pheno-seq – linking 3D phenotypes of clonal tumor spheroids to gene expression

Stephan M. Tirier[1,2], Jeongbin Park[1,3], Friedrich Preußer[1,2,4], Lisa Amrhein[5,6], Zuguang Gu[1,8], Simon Steiger[1,2], Jan-Philipp Mallm[2,7,8], Marcel Waschow[1,2], Björn Eismann[1,2], Marta Gut[9,10], Ivo G. Gut[9,10], Karsten Rippe[2,7], Matthias Schlesner[1,11], Fabian Theis[5,6], Christiane Fuchs[5,6,12], Claudia R. Ball[13], Hanno Glimm[13,14], Roland Eils[1,2,3,8,15], Christian Conrad[1,2,8,†]*

Affiliations

[1]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany.

[2]Center for Quantitative Analysis of Molecular and Cellular Biosystems (BioQuant), University of Heidelberg, Heidelberg, Germany.

[3]Digital Health Center, Berlin Institute of Health (BIH)/Charité-Universitätsmedizin Berlin, Berlin, Germany

[4]Max Delbrück Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, Berlin, Germany (*current address)

[5]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Munich, Neuherberg, Germany

[6]Department of Mathematics, Technische Universität München, Munich, Germany

[7]Division of Chromatin Networks, German Cancer Research Center (DKFZ), Heidelberg, Germany.

[8]Heidelberg Center for Personalized Oncology, DKFZ-HIPO, DKFZ, Heidelberg, Germany.

[9]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain.

[10]Universitat Pompeu Fabra, Barcelona, Spain.

[11]Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany.

[12]Faculty of Business Administration and Economics, Bielefeld University, Bielefeld, Germany.

[13]Department of Translational Oncology, NCT Dresden, University Hospital, Carl Gustav Carus, Technische Universität Dresden, Dresden and DKFZ, Heidelberg, Germany

[14]German Cancer Consortium, Heidelberg, Germany.

[15]Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology (IPMB) Heidelberg University, Heidelberg, Germany.

[†]Corresponding author: c.conrad@dkfz.de

Keywords: 3D cell culture, tumor cell heterogeneity, single cell genomics, gene expression deconvolution, spheroid image analysis

Abstract

3D-culture systems have advanced cancer modeling by reflecting physiological characteristics of *in-vivo* tissues, but our understanding of functional intratumor heterogeneity including visual phenotypes and underlying gene expression is still limited. Transcriptional heterogeneity can be dissected by single-cell RNA-sequencing, but these technologies suffer from low RNA-input and fail to directly correlate gene expression with contextual cellular phenotypes. Here we present 'pheno-seq' for integrated high-throughput imaging and transcriptomic profiling of clonal tumor spheroids derived from 3D models of breast and colorectal cancer. Specifically, we identify characteristic expression signatures that are associated with heterogeneous invasive and proliferative behavior including a rare cell subtype. Furthermore, we identify functionally relevant transcriptional regulators missed by single-cell RNA-seq, link visual phenotypes defined by heterogenous expression to inhibitor response and infer single-cell regulatory states by deconvolution. We anticipate that directly linking molecular features with patho-phenotypes of cancer cells will improve the understanding of intratumor heterogeneity and consequently prove to be useful for translational research.

Introduction

Three-dimensional (3D) cell culture systems provide a physiologically relevant context for *in-vitro* testing, manipulation and high-throughput screening applications[1]. Mimicking the 3D-tissue environment thus holds great promise for future diagnostics[2] and the analysis of functional differences between tumor cells in a single patient (intratumor heterogeneity)[3], a phenomenon increasingly recognized as an essential driver of tumorigenic progression, treatment resistance and relapse[4].

While visual characteristics of 3D-cultures such as shape and size can be highly informative for classification of tumor subtypes and disease states, most studies have so far focused on inter-patient differences rather than heterogeneous behavior of cells derived from a single patient[5]. Single-cell 3D-culture combined with microscopy and molecular analyses appears as a key strategy for investigating cellular heterogeneity *in-vitro* as it enables analysis of clonal behavior in defined spatial and temporal conditions[6,7]. Ideally, the visual phenotype of the emerging multicellular complex (spheroids, organoids, etc.) reflects the characteristics of the primary tumor and consequently informs about the functional outcome of heterogeneous cancer cell states. Informative subpopulation-specific oncogenic phenotypes include long-term proliferative potential[8], or more complex phenotypes such as deregulation of epithelial growth and invasiveness, a prerequisite for metastasis[9]. Likewise, identifying heterogeneous gene expression in functionally distinct tumor populations is of particular importance to infer drivers of cell state transitions and to uncover underlying signaling pathways[10].

Recent technological advances in microscopy enable high-throughput phenotyping to quantitatively characterize cellular heterogeneity[11], but methods to directly associate observed cellular properties in 3D cultures with system-wide gene expression are still lacking. While laser capture microdissection (LCM)[12] or multiplexed fluorescence *in-situ* hybridization (FISH)[13] techniques are principally suited to linking cellular phenotypes to gene expression, the required histological preparation, elaborated sample processing and pre-selection of transcript-specific probes limit the applicability and resolution of those methods. On the other hand, the variety of recently developed technologies for single-cell RNA-sequencing (scRNA-seq)[14] suffer from low-input material and do not provide a direct link to visual cellular phenotypes since the available protocols involve dissociation of cells and loss of their multicellular context.

Here, we present 'pheno-seq' to dissect heterogeneity in 3D cell culture systems by directly combining clonal cell culture, imaging and transcriptomic profiling. We developed an experimental and computational workflow for unbiased high-throughput pheno-seq, including i) automated dispensing and imaging of single spheroids in barcoded nanowells; ii) an automated image processing pipeline; and iii) 'PhenoSelect' software for interactive analysis and selection of spheroids. We demonstrate the power of pheno-seq in dissecting both cellular and molecular heterogeneity for established and patient-derived 3D-models of breast and colon cancer, respectively.

Results

Pheno-seq enables direct phenotype correlation and complements the identification of heterogeneous gene expression in the 3D breast cancer model MCF10CA

In breast cancer, normal epithelial cells undergo a stepwise transformation from local hyperplasia to premalignant carcinoma *in-situ* and invasive carcinoma[15]. Importantly, the switch from epithelial to invasive behavior requires gene expression programs that resemble those occurring during embryogenesis and wound healing, generally described as epithelial-to-mesenchymal transition (EMT)[9].

Single-cell-derived spheroids of the breast cancer cell line MCF10CA[16] show remarkable morphological heterogeneity when cultured in 3D, with cellular phenotypes reflecting characteristics of both carcinoma *in-situ* ('round' phenotype) and invasive carcinoma ('aberrant') (Supplementary Fig. 1a and b). To enable independent analysis of cells derived from both phenotypes, we developed a workflow to isolate single spheroids from reconstituted basement membrane (Matrigel) without perturbing their phenotypic identity (Fig. 1a and b). To functionally assess the observed heterogeneity, we reseeded and cultured cells from both phenotype classes independently which validated efficient isolation and revealed a high cell state stability (Supplementary Fig. 1c).

To identify gene expression differences between round and aberrant spheroids, we first generated and deeply sequenced microfluidics-based full-length scRNA-seq libraries of both phenotypes independently (166 cells in total, Fig. 1c, Supplementary Table 1). Notably, this strategy does not enable a direct phenotypic correlation as cells from multiple spheroids needed to be pooled.

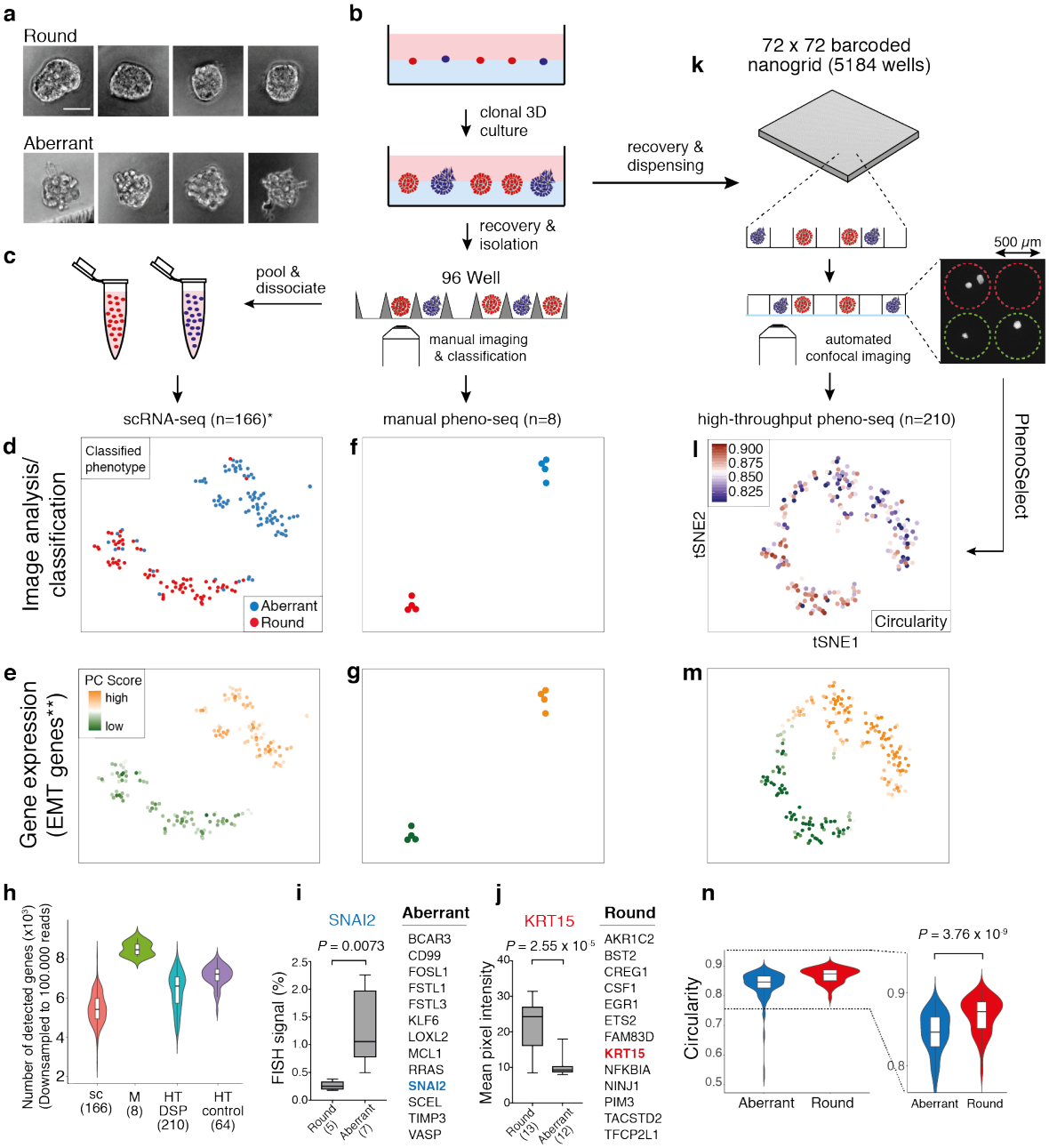


Figure 1 | pheno-seq enables direct image correlation and complements the identification of morphology-specific gene expression

(a) Brightfield images of clonal spheroids (MCF10CA phenotype classes 'round' and 'aberrant') after isolation from Matrigel (scale bar 50 µm). **(b)** Workflow overview for the isolation of clonal spheroids for inference of morphology-specific gene expression. **(c)** Indirect phenotype – gene expression correlation by scRNA-seq using single cells isolated from multiple spheroids with annotated morphology phenotype. **(d)** 2D tSNE visualization[17] of 166 scRNA-seq (*cell-cycle corrected) full-length expression profiles of cells from manually isolated round and aberrant spheroids with coloring based on manual phenotype annotation. **(e)** Same 2D tSNE visualization as shown in (d) but coloring based on PC scores for **HALLMARK_EMT gene set derived from the Molecular Signature Database (MSigDB)[18]. **(f and g)** 2D tSNE visualization of 8 full-length manual pheno-seq profiles based on manually isolated single spheroids. Same coloring as shown in (d) and (e). **(h)** Number of genes

detected in downsampled scRNA-seq and pheno-seq libraries (sc: scRNA-seq; M: manual pheno-seq; HT-DSP: high-throughput pheno-seq with dithio-bis(succinimidyl) propionate fixation; HT-control: HT-pheno-seq bottom control). Numbers of samples indicated on x-axis under respective method. **(i and j)** Selected genes only identified by manual pheno-seq and not by scRNA-seq and validation of phenotype-specific expression for SNAI2 (aberrant) and KRT15 (round). RNA-FISH for SNAI2: Plotted values reflect the faction of pixels that exceed the background threshold per spheroid. KRT15 immunofluorescence: Plotted values reflect mean pixel intensity per classified spheroid. Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values. Numbers of samples indicated on x-axis under respective phenotype class. Indicated are *P*-values from unpaired two-tailed Students t-test. **(k)** High-throughput (HT) pheno-seq workflow based on automated dispensing and confocal imaging of recovered spheroids in barcoded nanowells. **(l)** 2D tSNE visualization of 210 HT-pheno-seq 3´-end profiles with coloring based on image feature 'circularity'. For better visualization, all circularity values below 0.8 were set to minimum in the color code scheme. **(m)** Same 2D tSNE visualization as shown in (l) with coloring based on PC scores for **HALLMARK_EMT gene set as shown in (e) and (g). **(n)** Circularity plotted per cluster (k-means clustering, k=2) as shown in (l). Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR. Indicated *P*-value from unpaired two-tailed Students t-test.

