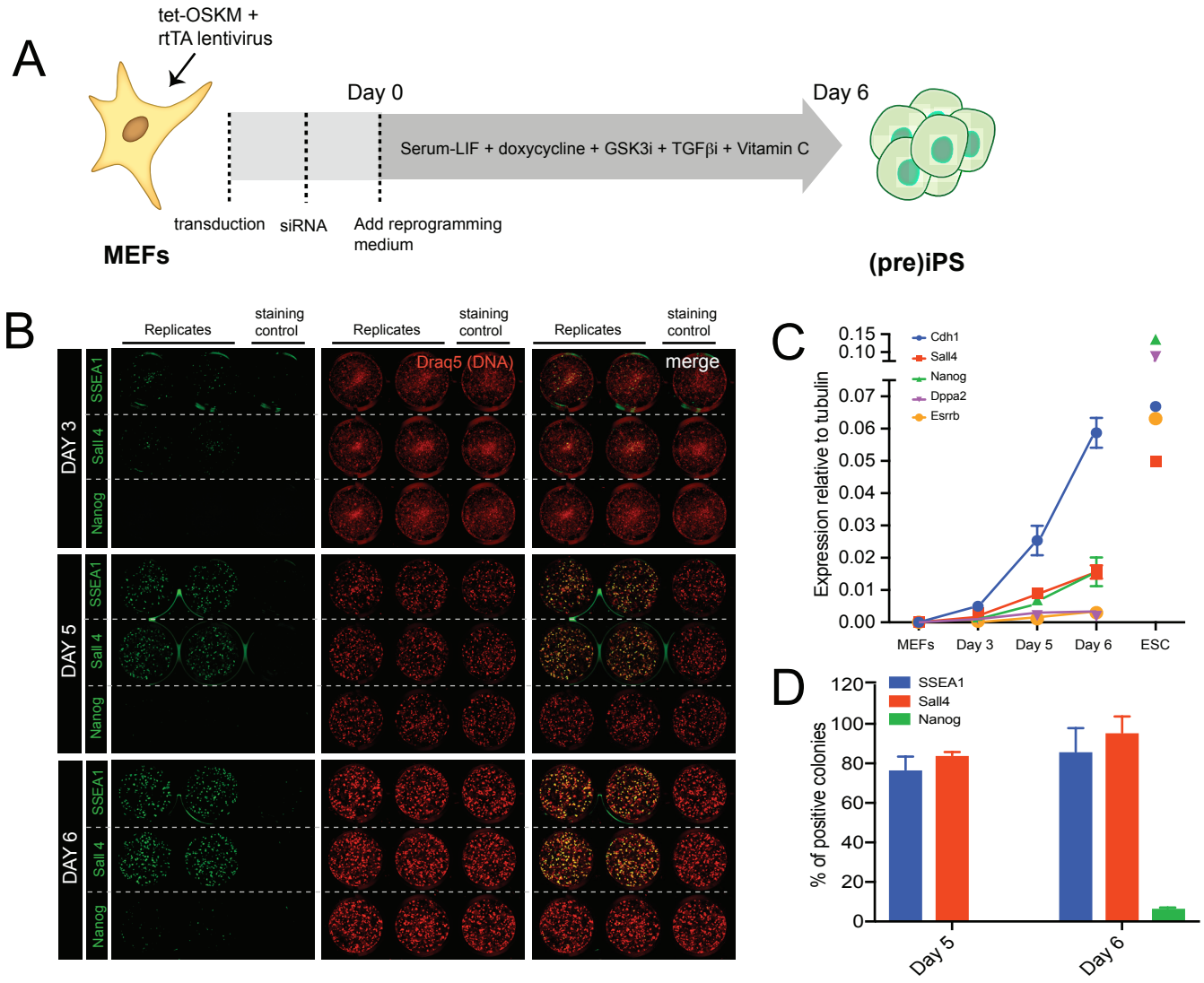

Figure S1. Characteristics of the reprogramming system. Related to Figure 1.
(A) Schematic representation of the reprogramming protocol used. (B) Expression dynamics of some pluripotency proteins determined at Day 3, Day 5 and Day 6 of reprogramming by in-cell western, imaged in 12-well plates. The staining control corresponds to a sample stained without primary antibody. (C) mRNA expression of early and later pluipotency markers of reprogramming timepoints. Expression levels are compared to those of Embryonic Stem Cells (ESC) at the right part of the plot. Data are presented as mean ± SD of biological replicates. (D) Quantification of positive colonies from B for each marker (Sall4, SSEA1 and Nanog) at Day 5 and Day 6. The calculation corresponds to the ratio of positive-colonies to Draq5. Day 3 was difficult to determine because of the fuzzy pattern. Data are presented as mean ± SD.

