses. In summary, the solution presented in our study is expected to facilitate organization and analysis

of the data from genetic screens that employ 3D culture systems. Furthermore, the Wilcoxon-MMD statistical framework provides an explorative platform that allows integration of phenotypic data with different biological knowledgebase for generation of new testable hypotheses on epithelial pathway functions.

## Material and Methods

**Construction of shRNA library, cell biological experiments and data acquisition.** The shRNA library targeting the Human Epithelial Integrity Regulator genes (hEIR), experimental methods related to the MCF10A 3D culture and image acquisition procedures are described in Marques (2016)<sup>14</sup>.

**Statistical tests.** Kolmogorov-Smirnov and Shapiro-Wilk normality tests<sup>46</sup> were performed using SPSS Statistics Ver. 22.0.0.1.

The non-parametric Wilcoxon rank sum test was used to define significant size and symmetry differences within the independent experimental sets followed by multiple test correction using the Benjamini Hochberg (FDR) method. The threshold  $FDR < 0.05$  was used to define hits. All tests were performed using the R software for statistical computing<sup>47</sup>.

**MMD analysis of the 3D population distributions.** *Preprocessing.* Observed morphometric parameters, area and circularity, were log-transformed and logit-transformed respectively. Transformed values were then scaled to have unit variance over the aggregate of all control and knockdown samples. The raw data processing was performed independently for the samples with or without oncogene challenge.

**Data analysis.** A non-parametric two-sample statistical test was adopted to calculate the maximum mean discrepancy (MMD) between two data populations of interest. The MMD statistic developed by Gretton 2012<sup>21</sup> restricts the function choice to a certain function class (unit ball in a reproducing kernel Hilbert space (RKHS)) and allows estimating MMD as averages of kernel evaluations over all the observed samples.

To test the difference between two data populations, the following MMD statistics was applied:

$$MMD = \sup_f E_x[f(x)] - E_y[f(y)] = \|\mu_x - \mu_y\|, \quad (1)$$

where  $x$  and  $y$  represent the populations distributions and  $f$  is defined to maximize the difference between them. The statistic is the distance between the mean embedding  $\mu_x$  and  $\mu_y$  of the two population distributions in the RKHS (they are expectations of the kernel feature space over the distributions). We used the MMD with Gaussian kernel where the length-scale is chosen with the median distance heuristic over aggregate samples. The null distribution for the hypothesis that  $x$  and  $y$  have the same distribution is generated by simulation (permutation testing) and the p-value is estimated from  $M$  permutations as:

$$p - value = \frac{\sum_i^M I(MMD_i^2 > MMD^2) + 1}{M + 1}, \quad (2)$$

where  $MMD_i^2$ ,  $i = 1, \dots, M$  are the permutation statistics and  $MMD^2$  is the observed test statistic.

Since multiple statistical tests were performed, a false discovery rate (FDR) control defining a p-value cut-off was applied as follow:

$$FDR = E\left[\frac{T}{T + F}\right], \quad (3)$$

where  $T$  is the number of true discoveries and  $F$  is the number of false discoveries ( $T + F$  is the total number of tests declared significant). A test is defined as interesting knockdowns if the FDR corrected p-value is  $< 0.05$ <sup>19</sup>.

**MMD distance based network construction.** To normalize MMD estimates across the independent experimental sets, the MMD statistics was extended as:

$$\begin{aligned} MMD &= \sup_f (E_x[f(x)] - E_A[f(a)] - (E_y[f(y)] - E_B[f(b)])) \\ &= \|(\mu_x - \mu_A) - (\mu_y - \mu_B)\| \end{aligned} \quad (4)$$

where  $A$  and  $B$  are the respective control populations for  $X$  and  $Y$ . If  $X$  and  $Y$  are from the same experimental set,  $\mu_A = \mu_B$  and the MMD statistic simplifies to the previous definition. Otherwise, the independent experiments correction will mean-centre the two samples using their respective control means in the RKHS. The networks were built greedily by connecting the two knockdowns with the smallest MMD and so on until a set threshold. After computing the MMD values for each pair of populations, we clustered the populations based on the MMD values using hierarchical clustering with Ward's criterion (R function `hclust` with `ward.D` method). Using Cytoscape software<sup>48</sup>, we produced pruned heatmaps and networks, where we only include populations that show different phenotype from the controls (network edges pruned by  $MMD \leq 0.1$ ). To integrate our population distances to known protein-protein interactions we used the STRINGdb (version 10) with confidence level 0.7 and source of evidence set as in<sup>49</sup>.