Combined transcriptomic analysis by testing annotated and *de-novo* identified gene sets for coordinated expression variability[17] across cells revealed two distinct clusters and a tight association of cells to their original phenotype class (Fig. 1d), whereas differential expression analysis[19] identified biologically relevant expression patterns. Cells derived from aberrant spheroids are characterized by the expression of EMT related genes (Fig. 1e), including vimentin (VIM), Beta-Actin (ACTB) and fibroblast activating protein (FAP). In contrast, cells isolated from round spheroids showed higher expression of genes involved in adherence and formation of tissue structures including desmoglein 3 (DSG3) and keratin 16 (KRT16) (Supplementary Fig. 2a and 3a). In order to validate if we could accurately detect gene expression specific for invasive phenotypes, we used whole mount immunofluorescence (IF) of individual marker genes, in particular the EMT marker VIM and the cytoskeleton component ACTB (Supplementary Fig. 4a and b).

Current scRNA-seq methods are affected by low RNA input[14] and dissociation bias[20]. Accordingly, we next tested whether expression profiling of manually isolated whole spheroids (manual pheno-seq) might serve as a complementary approach to identify transcriptional differences between clonal spheroid phenotypes. Despite the loss of single-cell resolution, we reasoned that our approach should enhance accuracy by enabling the direct correlation of image phenotype to transcriptome, and at the same

time provide more RNA material for cDNA library preparation. Therefore, we started with low spheroid sample numbers in a tube-based setup to evaluate our ability to detect relevant heterogeneous gene expression that is missed by scRNA-seq.

Profiling of only eight spheroids by manual pheno-seq yielded a similar phenotype-specific distinction defined by high and low expression of EMT-related genes (Fig. 1f and g, Supplementary Fig. 2b). While the sample number was approximately 20 times lower (166 single-cells vs. 8 single spheroids), the gene detection rate per sample was significantly higher compared to scRNA-seq (Fig. 1h), and differential expression analysis revealed over 50 phenotype-specific genes in each of the two classes that could not be detected by scRNA-seq (Fig. 1i and j, Supplementary Fig. 3b). Although we detected more differentially expresses genes by scRNA-seq, most likely due to the much higher sample number, only pheno-seq identified the transcriptional EMT master regulator SNAI2[21] (aberrant) and keratin 15 (KRT15, round), a basal-myoepithelial marker in the mammary gland[22] (Fig. 1i and j, Supplementary Fig. 3c).

Phenotype-specific expression of SNAI2 and KRT15 was validated by RNA-FISH and immunofluorescence (IF), respectively (Fig.1i and j, Supplementary Fig. 4c and d). We reasoned that SNAI2 could not be identified by scRNA-seq due to its low expression (Supplementary Fig. 3c), a frequent phenomenon for transcriptional regulators[23]. Although KRT15 is one of the top markers for round spheroids detected by pheno-seq, the existence of residual $KRT15^{+}$ cells in aberrant spheroids (Supplementary Fig. 4c) seemed to mask the identification of KRT15 as phenotype-specific when single-cell profiles were analyzed. Remarkably, differential expression of KRT15 and SNAI2 could not be restored from scRNA-seq data by generating pseudo pheno-seq profiles from averaged single-cell expression (Supplementary Fig. 2c), indicating for the additional influence of dissociation bias[20] on KRT15 mRNA abundance. In summary, pheno-seq enables the direct correlation of clonal spheroid phenotypes and transcriptomes and complements scRNA-seq methods to identify expression differences between 3D phenotypes already with low sample numbers.

High-throughput pheno-seq in barcoded nanowells enables combined quantitative analysis of image features and gene expression

A major limitation of both scRNA-seq and manual pheno-seq is the non-quantitative and biased selection of spheroid phenotypes based on visual inspection by eye. In

addition, increasing the number of spheroids per pheno-seq experiment is necessary to comprehensively understand associations between visual phenotypes and gene expression in a particular 3D-culture model. Therefore, we developed high-throughput (HT) pheno-seq by repurposing the nanowell-based iCELL8 scRNA-seq system[24], a platform for integrated imaging and gene expression profiling of single cells, for the processing of spheroid samples of up to 100 µm in size. Key modifications included cellular fixation[25], altered chip setup, higher-resolution microscopy, an automated image-processing pipeline and the 'PhenoSelect' software for interactive analysis and selection of spheroids for sequencing (Fig. 1k, Supplementary Fig. 5, 6 and 7). These substantial technical adaptions had only minor influences on the gene detection rate, which fell in between scRNA-seq and manual pheno-seq (Fig 1h, Supplementary Table 1). MCF10CA HT-pheno-seq yielded very similar results as described, with two distinct clusters driven by expression of genes involved in EMT (VIM^+) as well as tissue formation ($KRT15^+$) but at higher throughput per experiment (n = 210) compared to manual pheno-seq (Fig 1l and m, Supplementary Fig. 8a). Both pheno-seq methods show good concordance in identifying differentially expressed genes between spheroid phenotypes (Supplementary Fig. 8b), despite unbiased capture of spheroids by HT-pheno-seq as well as differences in sample size, read depth and library structure (3´-end vs. full-length, Supplementary Table 1).

In contrast to scRNA-seq, HT-pheno-seq allows measurements of RNA abundance and image features from the same spheroid, which enabled straightforward association of genetic programs and complex visual phenotypes based on the fluorescence signal derived from a cytoplasmic dye (CellTracker Red). These included the morphology-related feature 'circularity', informing about (de)regulation of lobular development (Fig. 1l and n), and spheroid size, demonstrating a higher proliferative activity of epithelial cells (Supplementary Fig. 8c). In addition, pheno-seq linked negatively skewed pixel intensity distributions to round phenotypes (Supplementary Fig. 8d), indicative of increased cell density in round 3D phenotypes that leads to an increased proportion of high pixel intensity values derived from the cytoplasmic signal. Hence, HT-pheno-seq represents a new method that, unlike scRNA-seq, directly and quantitatively links heterogeneous visual phenotypes to underlying gene expression in a single experiment.

HT-pheno-seq of a patient-derived colorectal 3D model links proliferative capacity to cell type-specific expression signatures

We next set out to assess the functional correlation between visual phenotypes and gene expression in a physiologically relevant and patient-derived 3D model originally isolated from a liver metastasis of a colorectal cancer (CRC) patient. Similar to the phenotypic heterogeneity in the MCF10CA spheroids described above, functionally distinct subpopulations in 3D-cultures of CRC patients have previously been identified[8]. The reported heterogeneity in proliferative potential seems to be largely independent of mutational subclone diversity[26], thereby supporting the presence of a differentiation-like hierarchy in CRC[27]. As reseeding of cells from different classes of spheroid sizes (20-40 µm and >70 µm) revealed significant differences in spheroid forming capacity (Supplementary Fig. 9a), we hypothesized that specific stem- and differentiation-related transcriptional signatures should underlie these heterogenous proliferative phenotypes. To investigate this hypothesis, we performed HT-pheno-seq based on clonal CRC spheroids cultured in an inverse pyramidal-shaped microwell setup (Fig. 2a; Supplementary Fig. 9b).

Analysis of 95 HT-pheno-seq gene expression profiles confirmed two transcriptionally distinct clusters (Fig. 2b). Image analysis of the respective spheroids revealed a strong difference in spheroid size composition between both clusters (Fig. 2c) that does not influence library complexity (Supplementary Fig. 10b). Differential expression analysis showed that the first cluster ('small' phenotype) is enriched for genes involved in ribosomal activity (GO_RIBOSOME, FDR q-value 2.41×10^{-45}) as well as intestinal secretory lineage markers, including Trefoil Factor 3 (TFF3), KRT18 and SPINK4[28] (Fig. 2d). In contrast, the second cluster ('big' phenotype) is characterized by the expression of genes previously described to be involved in (i) stem cell maintenance (including CD44, MYC, NOTCH1, APP, MSI1 and ITGA6)[28,29], (ii) the formation of cell-cell junctions (including EPCAM, CLDN4, CDH1) and (iii) WNT signaling (ZNRF3, LGR4, JUN) (Fig. 2d). This signature showed a very high overlap with genes correlated with the major intestinal stem cell marker LGR5, including CD44, NOTCH1 and SMOC2 (Supplementary Fig. 10a). We validated sphere size-dependent expression for selected markers by quantitative RNA-FISH (Fig. 2e, Supplementary Fig. 11a-c).

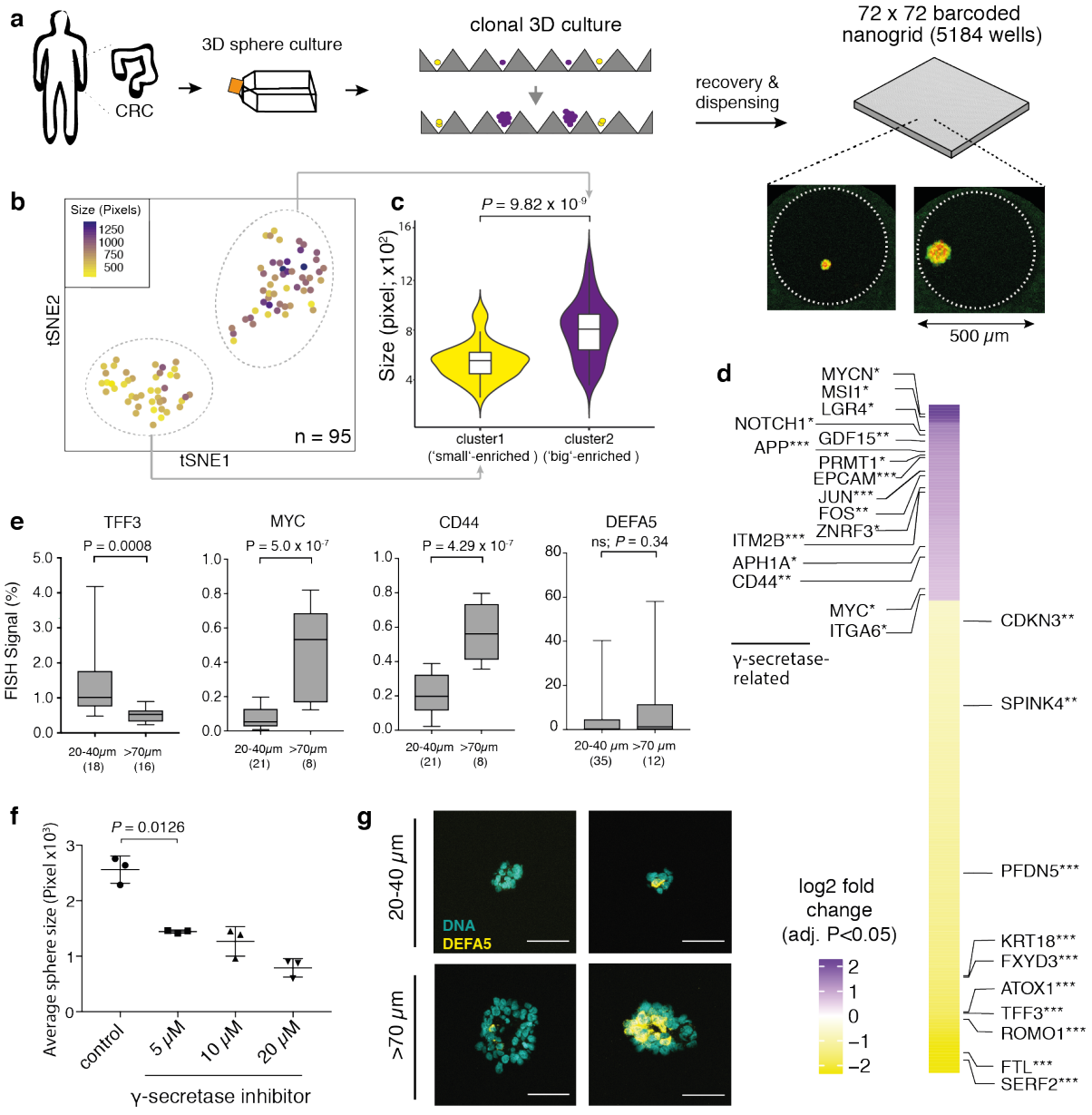


Figure 2 | pheno-seq of a 3D model of colorectal cancer links heterogeneous proliferative phenotypes to expression signatures enriched for cell type-specific markers

(a) Clonal 3D-culture in inverse pyramidal shaped microwells and recovery strategy for HT-pheno-seq of patient-derived CRC spheroids isolated from a liver metastasis. Yellow and purple indicate heterogeneous subpopulations with functional differences in proliferative potential[8]. **(b)** 2D tSNE visualization of 95 HT-pheno-seq expression profiles. Coloring by sphere size (pixel). **(c)** Spheroid size plotted per cluster. Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR). **(d)** Heatmap reflecting differential expression analysis of identified clusters in (b). Selected genes are listed beside the heatmap; Fold change > 1.5; adjusted *P*-value < 0.05; **P* < 0.05, ***P* < 0.01, ****P* < 0.001; 'small' cluster1: 313 differentially expressed genes; 'big' cluster: 130 differentially expressed genes. **(e)** Validation of pheno-seq by quantitative RNA-FISH for size-dependent differentiation marker TFF3 and stem cell markers CD44/MYC, and for size-independent DCS-cell marker DEFA5. Plotted values reflect the pixel fraction that exceeds the background threshold per

spheroid (Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values; *P*-values from unpaired Students t-test, ns: non-significant. Numbers of samples n indicated on x-axis under respective class). **(f)** Influence of γ-secretase inhibitor on spheroid growth. Plotted values represent average spheroid sizes after 10 days in culture in the presence of different concentrations of the γ-secretase inhibitor PF-03084014 (Three replicates; scatter dot plot center line: mean; whiskers: standard deviation; *P*-values from paired two-tailed Students t-test). **(g)** Example images (Z-projections) for RNA-FISH staining of DEFA5 corresponding to data shown in (e). DNA (Hoechst) counterstain visualization (Hoechst: cyan; RNA: yellow; scale bar 50 µm).