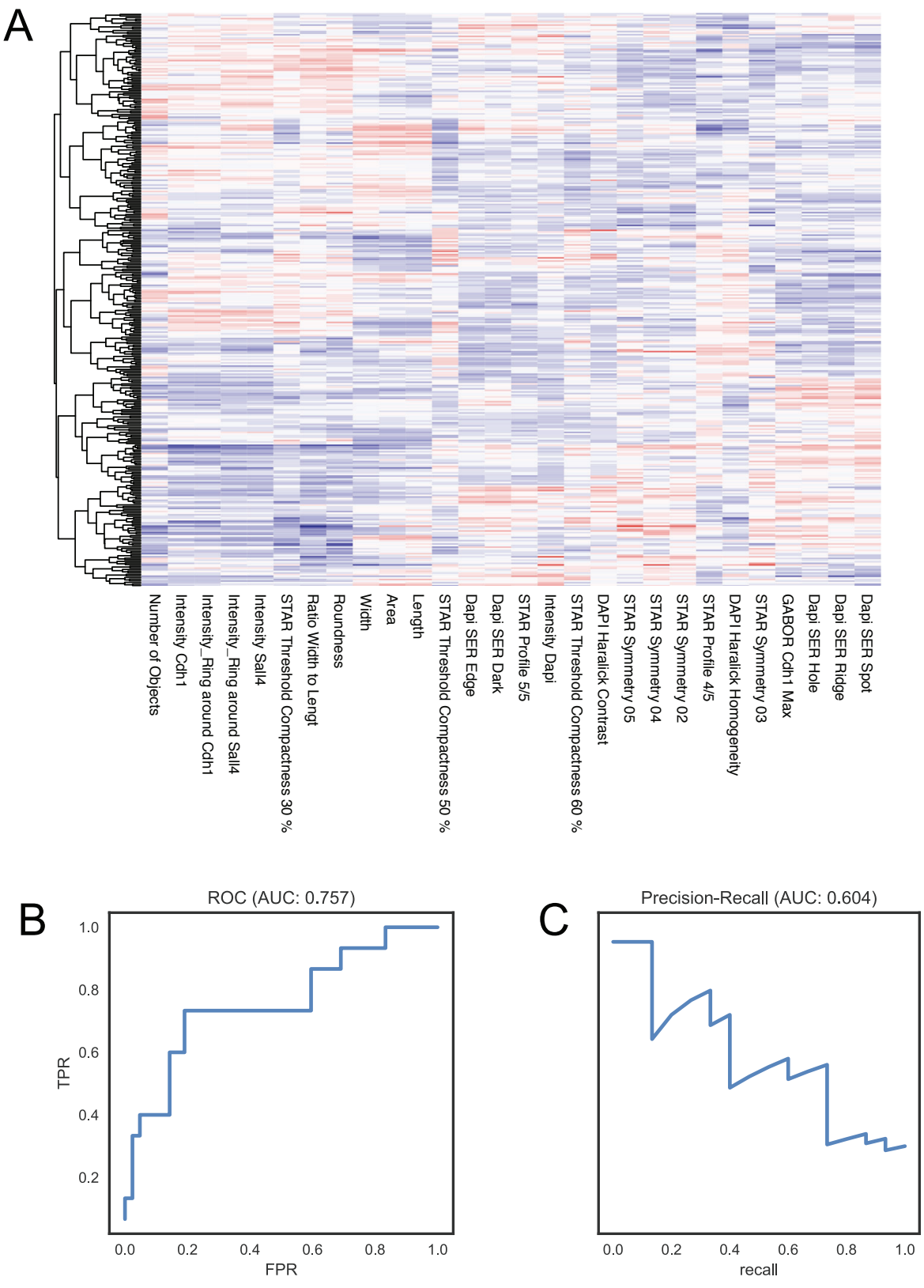

Figure S2. Unsupervised clustering and machine learning classification of high-content phenotypes. Related to Figure 2
(A) Hierarchical clustering based on Pearson correlation distance metrics applied to the whole screening dataset with filtered features.
The performance of the ensemble of the two machine learning classifiers in predicting reprogramming facilitators. Metrics were calculated based on the average probability in a leave-one-out cross-validation procedure. (B) The Receiver-Operator Curve (ROC), with an area under the curve (AUC) of 0.757. (C) The Precision-Recall curve with an AUC of 0.604.