## References

- Bryant, D. M. & Mostov, K. E. From cells to organs: building polarized tissue. *Nat Rev Mol Cell Biol* **9**, 887–901 (2008).
- Assemat, E., Bazellieres, E., Palesi-Pocachard, E., Le Bivic, A. & Massey-Harroche, D. Polarity complex proteins. *Biochim Biophys Acta* **1778**, 614–630 (2008).
- O'Brien, L. E., Zegers, M. M. & Mostov, K. E. Opinion: Building epithelial architecture: insights from three-dimensional culture models. *Nat Rev Mol Cell Biol* **3**, 531–537 (2002).
- Lee, M. & Vasioukhin, V. Cell polarity and cancer—cell and tissue polarity as a non-canonical tumor suppressor. *J Cell Sci* **121**, 1141–1150 (2008).
- Royer, C. & Lu, X. Epithelial cell polarity: a major gatekeeper against cancer? *Cell Death Differ* **18**, 1470–1477 (2011).
- Runkle, E. A. & Mu, D. Tight junction proteins: from barrier to tumorigenesis. *Cancer Lett* **337**, 41–48 (2013).
- Ivanov, A. I. & Naydenov, N. G. Dynamics and regulation of epithelial adherens junctions: recent discoveries and controversies. *Int Rev Cell Mol Biol* **303**, 27–99 (2013).
- Grassian, A. R., Coloff, J. L. & Brugge, J. S. Extracellular matrix regulation of metabolism and implications for tumorigenesis. *Cold Spring Harb Symp Quant Biol* **76**, 313–324 (2011).
- Kaushik, S., Pickup, M. W. & Weaver, V. M. From transformation to metastasis: deconstructing the extracellular matrix in breast cancer. *Cancer Metastasis Rev* **35**, 655–667 (2016).
- McCaffrey, L. M. & Macara, I. G. Epithelial organization, cell polarity and tumorigenesis. *Trends Cell Biol* **21**, 727–735 (2011).
- Bissell, M. J. & Hines, W. C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nat Med* **17**, 320–329 (2011).
- Choi, J., Ilich, E. & Lee, J. H. Organogenesis of adult lung in a dish: Differentiation, disease and therapy. *Dev Biol* **420**, 278–286 (2016).
- Muranen, T. *et al.* Inhibition of PI3K/mTOR leads to adaptive resistance in matrix-attached cancer cells. *Cancer Cell* **21**, 227–239 (2012).
- Marques, E. *et al.* Par6G suppresses cell proliferation and is targeted by loss-of-function mutations in multiple cancers. *Oncogene* **35**, 1386–1398 (2016).
- Gobeil, S., Zhu, X., Doillon, C. J. & Green, M. R. A genome-wide shRNA screen identifies GAS1 as a novel melanoma metastasis suppressor gene. *Genes Dev* **22**, 2932–2940 (2008).
- Leung, C. T. & Brugge, J. S. Outgrowth of single oncogene-expressing cells from suppressive epithelial environments. *Nature* **482**, 410–413 (2012).
- Vegas, E., Oller, J. M. & Reverter, F. Inferring differentially expressed pathways using kernel maximum mean discrepancy-based test. *BMC Bioinformatics* **17**(Suppl 5), 205 (2016).
- Borgwardt, K. M. *et al.* Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics* **22**, e49–57 (2006).
- Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**, 279–284 (2001).
- Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A Kernel Two-Sample Test. *J. Mach. Learn. Res.* **13**, 723–773 (2012).
- Campanale, J. P., Sun, T. Y. & Montell, D. J. Development and dynamics of cell polarity at a glance. *Journal of Cell Science* **130**, 1201–1207 (2017).
- Hirsh, E. & Sharan, R. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics* **23**, e170–176 (2007).
- Insolera, R., Chen, S. & Shi, S. H. Par proteins and neuronal polarity. *Dev Neurobiol* **71**, 483–494 (2011).
- Szklarczyk, D. *et al.* STRINGv10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447–452 (2015).
- Tervonen, T. A. *et al.* Faulty epithelial polarity genes and cancer. *Adv Cancer Res* **111**, 97–161 (2011).
- Goldstein, B. & Macara, I. G. The PAR proteins: fundamental players in animal cell polarization. *Dev Cell* **13**, 609–622 (2007).
- Petrie, R. J., Doyle, A. D. & Yamada, K. M. Random versus directionally persistent cell migration. *Nat Rev Mol Cell Biol* **10**, 538–549 (2009).