In the cluster enriched for big spheres, we identified several genes related to the γ-secretase machinery (Fig. 2d), a key component of the NOTCH pathway and target of new therapies aiming to disrupt cancer stem cell signaling[30]. Importantly, selective targeting of the γ-secretase by a small molecule inhibitor in concentration ranges that have been shown to force colonic stem cells into differentiation[31] showed a inhibitory effect on spheroid growth (Fig. 2f, Supplementary Fig. 11d). This finding suggests a similar signaling dependency of the normal and transformed intestinal stem cell niche and shows the potential of pheno-seq to identify relevant signaling components required for proliferative capacity.

Moreover, we determined an expression signature primarily driven by the expression of deep crypt secretory (DCS) cell markers DEFA5 and DEFA6 that is independent of the size-related clusters shown above (Supplementary Fig. 10a). DCS cells represent a post-mitotic secretory subpopulation at the bottom of intestinal crypts that serves as niche for $LGR5^+$ stem cells[31]. In line with pheno-seq results, we validated high-expressing $DEFA5^+$ cells as rare subpopulation with spheroid size-independent relative expression by RNA-FISH (Fig. 2e and g, Supplementary Fig. 10c). For a cellular subtype with limited proliferative potential within the putative CRC differentiation hierarchy, we would have expected a similar association of relative expression and size as observed for the $TFF3^+$ secretory signature above. Therefore, we suggest that DCS-like cells exhibit a heterogeneous growth phenotype (high- and low-cycling) that might relate to the delayed-contributing subpopulation in CRC previously described[8]. Thus, pheno-seq is able to directly assign heterogeneous proliferative phenotypes to expression signatures enriched for specific intestinal cell-type markers, results that are not directly obtainable from scRNA-seq data.

Single-cell deconvolution by combining image analysis and maximum likelihood inference

The pheno-seq method enables the direct association of spheroid 3D phenotypes and gene expression at a depth that cannot be reached by current single-cell methods alone. However, this accuracy comes at the cost of lower cellular resolution. The gene expression signatures identified from CRC spheroids inform about general phenotype-specific expression and trends in subtype composition but might derive from multiple cellular subtypes present within the same spheroids. While these results are highly valuable for understanding growth behavior in clonal cell culture systems (Fig. 2, Supplementary Fig. 10 and 11), obtaining 'real' single-cell information from pheno-seq data would be of high relevance to distinguish between genes that are generally associated with spheroid phenotypes and those who are robustly expressed in a single-cell subpopulation. Therefore, we aimed to computationally infer single-cell regulatory states by deconvolution of gene expression data using both image analysis and a maximum likelihood inference approach.

First, we generated a 3D high-resolution imaging reference dataset by light-sheet microscopy from spheroids of different sizes, which we used to determine the relationship of spheroid size and nuclei counts to estimate cell numbers from CRC spheroid pheno-seq imaging data (Supplementary Fig. 12a). As the original pheno-seq data exhibited a low correlation between library complexity and estimated cell numbers, we downsampled the data to achieve a constant number of mRNA counts per estimated single cell content (Supplementary Fig. 12b). As expected, this transformation introduces a positive overall shift of correlations between gene expression and cell numbers (Supplementary Fig. 12c), which can be mainly explained by housekeeping genes with a constant number of mRNA molecules per cell (Supplementary Fig. 13a). However, heterogeneously expressed genes such as the differentiation markers TFF3 and DEFA5 do not exhibit any correlation with cell numbers (Supplementary Fig. 13b and c), validating our normalization approach.

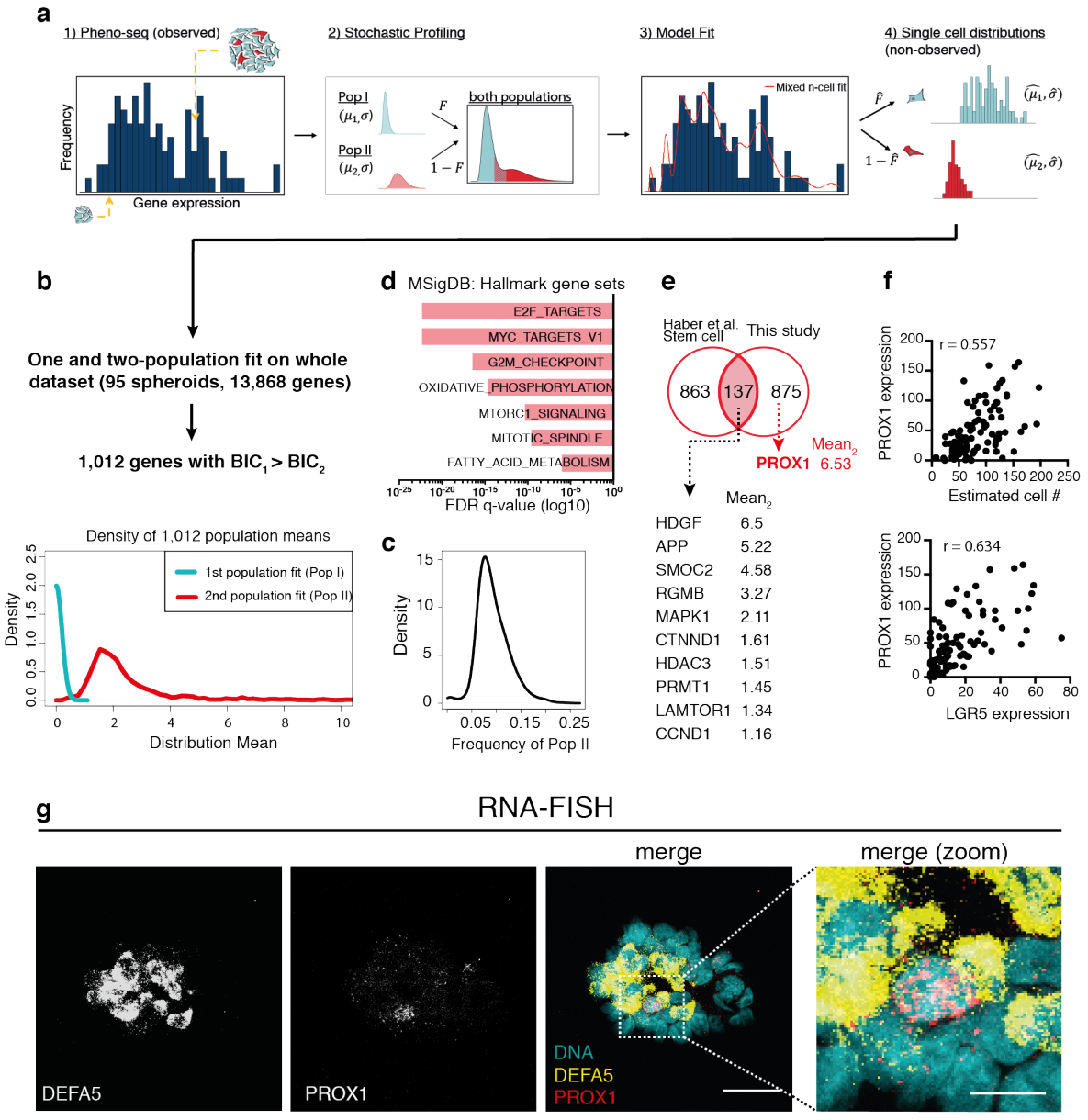


Figure 3 | Single-cell deconvolution of CRC spheroid pheno-seq data by maximum likelihood inference identifies PROX1 as potential CRC stem cell marker.

(a) Concept of adapted maximum likelihood approach[32] based on estimated cell numbers and transformed pheno-seq data: 1) Acquired and transformed pheno-seq data based on estimated cell numbers build a distribution of measurements for inference by the model. Coloring of cells in spheroids: red = stem-like; cyan = differentiated. 2) Assumptions on single cell distributions: Model of heterogeneous gene regulation in which single cells are supposed to exhibit gene expression at low (Pop I) or high (Pop II) levels with a common coefficient of variation. The four parameters of the model are the log-mean expression for each subpopulation (μ_1 and μ_2), the proportion of cells in the high subpopulation (F), and the common log-SD of expression (σ). 3) Based on the model in step 2, a likelihood function is derived that takes different numbers of cells per spheroid into account. The likelihood function is then maximized by searching through the four parameters of the model to identify those that are most likely given the experimental observations. 4) These four parameters define the

inferred single cell distributions of the low and high-level populations. **(b)** 1,012 genes show an improved two-population fit compared to a one population fit (BIC: Bayesian information criterion). Densities of the means of the first (Pop I: low regulatory state) and second population (Pop II: high regulatory state) for all identified 1,012 genes. **(c)** Frequency distribution of cells with high regulatory state (Pop II) of identified 1,012 genes. **(d)** Gene set enrichment analysis for two-population genes based on Hallmark gene sets derived from the MSigDB. Bar plot showing top enriched gene sets ranked by FDR q-values. **(e)** Venn-diagram showing overlap between identified two-population genes and murine small intestinal stem cell signature from scRNA-seq study[28]. Selected genes are listed below ordered by mean for high-state population Pop II ($Mean_2$). **(f)** Scatter plots for PROX1 expression plotted against estimated cell numbers (upper) and against expression of major intestinal stem cell marker LGR5 (lower) as well as associated Pearson's correlation coefficients (r). **(g)** RNA-FISH co-staining of CRC spheroids for PROX1 and DEFA5 and Hoechst counterstaining for visualization of DNA. Merged images: DNA: cyan; DEFA5 yellow; PROX1: red. Images represent Z-projections (scale bar 30 µm and 10 µm for magnified merged image).

To identify genes whose expression was likely to be informative for heterogeneous single-cell regulatory states, we used a maximum likelihood inference approach initially developed to deconvolve cell-to-cell heterogeneities from random 10-cell samples[32] (Fig. 3a). The adapted algorithm uses the estimated cell numbers per spheroid to fit two log-normal distributions (LN-LN model) to given 'mixed-n' datasets in order to identify genes with bimodal expression pattern at the single-cell level (Stochastic Profiling, see Methods). Importantly, this approach unbiasedly pinpoints genes that show a heterogeneous and robust expression within spheroids at the single-cell level, instead of comparing gene expression between spheroids.

Whilst the deconvolution technique assumes that cellular subtypes are identically distributed across samples, pheno-seq is based on clonal spheroids whose cell number, subtype composition and expression profile is dependent on the state of the founding cell. Based on the cancer stem cell model and the CRC differentiation hierarchy confirmed above, we assume that continuously growing spheroids ('big' phenotype) harbor all cellular subtypes present in this system, including stem-like cells, whereas small spheroids with limited proliferative capacity and low cell numbers are more homogeneous and contain only differentiated subtypes. Thus, inferred regulatory states should be enriched for genes specific for the stem-like compartment, as these represent the major source of heterogeneity across all spheroids at the single-cell level.

Deconvolution of the entire CRC pheno-seq dataset revealed 1,012 genes that show an improved two-population fit compared to a one-population fit, assessed by the Bayesian information criterion (BIC) to calculate the quality of the fit relative to the number of inferred parameters (Fig. 3b). Most of the fits resulted in a highly-expressing cellular fraction of 5 – 15% (Fig. 3c) thereby matching the proportion of cells with spheroid forming capacity in this model[8]. Interestingly, the positive shift of correlations between gene expression and cell numbers (before and after downsampling) is much more pronounced in the set of two-population genes compared to the set of non-two-population genes (Supplementary Fig. 13d), suggesting that many of the inferred two-population genes are involved in proliferative potential. Indeed, gene set enrichment analysis reveals a high proportion of MYC targets as well as genes involved in the regulation of cell growth and proliferation (Fig. 3d). In addition, high enrichment of genes involved in oxidative phosphorylation indicates for heterogeneous mitochondrial activity at the single-cell level, a phenomenon recently described for intestinal stem cells and niche-forming Paneth cells in the small intestine[33]. Strikingly, a high number of identified genes are overlapping with a recently identified stem cell signature of the small intestine revealed by massively parallel scRNA-seq[28], including SMOC2, APP, PRMT1, RGMB, MAPK1 and CTNND1, respectively (Fig 3e).