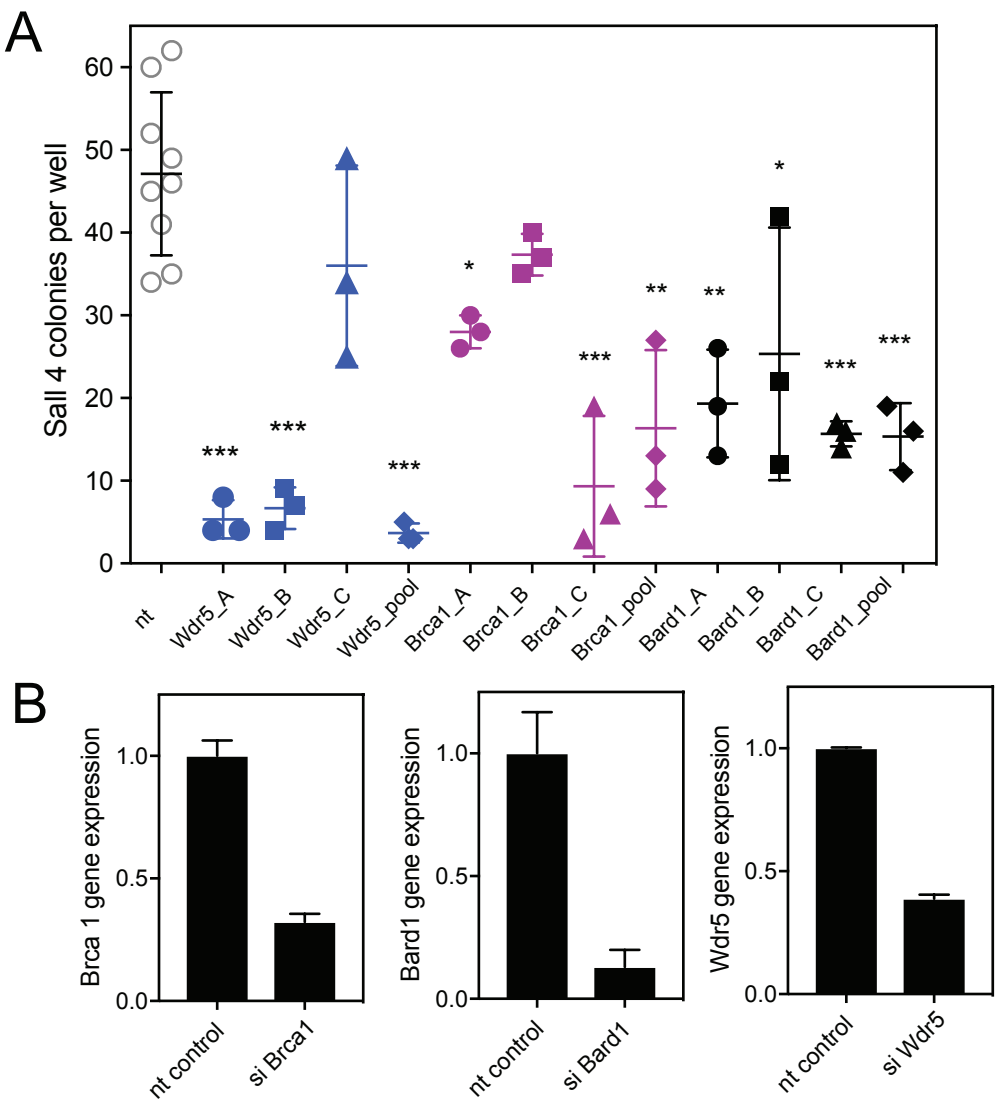

Figure S3. Analysis of Wdr5, Brca1 and Bard1 siRNA target specificities. Related to Figure 3.
(A) Sall4-positive colony formation assay for deconvolution of 3 siRNA sequences targeting Wdr5 (blue), Brca1(magenta) or Bard1. In all cases, at least two out of the three sequences elicit the same colony-phenotype as the pooled siRNAs at day 6. Each dot represents one independent transfection (B) Knockdown efficiencies of siWdr5, siBrca1 or siBard1 for their targets analyzed at day 3 by RT-qPCR. Data represent mean ± SD