- Goulas, S., Conder, R. & Knoblich, J. A. The Par complex and integrins direct asymmetric cell division in adult intestinal stem cells. *Cell Stem Cell* **11**, 529–540 (2012).
- Sheikh, A. *et al.* The role of Wnt signaling pathway in carcinogenesis and implications for anticancer therapeutics. *Hered Cancer Clin Pract* **12**, 13 (2014).
- Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473 (2017).
- von Eyss, B. *et al.* A MYC-Driven Change in Mitochondrial Dynamics Limits YAP/TAZ Function in Mammary Epithelial Cells and Breast Cancer. *Cancer Cell* **28**, 743–757 (2015).
- Reginensi, A. *et al.* A critical role for NF2 and the Hippo pathway in branching morphogenesis. *Nat Commun* **7**, 12309 (2016).
- Zhang, N. *et al.* The Merlin/NF2 tumor suppressor functions through the YAP oncoprotein to regulate tissue homeostasis in mammals. *Dev Cell* **19**, 27–38 (2010).
- Yu, F. X., Zhao, B. & Guan, K. L. Hippo Pathway in Organ Size Control, Tissue Homeostasis, and Cancer. *Cell* **163**, 811–828 (2015).
- Petrilli, A. M. & Fernandez-Valle, C. Role of Merlin/NF2 inactivation in tumor biology. *Oncogene* **35**, 537–548 (2016).
- Guerrant, W. *et al.* YAP Mediates Tumorigenesis in Neurofibromatosis Type 2 by Promoting Cell Survival and Proliferation through a COX-2-EGFR Signaling Axis. *Cancer Res* **76**, 3507–3519 (2016).
- Bernacki, S., Karimi, M., Hilson, P. & Robertson, N. Virus-induced gene silencing as a reverse genetics tool to study gene function. *Methods Mol Biol* **655**, 27–45 (2010).
- Gartel, A. L. & Kandel, E. S. RNA interference in cancer. *Biomol Eng* **23**, 17–34 (2006).
- Cho, K. H. *et al.* Reverse engineering of gene regulatory networks. *IET Syst Biol* **1**, 149–163 (2007).
- Shah, A. N., Davey, C. F., Whitebirch, A. C., Miller, A. C. & Moens, C. B. Rapid reverse genetic screening using CRISPR in zebrafish. *Nat Methods* **12**, 535–540 (2015).
- MB, T. & KH, L. An Introduction to Reverse Genetic Tools for Investigating Gene Function. *The Plant Health Instructor* (2005).
- Nelson, C. M. & Bissell, M. J. Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. *Annu Rev Cell Dev Biol* **22**, 287–309 (2006).
- Shaham, U. *et al.* Removal of Batch Effects using Distribution-Matching Residual Networks. *Bioinformatics* (2017).
- Kampmann, M. *et al.* Next-generation libraries for robust RNA interference-based genome-wide screens. *Proc Natl Acad Sci USA* **112**, E3384–3391 (2015).
- Ghasemi, A. & Zahediasl, S. Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab* **10**, 486–489 (2012).
- R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria 2013).
- Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
- Franceschini, A. *et al.* STRINGv9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808–815 (2013).

## Acknowledgements

We thank Biomedicum Functional Genomics Unit (FuGu), Biomedicum Imaging Unit (BIU) for services; Tiina Raatikainen and Tarja Välimäki for technical support; Julia Casado for reading and critical comments on the manuscript. This work was funded by the Academy of Finland, TEKES and the Finnish Cancer Organizations; E.M. was funded by the Integrative Life Sciences Doctoral Program (ILS), DSHealth doctoral school and Biomedicum Helsinki Foundation (Young scientist's grants).

## Author Contributions

E.M. generated raw data, designed and performed the experiments, performed the statistical analysis, wrote the manuscript. T.P. performed the MMD statistical analyses and contributed to writing of the manuscript. S.K. designed and supervised the statistical analysis. J.K. supervised the study, designed the experiments and wrote the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-22293-x>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018