Here, we identified the transcriptional regulator PROX1 as gene with a high population (Pop II) mean (Fig. 3e) that is strongly correlated with cell numbers and with expression of the major stem cell marker LGR5 (Fig. 3f). In the normal intestinal epithelium, PROX1 is expressed in the enteroendocrine lineage[34]. However, two studies based on mouse tumor models suggest a role for PROX1 in cancer stem cell maintenance and metastatic outgrowth[35,36]. In line with these observations, we validated PROX1^{+} cells by RNA-FISH as a rare subpopulation in a patient-derived human tumor model (Fig. 3g). Furthermore, PROX1^{+} cells are framed by DEFA5^{+}-positive DCS-like cells, suggesting a similar niche dependency for normal stem cells and CRC stem-like cells at distant sites of neoplasia. Taken together, deconvolution of pheno-seq data provides information about gene expression patterns at the single cell level.

Discussion

As applications of 3D-cultures are emerging even into clinical settings[2], there is increasing need to directly link oncogenic visual phenotypes to underlying system-wide gene expression, which cannot be achieved by imaging or scRNA-seq alone. In this study, we present pheno-seq as new and complementary approach that directly combines high-throughput imaging and next generation sequencing to functionally explain heterogeneous visual phenotypes in 3D-culture systems. At the same time, pheno-seq bridges the gap between single-cell and bulk expression profiling. In principle, pheno-seq can be applied to any 3D-culture system given that the phenotypic identity is maintained upon spheroid isolation. We expect that this combination of functional single cell growth assay with combined image and gene expression profiling will be widely applied in cancer biology, ranging from primary to circulating tumor cells (CTCs[37]).

Importantly, we show that pheno-seq is able to link cell type-specific genes to heterogeneous growth phenotypes even in highly complex cell culture systems. In addition, we show that deconvolution by maximum likelihood inference provides an additional layer of information by revealing single-cell regulatory states that are likely to be associated with a distinct stem-like population, thereby further supporting a differentiation-like hierarchy in CRC. Based on our results, future studies should shed light on additional functional characteristics and dependencies of the stem-like compartment, the implication of the heterogeneous growth phenotype of DCS-like cells, potential cancer cell plasticity and the impact of subtype-specific metabolic preferences. Complementary single-cell-derived information might be added to further deconvolve pheno-seq expression profiles to fully understand cell type composition and differentiation trajectories.

We envision pheno-seq to become even more powerful with increasing resolution and content of imaging, employing enhanced 3D-image acquisition, integrated staining by IF or live-dyes, and time-lapse microscopy, respectively. Pheno-seq can also be easily extended to other low-input, next-generation sequencing modalities such as chromatin accessibility sequencing. Furthermore, pheno-seq might be applied to pooled-screening approaches[38] or to resolve transcriptional changes that are associated with morphological transitions in non-synchronized developmental processes. Thus, pheno-seq will widely impact the way how we study functional heterogeneity in a variety of biological and clinical applications.

Acknowledgments

We thank David Ibberson (CellNetworks Deep Sequencing Core Facility, Heidelberg University) for NGS services, Daniel Liber and Marizela Kulisic (TakaraBio) for technical support for the iCELL8 system, Lorenz Maier (Theoretical Bioinformatics, DKFZ) for help with KNIME, Katharina Jechow (Theoretical Bioinformatics, DKFZ) for technical laboratory support, Claudia Ernst and Niels Grabe (Hamamatsu TIGA Center, Heidelberg University) for help with histological preparation, Naveed Ishaque (Theoretical Bioinformatics, DKFZ) for assistance in RNA-seq data analysis and Dominik Niopek, Luca Tosti, Julia Neugebauer, Teresa Krieger and Lorenz Chua (Theoretical Bioinformatics, DKFZ) for critically revising the manuscript. Primary human colon cancer samples were obtained from Heidelberg University Hospital in accordance with the declaration of Helsinki. Informed consent on tissue collection was received from each patient, as approved by the University Ethics Review Board. ST is recipient of the stipend for the PhD program of the Helmholtz International Graduate School for Cancer Research (DKFZ, Heidelberg). This study was supported by the Helmholtz International Graduate School for Cancer Research, the iMed Program (Helmholtz Association), the BMBF-funded Heidelberg Center for Human Bioinformatics (HD-HuB) within the German Network for Bioinformatics Infrastructure (de.NBI) (#031A537A, #031A537C), the DFG (SFB873), the EU framework programme Horizon2020 (TRANSCAN-2 ERA-NET), the German Cancer Aid (Colon-Resist-Net), NCT3.0_2015.4 TransOnco. and NCT3.0_2015.54 DysregPT, the German Research Foundation (DFG) within the Collaborative Research Centre 1243, Subproject A17, the BMBF (grant # 01ZX1711A) and the Helmholtz Association (Incubator grant sparse2big, grant # ZT-I-0007). DKFZ-HIPO provided technical support and funding through Grant No. HIPO-H012.

Author contributions

SMT and CC conceived the study, SMT, CC, HG and RE designed experiments; SMT performed 3D cell culture experiments, IF/RNA-FISH stainings and iCELL8 sample and library preparation; SMT and FP performed confocal microscopy; BE performed light-sheet microscopy; FP developed the HT-pheno-seq imaging protocol, the image processing pipeline and PhenoSelect; FP and MW performed image analysis; JPM and SMT performed Fluidigm C1 experiments and JPM generated sequencing libraries; JP, LA, ZG, SMT and SS analyzed RNA-seq data; LA, CF and

FJT developed and applied the adapted maximum likelihood inference deconvolution approach for pheno-seq data; CB and HG generated and characterized the colon spheroid cultures and contributed experimental and clinical expertise; KR, MS, MG and IG provided advice on single-cell sequencing experiments and analysis. MG, IG contributed NGS expertise and sample processing. SMT and CC wrote the manuscript. All authors revised and approved the manuscript.

Competing financial interests

The authors declare no competing financial interests

References

1. Pampaloni, F., Reynaud, E. G. & Stelzer, E. H. K. The third dimension bridges the gap between cell culture and live tissue. *Nat. Rev. Mol. Cell Biol.* **8,** 839–845 (2007).
2. Sachs, N. *et al.* A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. *Cell* **172,** 373–386.e10 (2018).
3. Neal, J. T. & Kuo, C. J. Organoids as Models for Neoplastic Transformation. (2016). doi:10.1146/annurev-pathol-012615-044249
4. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501,** 355–64 (2013).
5. Borten, M. A., Bajikar, S. S., Sasaki, N., Clevers, H. & Janes, K. A. Automated brightfield morphometry of 3D organoid populations by OrganoSeg. *Sci. Rep.* **8,** 5319 (2018).
6. Roerink, S. F. *et al.* Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* 1 (2018). doi:10.1038/s41586-018-0024-3
7. Bithi, S. S. & Vanapalli, S. A. Microfluidic cell isolation technology for drug testing of single tumor cells and their clusters. *Sci. Rep.* 1–12 (2017). doi:10.1038/srep41707
8. Dieter, S. M. *et al.* Distinct Types of Tumor-Initiating Cells Form Human Colon Cancer Tumors and Metastases. *Cell Stem Cell* **9,** 357–365 (2011).
9. Ye, X. & Weinberg, R. A. Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. *Trends Cell Biol.* **25,** 675–686 (2015).
10. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer* **17,** 557–569 (2017).
11. Bray, M. *et al.* Cell Painting , a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11,** 1757–1774 (2016).
12. Nichterwitz, S. *et al.* Laser capture microscopy coupled with Smart-seq2 (LCM-seq) for robust and efficient transcriptomic profiling of mouse and human cells. *Nat. Commun.* **7,** 1–11 (2016).
13. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348,** aaa6090 (2015).
14. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9,** 75 (2017).
15. Debnath, J. & Brugge, J. S. Modelling glandular epithelial cancers in three-dimensional

cultures. *Nat. Rev. Cancer* **5,** 675–688 (2005).

16. Santner, S. J. *et al.* Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells. *Breast Cancer Res. Treat.* **65,** 101–110 (2001).
17. Fan, J.-B. J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 241–244 (2016). doi:10.1038/nmeth.3734
18. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).
19. Kharchenko, P. V, Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–2 (2014).
20. Brink, S. C. van den *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14,** (2017).
21. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. Emt: 2016. *Cell* **166,** 21–45 (2016).
22. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.* **8,** (2017).
23. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Aspects Med.* **59,** 114–122 (2018).
24. Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* **8,** 228 (2017).
25. Attar, M. *et al.* A practical solution for preserving single cells for RNA sequencing Fixation protocol. *Sci. Rep.* **8,** 2151 (2018).
26. Giessler, K. M. *et al.* Genetic subclone architecture of tumor clone-initiating cells in colorectal cancer. *J. Exp. Med.* **214,** 2073–2088 (2017).
27. Dieter, S. M., Glimm, H. & Ball, C. R. Colorectal cancer-initiating cells caught in the act. *EMBO Mol. Med.* **9,** 856–858 (2017).
28. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* (2017). doi:10.1038/nature24489
29. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29,** 1120–1127 (2011).
30. Takebe, N. *et al.* Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. *Nat. Rev. Clin. Oncol.* **12,** 445–64 (2015).
31. Sasaki, N. *et al.* Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. *Proc. Natl. Acad. Sci. U. S. A.* **4,** 201607327 (2016).
32. Bajikar, S. S., Fuchs, C., Roller, A., Theis, F. J. & Janes, K. a. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E626-35 (2014).
33. Rodríguez-Colman, M. J. *et al.* Interplay between metabolic identities in the intestinal crypt supports stem cell function. *Nature* 1–13 (2017). doi:10.1038/nature21673
34. Yan, K. S. *et al.* Intestinal Enteroendocrine Lineage Cells Possess Homeostatic and Injury-Inducible Stem Cell Activity. *Cell Stem Cell* **21,** 78–90.e6 (2018).
35. Wiener, Z. *et al.* Prox1 promotes expansion of the colorectal cancer stem cell population to fuel tumor growth and ischemia resistance. *Cell Rep.* **8,** 1943–1956 (2014).

36. Ragusa, S. *et al.* PROX1 promotes metabolic adaptation and fuels outgrowth of Wnthigh metastatic colon cancer cells. *Cell Rep.* **8,** 1957–1973 (2014).

37. Khoo, B. L. *et al.* Expansion of patient-derived circulating tumor cells from liquid biopsies using a CTC microfluidic culture device. *Nat. Protoc.* **13,** 34–58 (2018).

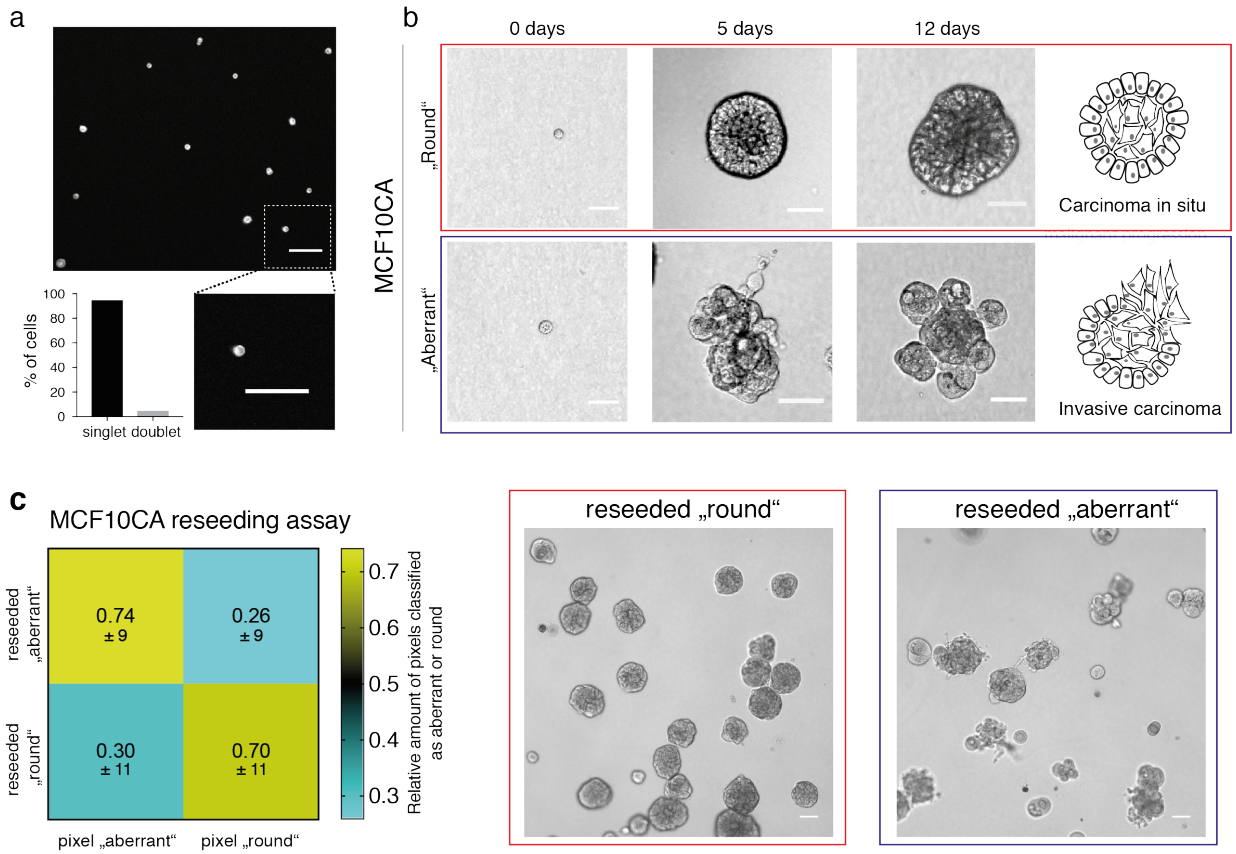
38. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome read-out. *Nat. Methods* **12,** (2016).