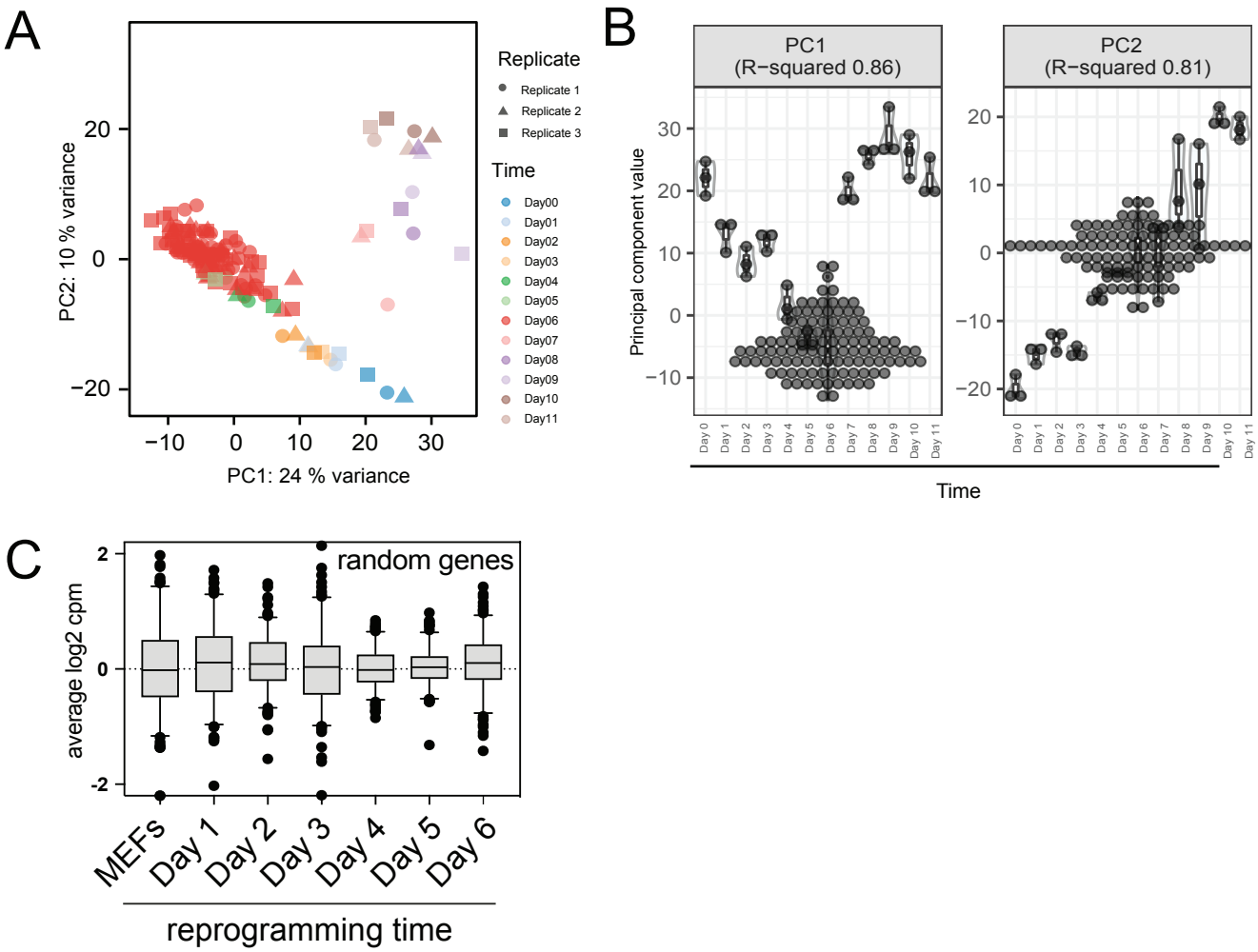

Figure S4. PCA analysis on CELSeq datasets. Related to Figure 3.
(A) PCA analysis of knockdowns (Day 6 red) together with timecourse. (B) Correlation (r-squared value) with PC1 and PC2 dataset in A. PC1 and PC2 highly correlate with time. (C) Boxplots representing expression of a set of 150 random genes in reprogramming time. The values are normalized in log2-counts per million reads (log2-cpm).