Supplementary tables and figures

3D-culture model	MCF10CA	MCF10CA	MCF10CA	MCF10CA	MCF10CA	CRC spheroid
Method	scRNA-seq	Pseudo-pheno-seq	Manual pheno-seq	HT-pheno-seq (control)	HT-pheno-seq (DSP)	HT-pheno-seq
Library structure	Full-length C1	Full-length C1	Full-length Tube-based	3`-end iCELL8	3`-end iCELL8	3`-end iCELL8
Number of samples after library QC	166	8	8	64	210	95
Mean total read count per sample	3,820,057	3,685,536	9,965,986	485,975	803,669	1,304,480
Mean detected genes (> 0) per sample (all reads)	8,844	15,783	12,360	8,458	8,221	9,891
Mean detected genes (> 0) per sample (down-sampled to 100k reads)	5,554	13,374	8,411	7,051	6,377	7,412

Supplementary table 1 | Dataset overview and QC metrics

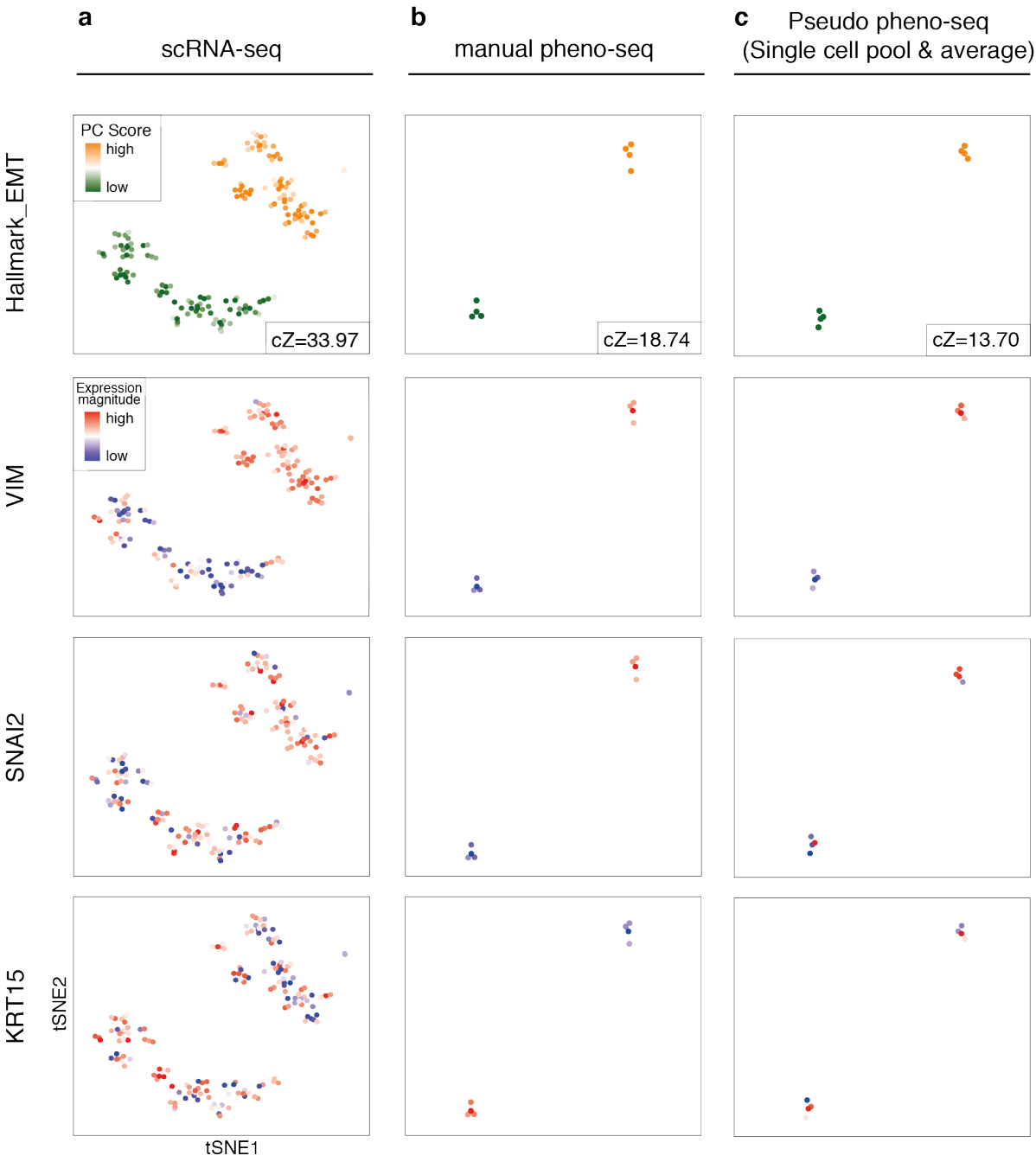
(HT-pheno-seq: high-throughput pheno-seq; control: bottom control with default chip and imaging settings; DSP: Fixation with dithio-bis(succinimidyl propionate crosslinker; C1: Fluidigm C1)



Supplementary figure 1 | Heterogeneous 3D breast cancer model MCF10CA

(a) Assessing single-cell seeding efficiency by image analysis. Upper: Example image of CellTacker Red stained and seeded cells (scale bar: 100 µm). Lower right: Magnified image that corresponds to dashed box in upper image. Lower left: Quantified cell singlets and doublets after seeding (289 objects in total). **(b)** Brightfield microscopy images of heterogeneous MCF10CA spheroids after 0, 5 and 12 days of culture in Matrigel, thereby reflecting histological characteristics of steps during malignant progression of breast cancer (Brightfield, scale bar 50 µm). Red box: 'round' phenotype; Blue box: 'aberrant' phenotype.

(c) Independent reseeding of isolated 'round' and 'aberrant' spheroid phenotypes and quantification after regrowth by 'ilastic' machine learning based on pixel classification. Left: Spheroid classification confusion matrix. Heatmap reflecting classified pixels as aberrant or round after reseeding (four replicates, indicated are relative pixel numbers and standard error of the mean below). Right: Example images of reseeded MCF10CA 'round' and 'aberrant' spheroids 5 days after reseeding (scale bar: 50 µm)



Supplementary figure 2 | Comparison of scRNA-seq and manual pheno-seq by tSNE visualizations

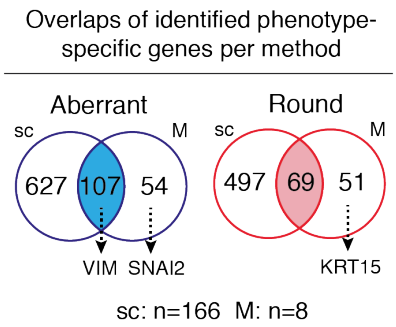
(a, b, c) PAGODA 2D tSNE embedding of MCF10CA scRNA-seq (a), manual pheno-seq (b) and Pseudo pheno-seq (c, based on averaged single cell data) datasets colored by PC scores for Hallmark_EMT gene sets (incl. associated cZ scores as measure of gene set overdispersion) and by expression magnitude of phenotype markers VIM, SNAI2 and KRT15.

a

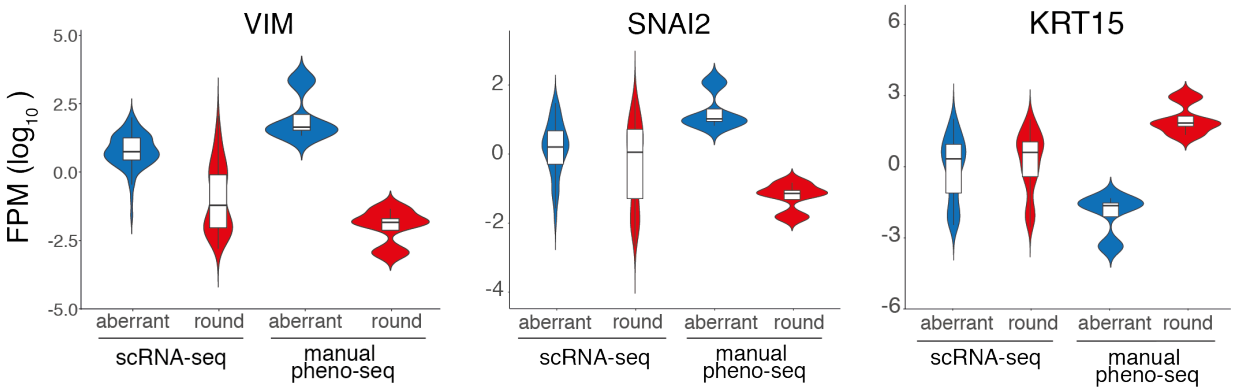
	scRNA-seq	manual pheno-seq
Aberrant HALLMARK_EMT	1.71×10^{-28}	4.18×10^{-27}
Round GO_TISSUE_DEVELOPMENT	1.71×10^{-43}	6.01×10^{-13}

FDR q-values

b



c

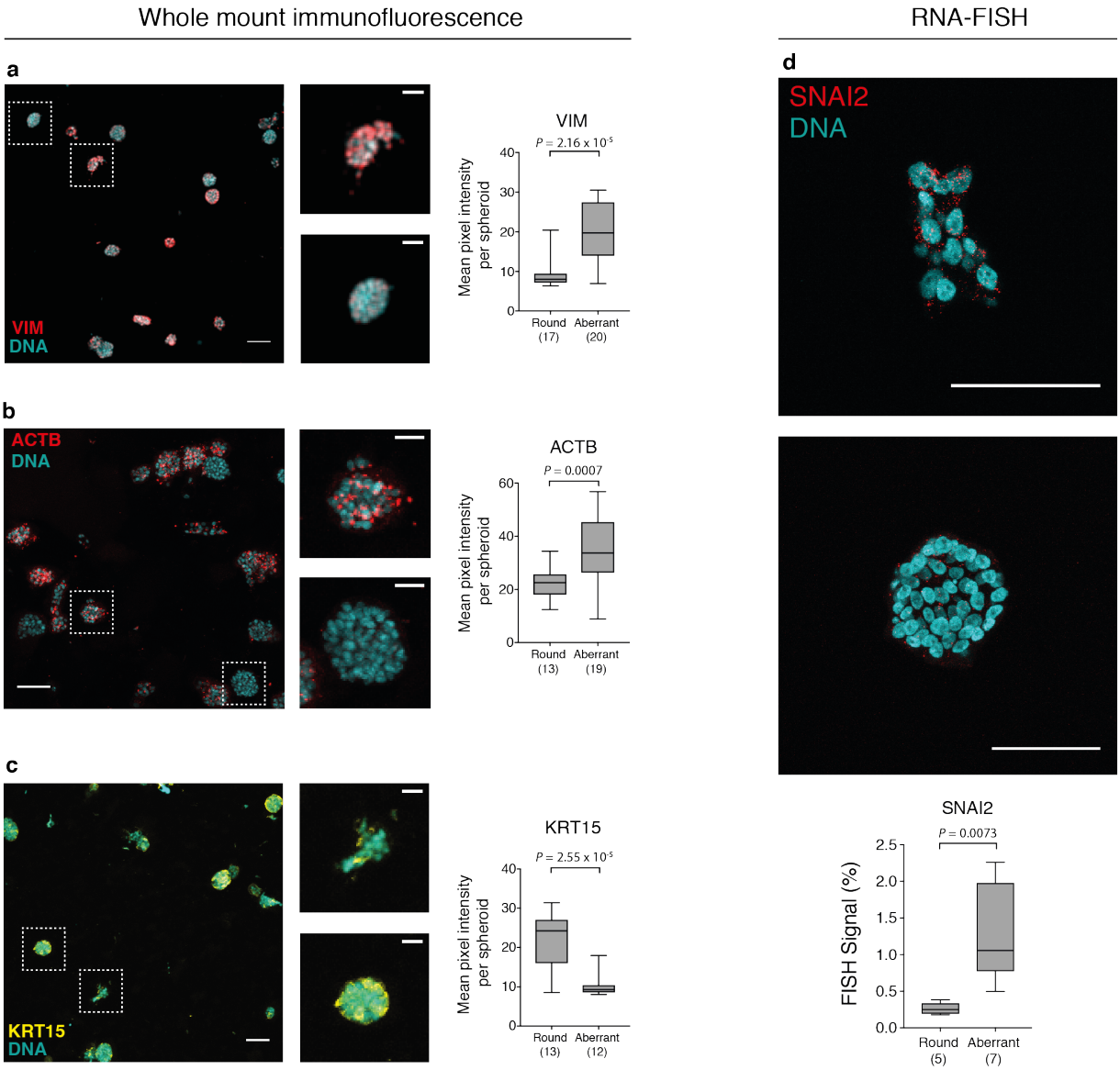


Supplementary figure 3 | pheno-seq identifies highly relevant gene expression that is missed by scRNA-seq

(a) Gene set enrichment analysis based on differentially expressed genes identified by scRNA-seq and manual pheno-seq. Listed are FDR q-values for enrichments of biologically relevant HALLMARK_EMT and GO_TISSUE_DEVELOPMENT gene sets (derived from MSigDB).