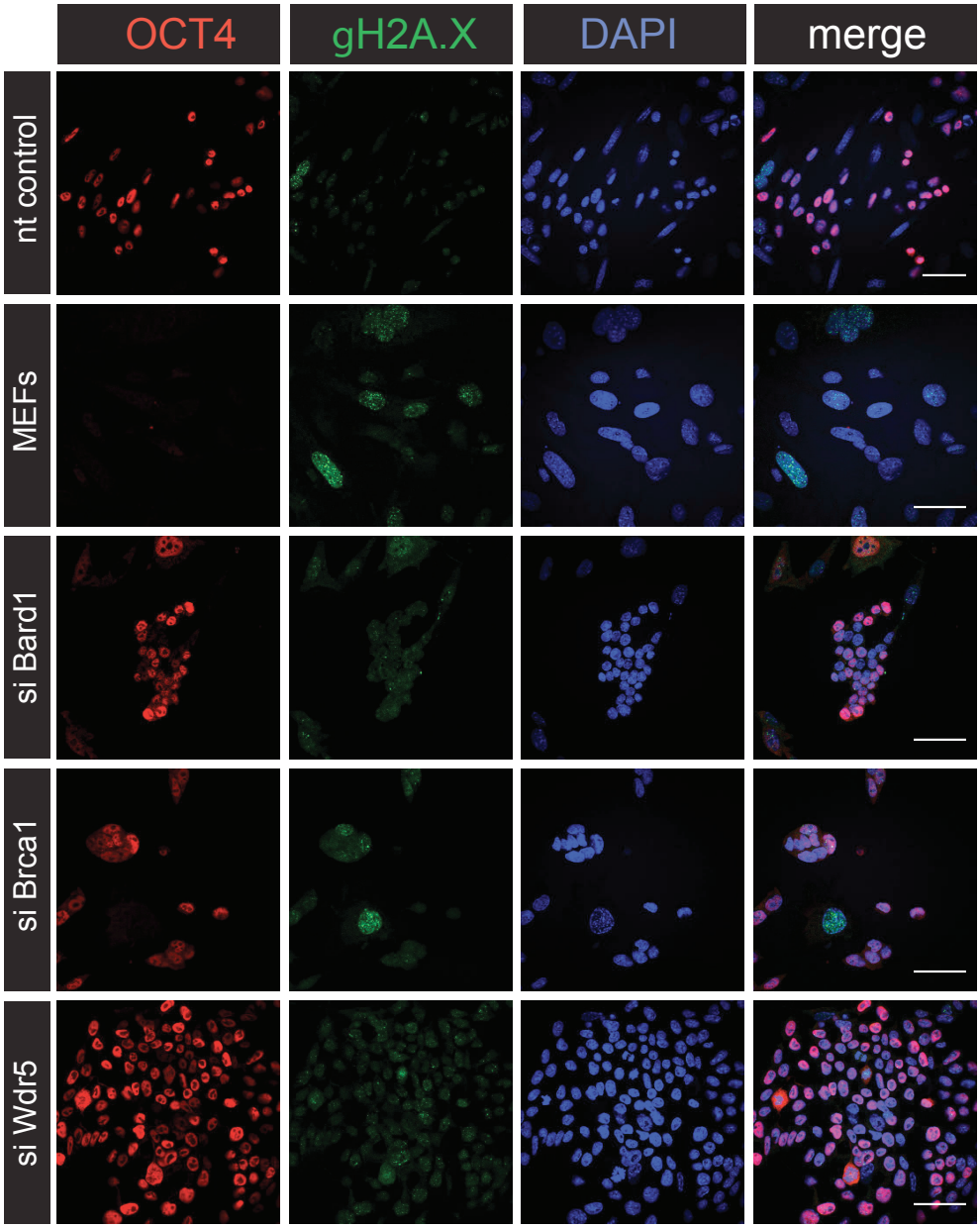

Figure S5. γH2A.X detection in reprogramming cells also stained for Oct4. Related to Figure 5.
Confocal microscopy images showing staining of exogenous Oct4 and γH2A.X, counterstained with DAPI at day 3 in perturbed reprogramming populations (siBrca1, siBard1, siWdr5), normal reprogramming (nt control) and MEFs (mouse embryonic fibroblasts). Scale bar is 100 μm.

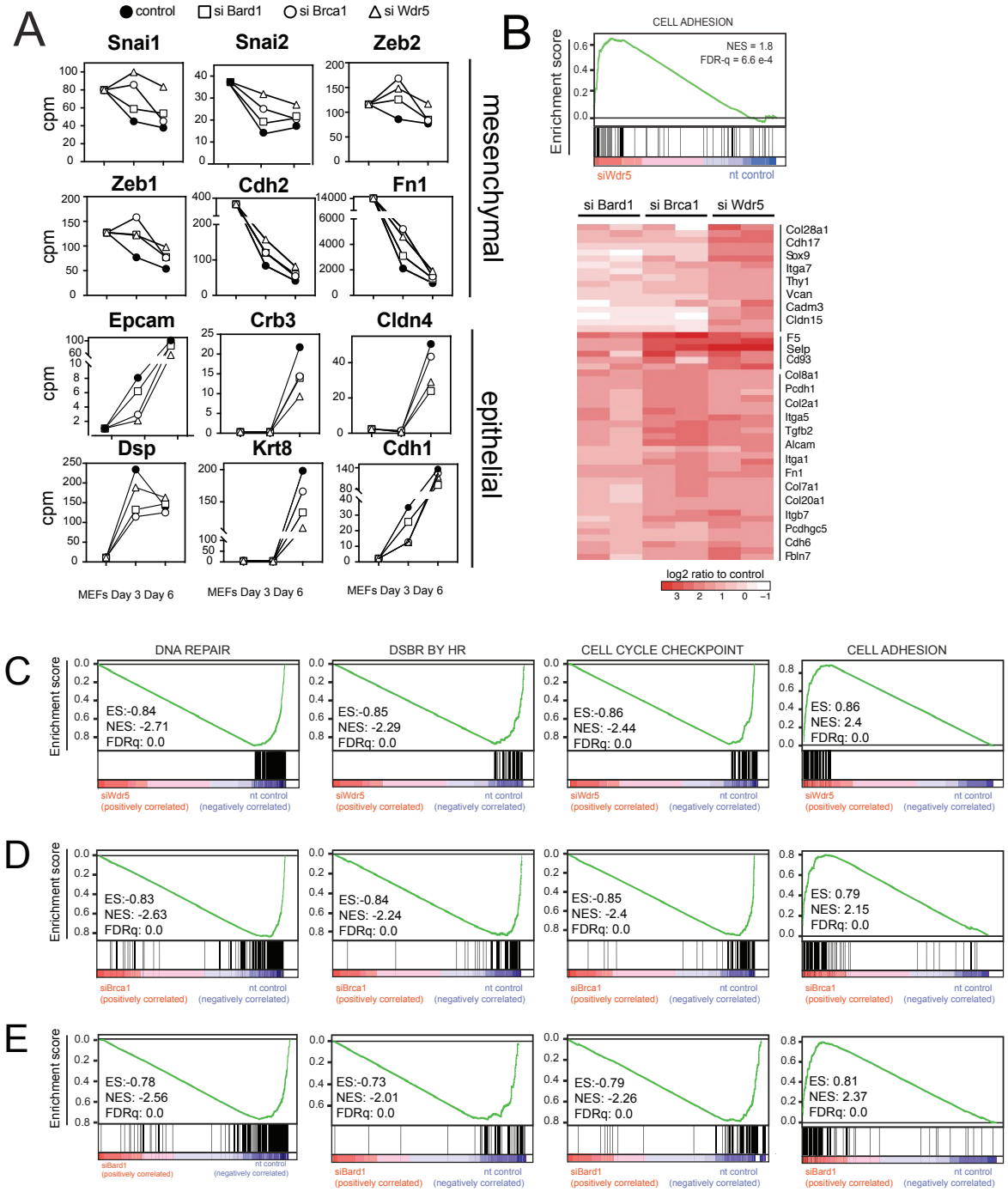

Figure S6. MET gene expression changes are related to cell adhesion molecules. Related to Figure 6
(A) Epithelial and mesenchymal gene expression in MEFs, Day 3 and Day 6 of reprogramming, comparing nt control, Bard1, Brca1 and Wdr5 depleted cells. cpm= counts per million reads. Data are presented as mean of two replicates from RNA-seq. (B) Leading edge analysis for Wdr5 vs control, using the leading edge genes extracted from Wdr5 vs control itself. Enrichment score for cell adhesion genes, comparing siWdr5 vs nt control (upper). Heatmap representing the log2-ratio compared to the control of cell adhesion genes in Brca1, Bard1 and Wdr5 depleted cells at day 3 of reprogramming (lower). (C-E) Leading edge analysis for Wdr5, Brca1 or Bard1 knock-downs vs control, using the core genes extracted from Wdr5 vs control. The green curve represents the enrichment score. Enrichment Score (ES), Normalized Enrichment Score (NES), False Discovery Rate q value (FDRq).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