(b) Venn-Diagrams reflecting overlaps of identified phenotype-specific genes between scRNA-seq and manual pheno-seq based on differential expression analysis (fold change > 1.3; adjusted p-value < 0.1).

(c) Violin plots showing expression of individual genes (VIM, SNAI2, KRT15) per identified phenotype-specific clusters for scRNA-seq and manual pheno-seq. Expression magnitude is plotted as Fragments per Million (FPM, $\log_{10}$). Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR.

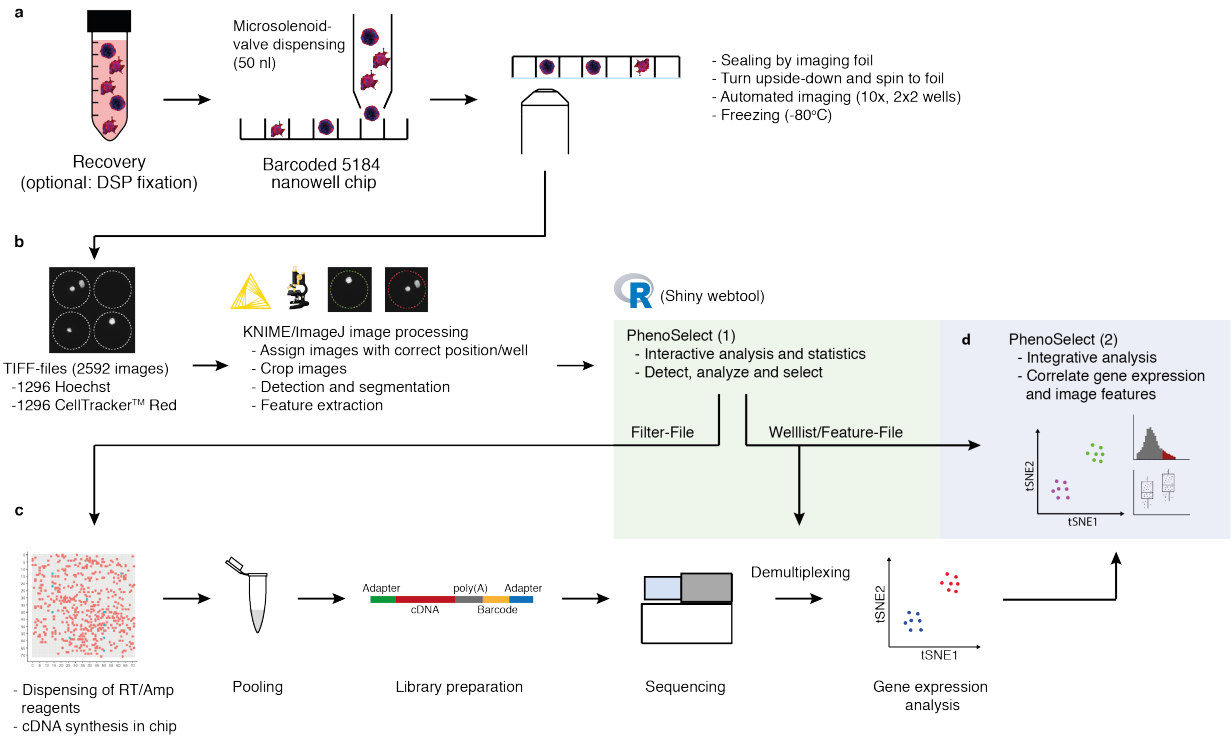


Supplementary figure 4 | Validation of RNA-seq data by quantitative fluorescence microscopy.

(a, b, c) Immunofluorescence staining with primary antibodies targeting VIM (a), ACTB (b) and KRT15 (c). Images represent Z-projections of whole mount spheroid immunofluorescence. Plotted values reflect the mean pixel intensity per classified spheroid of the respective class. Dashed boxes in overview images (scale bar 100 µm) correspond to magnified images beside (scale bar 30 µm).

(d) RNA-FISH with probe targeting SNAI2 mRNA (scale bar 100 µm). Images represent Z-projections and plotted values reflect the pixel percentage that exceeds the threshold per spheroid of the respective class after background correction.

(a-e) All samples are counterstained with Hoechst dye to visualize nuclei (Hoechst: cyan; Labelled antibodies for round specific markers: yellow; labelled antibodies and RNA-FISH probe for aberrant specific markers: red). (Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values; Indicated *P*-values from unpaired two-tailed Students t-test; Numbers of samples indicated on x-axis under respective class).



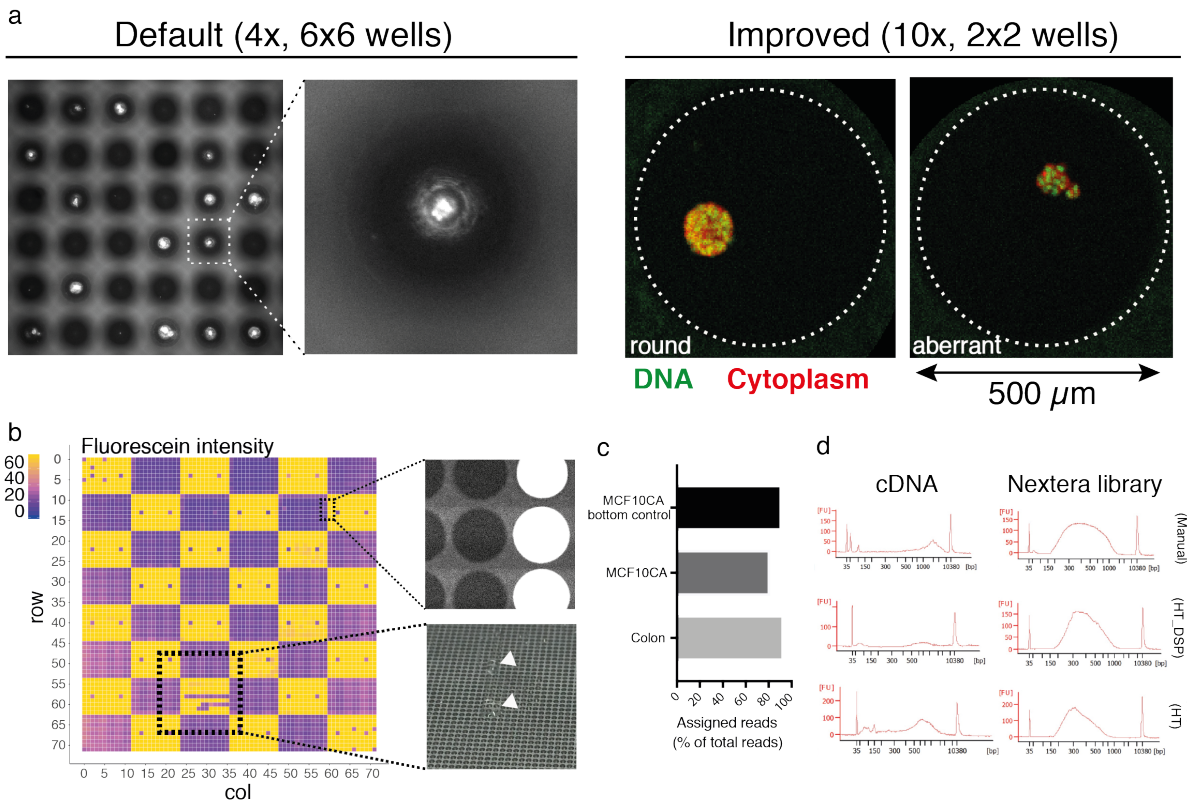
Supplementary figure 5 | Detailed HT-pheno-seq workflow.

(a) After staining and recovery (optional: DSP fixation), spheroids are distributed into a nanowell chip by a Microsolenoid-valve dispenser (50 nl per well). To improve imaging quality, spheroids are centrifuged upside-down to the foil and automatically imaged by an inverted confocal microscope. The chip is then frozen at -80ºC.

(b) Images are processed using a custom-made image processing pipeline in KNIME/ImageJ. A Shiny-based web-app (PhenoSelect) enables interactive analysis and selection based on quantified image features.

(c) A filter-file generated by PhenoSelect is used to dispense RT/Amp reagents only in selected wells. cDNA generation and amplification are performed in the chip. After pooling of barcoded cDNA, 3´-library generation and next generation sequencing, resulting raw data can be de-multiplexed using internal barcodes listed in the welllist/feature-file generated with PhenoSelect.

(d) Combined image analysis enables combined analysis of gene expression and image features.



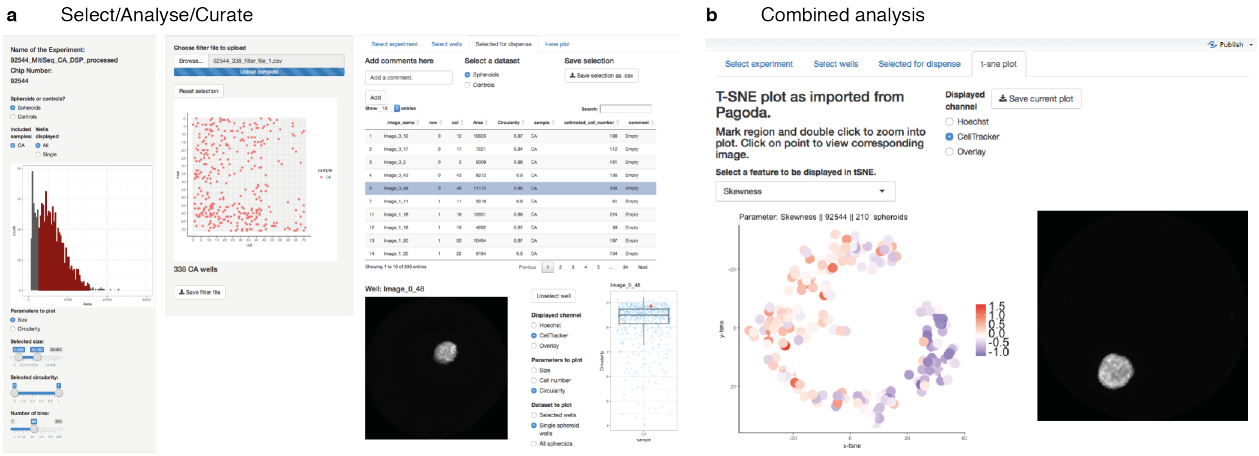
Supplementary figure 6 | Technical adaptions and controls for high-throughput pheno-seq.

(a) Comparison of images acquired by the default microscope with 4x objective, capturing 6x6 wells per image (spheroid nuclei are stained with Hoechst dye), and higher resolution microscopy (Confocal Leica SP8) with 10x objective, capturing 2x2 wells per image (spheroids are stained with Hoechst dye and CellTracker Red CMTPX).

(b) Leakage analysis by patterned Fluorescein dispensing. Average fluorescence intensity is plotted onto 72x72 well grid that corresponds to nanowell chip architecture (left). For better visualization, all average intensity values exceeding 77 were set to maximum in the color code scheme. Top right: Example image showing the border between wells that have been filled with PBS or PBS with Fluorescein. Lower right: Macroscopic image of nanowell surface with droplets, showing rare dispensing errors that are also reflected by absence of Fluorescein signal at the respective position.

(c) High percentage of reads that only map to selected well barcodes excludes severe leakage of barcoded Poly-T primers upon centrifugation of spheroids to the foil.