MEF Reprogramming

MEF medium consisted of DMEM (High glucose, GlutaMAX, Thermo Scientific), 15% FBS, 1% Penicilin/Streptomycin, 1% non-essential amino acids (Thermo Scientific) and 100 nM ß-mercaptoethanol (Sigma).
In case of siRNA transfection, transduced cells were then incubated with the siRNA transfection mix for one more day before starting reprogramming. Reprogramming initiated by adding medium based on DMEM, 10% FBS (Hyclone), supplemented with 2μg·mL-1 doxycycline, 3 μM Chiron (GSK3-inhibitor), 0.25 μM Alk5i (TGF-ß inhibitor), 50 μg·mL-1 ascorbic acid and $1 \cdot 10^3$ $U \cdot mL^{-1}$ LIF. Considering this point as day 0, the protocol continued for 6 days for most of the experiments, unless specified otherwise.

High content screening analysis and hit-selection

Besides the geometric and basic morphology features, texture features for whole cell regions were calculated using the SER (Spot, Edges and Ridges) method. There are eight SER features; Spot, Hole, Edge, Ridge, Valley, Saddle, Bright and Dark. Another parameter used in the SER method is the scale, which is smoothing technique of filtered images. The texture features were normalized using the Kernel method, which means that the images are pixelwise divided by the smoothed original image.
STAR advanced morphological features (Symmetry properties, Threshold compactness, Axial properties, Radial properties) were calculated. These set of features measure the symmetric distribution of the fluorescence intesities or textures and how compact they are within the region of interest. For the STAR profiles, the image region is subdivided in different sections and measured.
For data visualization and clustering, the quadruplicate values per knockdown were averaged, including the controls. Hierarchical clustering and visualization was performed with pheatmap R package.
To select the top-hits we used a combined approach. One part of it was based on a top ranking score, which was calculated by correlation of knockdowns with positive controls (siOct4, siMyc, siTrp53) and anti-correlation with non-targeting controls. Some of the top-ranking siRNAs were selected this way. The other part of the selection was based on a machine learning prediction. For that, we defined a set of known reprogramming facilitators based on literature (Table S4) and used these to train an ensemble of classifiers based on all features (Table S2). This ensemble consisted of an L2 penalized logistic regression and a random forest classifier. The hyperparameter of the L2 regularization of the logistic regression was set using three-fold cross-validation. The random forest model was trained with 1000 trees and a maximum of 10 features. The final probability was calculated as the mean of the predicted probability of the two classifiers. The performance was assessed by the ROC AUC (0.757) and the Precision-Recall AUC (0.604) based on leave-one-out cross-validation.

In-cell western, for colony counts

Samples were fixed with 4 % PFA for 15 minutes, followed by 0.3% triton X-100 permeabilization and 1 % BSA blocking for 30 minutes. After that, they were incubated with either Sall4 (Abcam; ab29112), SSEA1(R&D Systems; MAB-2155), or Nanog (eBioscience; 14-5761-80). Cells were washed twice with 1% PBS and stained with appropriate IRDye 800CW secondary antibodies (LI-COR) and counterstained with Draq5 (Thermo Scientific). As a staining control, a sample with only the secondary antibody was used to adjust the brightness and contrast properly, to avoid false-positives.
Images were acquired with an Odyssey CLx Infrared Imaging System (LI-COR) and Image Studio v5.0 software and adjusted for brightness and contrast with ImageJ and Adobe Photoshop CS6.