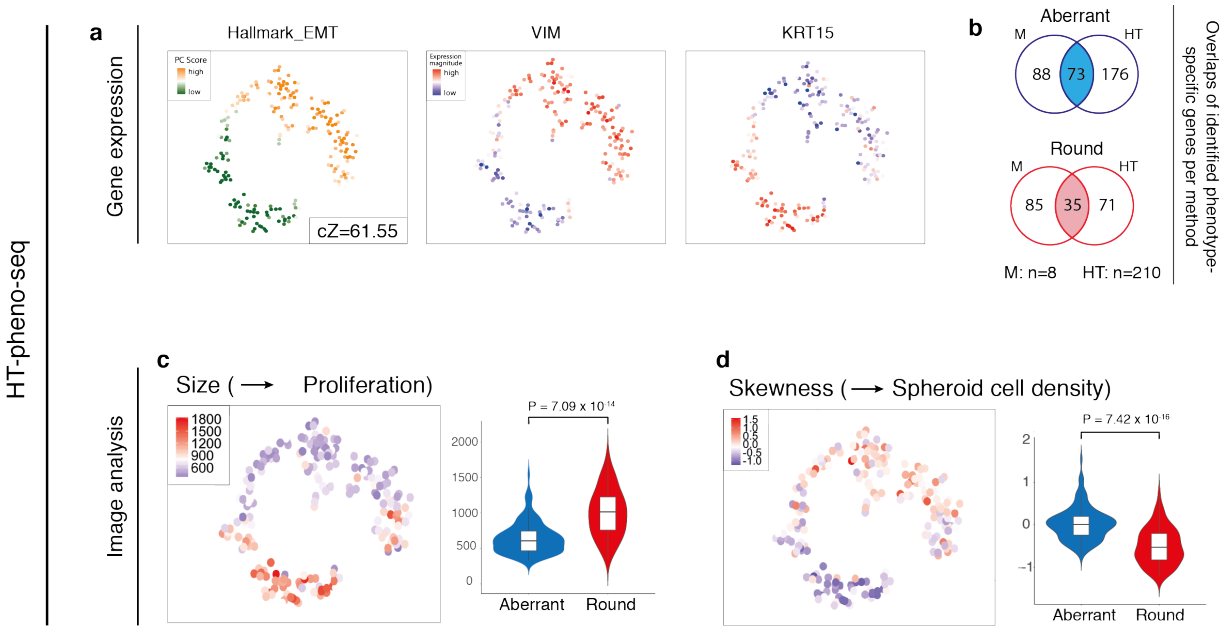
(d) cDNA and Nextera XT library Bioanalyzer traces show compatibility of HT-pheno-seq with nanowell based system.



Supplementary figure 7 | PhenoSelect software for interactive analysis and selection of spheroids for high-throughput pheno-seq.

(a) Primary selection of single spheroids based on individual thresholds as well as analysis and curation of selected spheroids. Filter- and welllist/feature-file are generated at this point and can be reloaded or adapted at any time.

(b) PhenoSelect enables import of externally generated tSNE maps (PAGODA) for combined analysis of image features and gene expression.

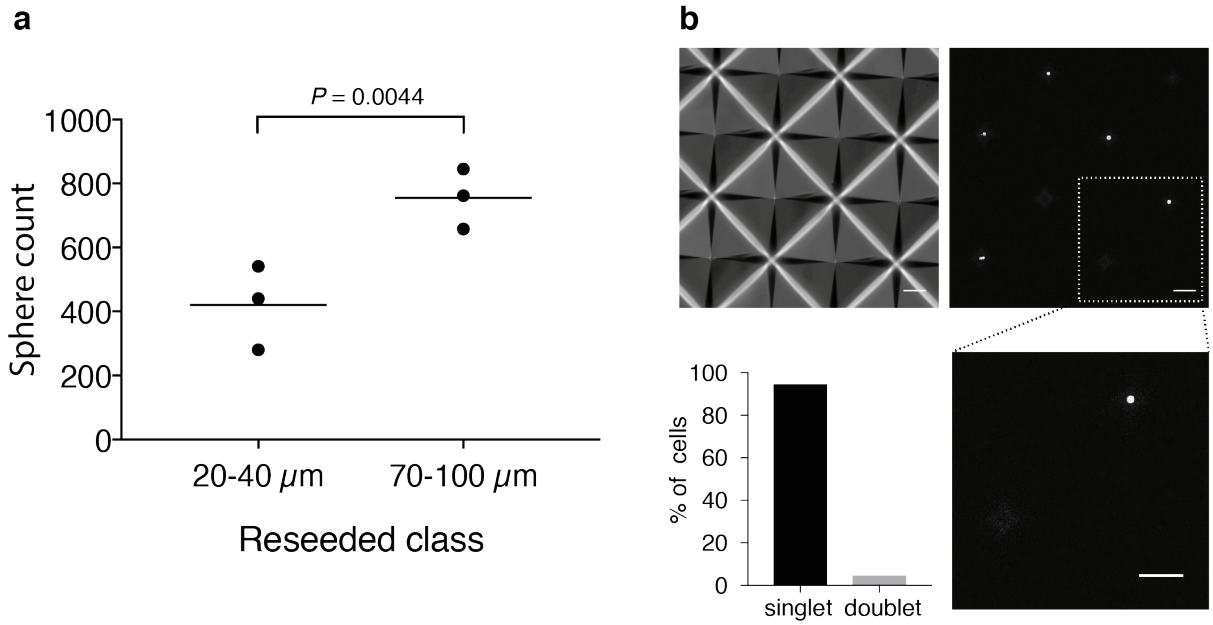


Supplementary figure 8 | High-throughput pheno-seq enables combined quantitative analysis of gene expression and image features of clonal spheroids

(a) PAGODA 2D tSNE embedding of MCF10CA HT-pheno-seq dataset colored by PC scores for Hallmark_EMT gene sets (incl. associated cZ scores as measure of gene set overdispersion) and by expression magnitude of major phenotype markers VIM and KRT15.

(b) Venn-Diagrams reflecting overlaps of identified phenotype-specific genes between manual pheno-seq and HT-pheno-seq based on differential expression analysis (fold change > 1.3; adjusted p-value < 0.1).

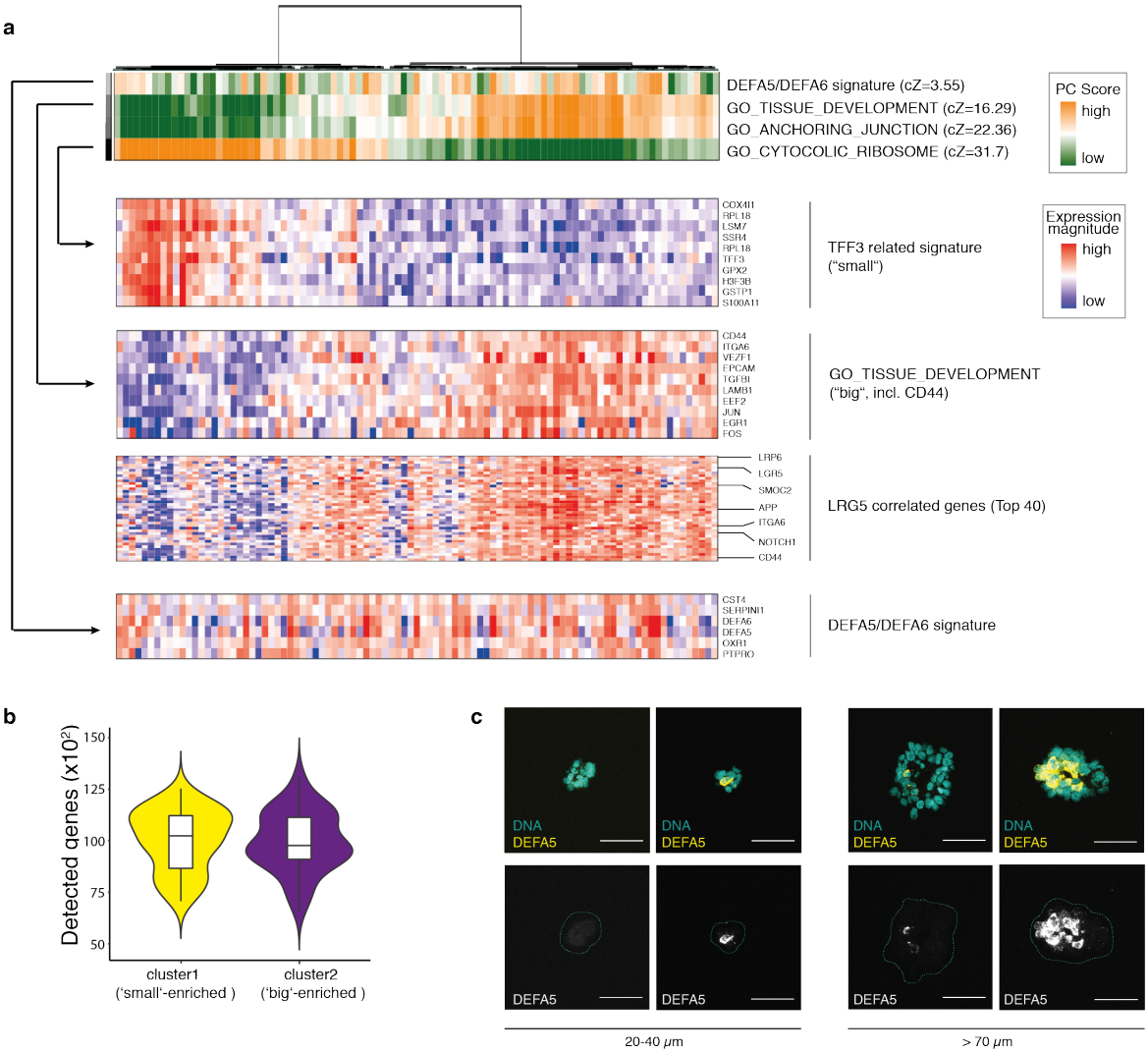
(c and d) Same t-SNE map as shown in (a) but colored based on combined image analysis for image features 'size' (c) and 'skewness' (d). Right: Violin plots show image feature quantification per cluster (k-means clustering: k=2; violin center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR; Indicated *P*-values from unpaired two-tailed Students t-test). Image feature associations can be interpreted according to the biological background (e.g. proliferation and cell density)



Supplementary figure 9 | Experimental basis for HT-pheno-seq of 3D model for colorectal cancer

(a) Functional reseeding assay with cells isolated from different spheroid size classes reveals association of spheroid size and the long-term proliferative capacity of associated cells (20-40 µm and >70-100 µm). Plotted are spheroid counts 10 days after reseeding (three replicates, center-line: mean; indicated *P*-value of paired two-tailed Students t-test).

(b) Assessing single-cell seeding efficiency by image analysis. Left: Example image of CellTacker Red stained cells seeded in microwells (scale bar: 100 µm). Lower right: Magnified image that corresponds to dashed box in upper image. Lower left: Quantified cell singlets and doublets after seeding (three wells, four images per well, 70 objects in total).