Double knockdowns and deconvolution transfections

Reprogramming was started in 12-well or 48-well plate formats, with transfection reagents and number of cells scaled according to previous descriptions (see Experimental Procedures). For the double knockdown, a mixture of two pooled siRNAs was used, with a final concentration of 40 nM. The corresponding single knockdowns were performed with 20 nM siRNA target + 20 nM nt control siRNA, to compensate for the lower individual siRNA dose in the double knockdowns.
As for the siRNA deconvolution experiments, each of the 3 siRNA target sequences was transfected at 40 nM final, and the pooled siRNAs were combined to give the same final concentration.
After six days of reprogramming, cells were fixed and stained (see in-cell western above). A CellProfiler script (Jones et al., 2008), kindly adapted by Jessie van Buggenum (https://github.com/jessievb/automated_CFA) was used to quantify Sall4-colony numbers automatically.

RNA isolation and RT-qPCR

Total RNA was isolated with Quick- RNA™ MicroPrep (Zymo Research) following the manufacturer´s instructions. RNA was eluted in RNAse-free water. Concentration was measured by absorbance with Nanodrop and we used Bioanalyzer (Agilent) to assess the quality and integrity of samples. cDNA was synthesized from 120-180 ng RNA with Superscript lll kit (Thermo Scientific). Each RT-qPCR reaction was done with 1-2 ng diluted cDNA and SYBR green mix (iQ-SYBR-Green Supermix, Biorad) and 10 μM primer pairs. Relative gene expression was calculated using the ΔΔCt method, using the housekeeping gene Gapdh as a reference.

CELseq2-RNAseq

The following adaptations were done to the original sample preparation: 100 pg purified RNA was directly added to a reverse transcription mixtures containing Maxima H Minus (ThermoFisher) reverse transcriptase and CELseq2 primers with a 6-nucleotide sample barcode and 8-nucleodide UMI. After reverse transcription samples were pooled and purified using AmpureXP beads (Beckman Coulter). Second strand synthesis and following steps were performed according to the original protocol.

The matrix with all the counts was analyzed with scater R package v. 1.3.49 (McCarthy et al., 2017) in order to assess the overall quality of the samples and to filter out those with low quality, based on default parameters. Scran v.1.6.9 (Lun et al., 2016) and scater R packages were used to correct for batch effects by regression, and also to normalize (log2-cpm).

Principal Component Analyses were conducted in scater and in gplots R packages. Most relevant Principal Components we found were PC1 and PC2. To analyze the most variable genes in the knockdowns (Figure 3B), the top 200 features (genes) correlating with either PC1 or PC2 were extracted. This set of genes was used for hierachical clustering based on the Pearson correlation of the knockdowns seen in Figure 3B.

RNA-seq analysis: Gene Ontology and GSEA anaysis

For differential gene expression analysis, negative control was compared to the knockdowns using DESeq2 v.1.18.1 (Love et al., 2014). We filtered differential genes with a cutoff of log2-foldchange>1 and padj<0.05 as for gene ontology analysis with DAVID v.6.7

(Huang da et al., 2009).

For heatmap visualization and GSEA v.3.0, data log-cpm with EdgeR package v. 3.20.9 (Robinson et al., 2010).

All gene sets used for Gene Set Enrichment Analysis (Subramanian et al., 2005) , were obtained by process search in the Gene Ontology database. Normalized and low-count filtered RNA-seq expression datasets were used as an input for the initial enrichment analysis , which initially compared siWdr5 vs. control (Figure 5). The basic parameters used were 1000 gene-set permutations to calculate the enrichment score by the log2-ratio of classes metric.

Subsequently, leading edge analysis (Mootha et al., 2003) was performed for Wdr5 vs control. Leading edge refers to the core genes contributing to the enrichment score. These genes were extracted from each of the sets (cell cycle checkpoint, DNA repair, DNA repair homologous recombination and cell adhesion). Those core genes were further used to determine the leading edge genes of either siBrca1, siBard1 or siWdr5 vs. control (Figure S5). The overlap of leading edge genes between the 3 knockdowns is what we have depicted in Figure 5 C.

SUPPLEMENTAL REFERENCES

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4, 44-57.

Jones, T.R., Kang, I.H., Wheeler, D.B., Lindquist, R.A., Papallo, A., Sabatini, D.M., Golland, P., and Carpenter, A.E. (2008). CellProfiler Analyst: data exploration and analysis software for complex image-based screens. BMC bioinformatics 9, 482.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15, 550.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 5, 2122.

McCarthy, D.J., Campbell, K.R., Lun, A.T., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 33, 1179-1186.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics 34, 267-273.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102, 15545-15550.