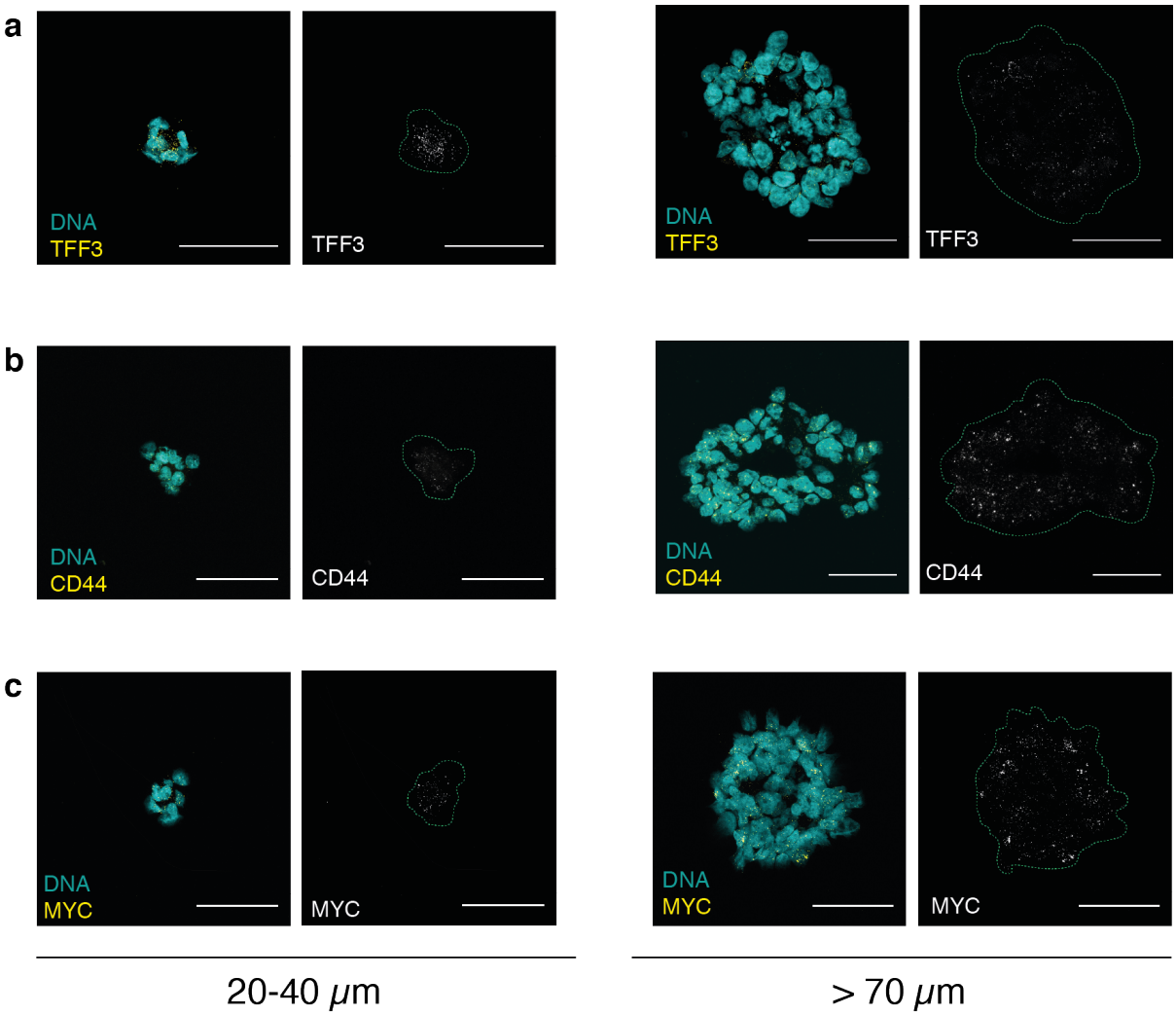


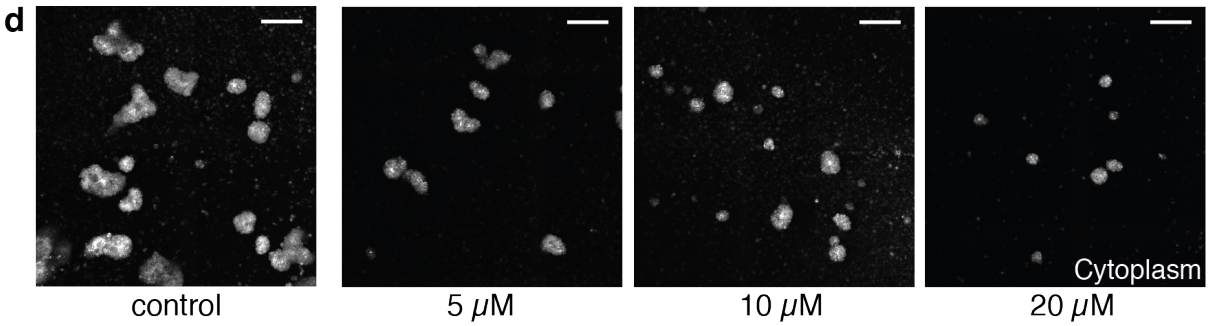
Supplementary figure 10 | LGR5 correlated genes overlap with expression signature of big spheroids and indication for heterogeneous growth phenotype of DCS-like subpopulation

(a) RNA-seq analysis of CRC spheroid HT-pheno-seq data. Dendogram shows overall clustering (left: 'small', right: 'big') and the rows below represent top four significant aspects of heterogeneity detected by PAGODA based on Hallmark and GO gene sets derived from the MSigDB as well as on *de-novo* identified gene sets. High aspect scores (PC Scores) correspond to high expression of associated gene sets. Associated top gene sets are listed next to rows (including cZ scores as measure of gene set overdispersion). Expression patterns below reflect top loading genes for selected gene sets that are associated with respective aspects. One exception is the expression pattern of genes exhibiting the highest correlation to the major intestinal stem cell marker LGR5 (Pearson's correlation, top 40). Bottom: Signature that is dominated by the DCS cell markers DEFA5 and DEFA6 is independent of the major (size-associated) clustering.

(b) Detected genes plotted per cluster shown in Fig. 2b. Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR.

(c) Example images (Z-projections) for RNA-FISH staining for DEFA5 of big (>70 μm) and small (20-40 μm) spheroids with (top) and without (lower) Hoechst counterstain visualization (Hoechst: cyan; RNA: yellow). Dashed line in images without Hoechst visualization represents spheroid border (scale bar 50 μm).

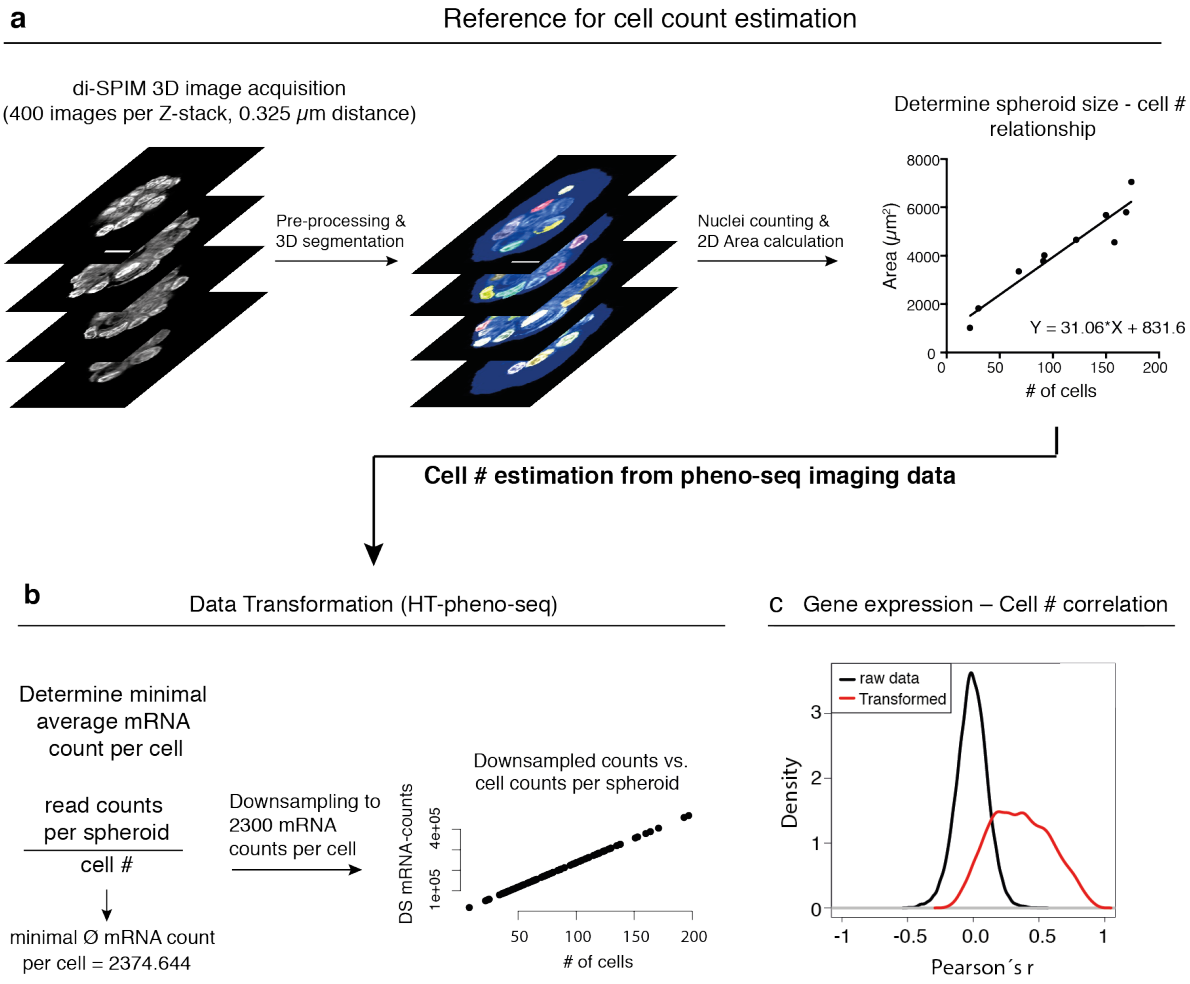




Supplementary figure 11 | Validation of CRC pheno-seq data by RNA-FISH and γ-secretase inhibition

(a, b, c) RNA-FISH example images of different spheroid size classes for differentiation marker TFF3 (a) and cancer stem cell markers CD44 (b) and MYC (c) (plotted data also shown in Fig. 2e). Z-projections for RNA-FISH staining of big (>70 µm) and small (20-40 µm) spheroids with (left) and without (right) Hoechst counterstain visualization (Hoechst: cyan; RNA: yellow). Dashed line in images without Hoechst visualization represents spheroid border (scale bar 50 µm).

(d) Example images of CellTracker Red stained spheroids after 10 days in culture under different γ-secretase inhibitor (PF-03084014) treatment conditions (scale bar 200 µm).

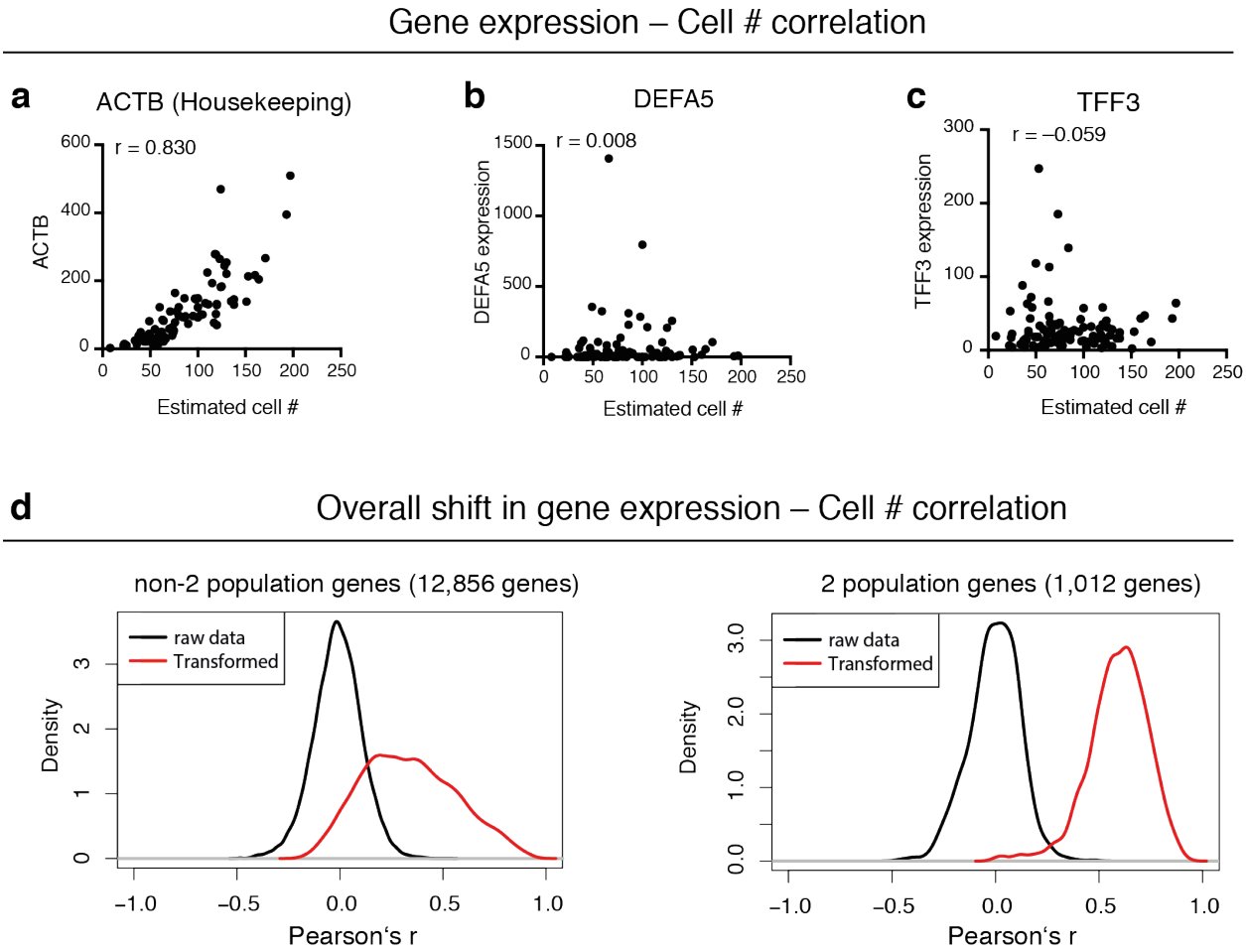


Supplementary figure 12 | Estimation of cell numbers from pheno-seq data using a high-resolution reference dataset and data transformation

(a) A two color (Hoechst and CellTracker Red) high-resolution 3D image reference dataset (10 spheroids) is generated by using dual-view inverted selective plane microscopy (di-SPIM). 3D Segmentation and image analysis enables counting of nuclei and the calculated cell number – spheroid size relationship is used to estimate cell numbers from pheno-seq data.

(b) Correcting for lost correlation of cell numbers and library complexity by approximating total mRNA abundances in spheroids of different sizes. Raw mRNA counts are divided by estimated cell numbers and the calculated minimal average mRNA count is used to transform the data by downsampling counts to 2300 counts per cell in the whole CRC phenoSeq dataset. This strategy results in a perfect correlation of cell numbers and mRNA counts (estimated cell number plotted against transformed mRNA counts).

(c) Pearson's correlation coefficients (r) distributions of gene expression and cell numbers for all 13,868 genes before and after data transformation.



Supplementary figure 13 | Gene-specific and global correlation analysis of gene expression and estimated cell numbers after data transformation

(a) Scatter plots of estimated cell numbers plotted against downsampled mRNA counts (see Supplementary Fig. 9 for data transformation) and associated Pearson's correlation coefficients (r) for housekeeping gene ACTB and differentiation markers TFF3 and DEFA5.

(b) Distribution of Pearson's correlation coefficients (r) distributions of gene expression and cell numbers for all genes before and after data transformation subdivided into non-2 population genes (left) and 2-population genes (right) as identified by maximum likelihood inference (see Fig. 3).