Stem Cell Reports, Volume *2*

Supplemental Information

Conserved Two-Step Regulatory Mechanism of Human Epithelial Differentiation

Jayant K. Rane, Alastair P. Droop, Davide Pellacani, Euan S. Polson, Matthew S. Simms, Anne T. Collins, Leo S.D. Caves, and Norman J. Maitland

Inventory of supplemental information:

1. Figure S1: qRT-PCR analysis showing the expression of the candidate genes in benign and malignant primary prostate epithelial cultures. This figure is in addition to figure 2D from the main text.
2. Figure S2: qRT-PCR analysis showing the expression status of retinoic acid receptors in benign and malignant primary prostate epithelial cultures. This figure is in addition to figure 3B from the main text.
3. Figure S3: qRT-PCR analysis showing the expression status of *LCN2, CEACAM6, S100P*, and *TMPRSS2* in basal and luminal cells enriched from benign and malignant primary prostate epithelial cultures. This figure is in addition to figure 4A from the main text.
4. Figure S4: Context dependent changes in the expression of candidate genes. This figure is in addition to figure 4C from the main text.
5. Table S1: Enrichment for GO terms associated with low variance genes. This figure is in addition to figure 2B from the main text.
6. Supplementary experimental procedures

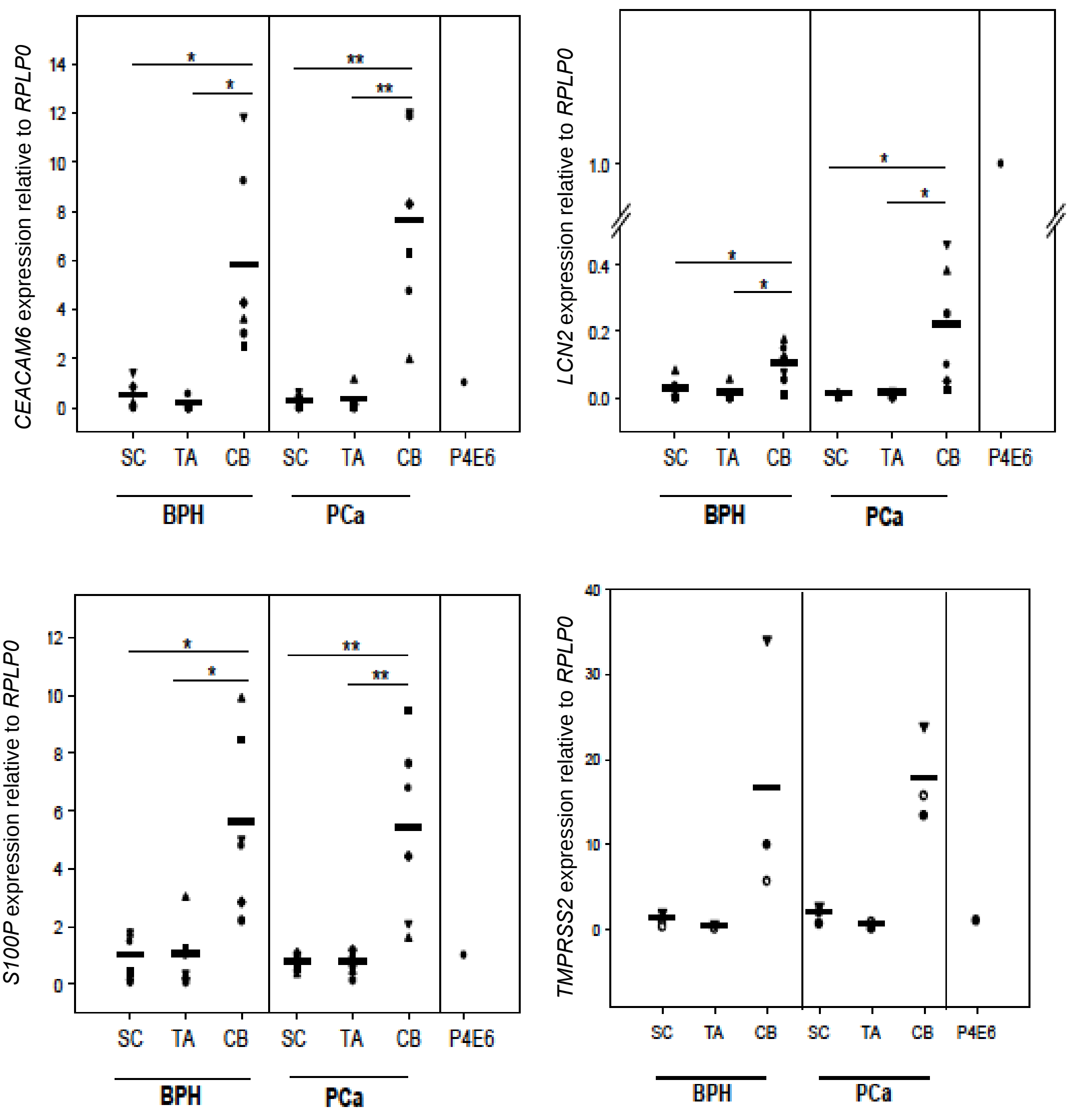

Figure S1: qRT-PCR analysis showing the expression of the candidate genes in benign and malignant primary prostate epithelial cultures. Horizontal black line indicates mean. SC: stem cells, TA: transit amplifying cells, CB: committed basal cells, BPH: benign prostatic hyperplasia, PCa: prostate cancer. Each point represents RNA from an individual patient biopsy. For *CEACAM6, LCN2*, and *S100P*, n = BPH (6) and PCa (6). For *TMPRSS2*, n=BPH (3) and PCa (3). All biological replicates.

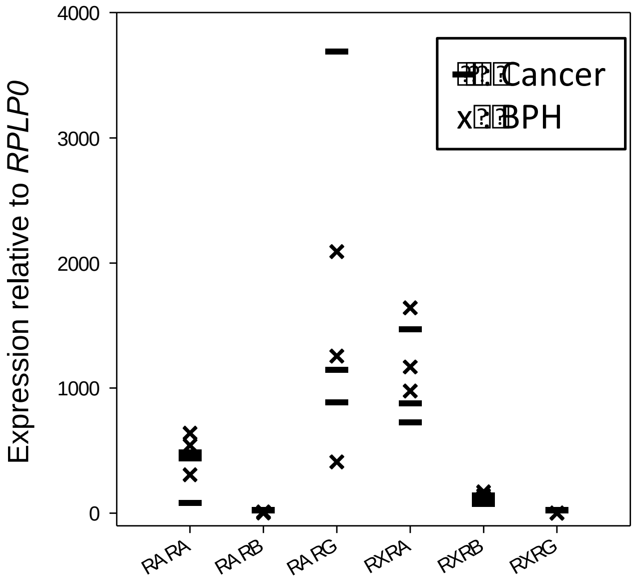

Figure S2: qRT-PCR analysis showing the expression status of retinoic acid receptors in benign and malignant primary prostate epithelial cultures. Horizontal black line indicates mean. BPH: benign prostatic hyperplasia. N = BPH (3) and PCa (3) (biological replicates).

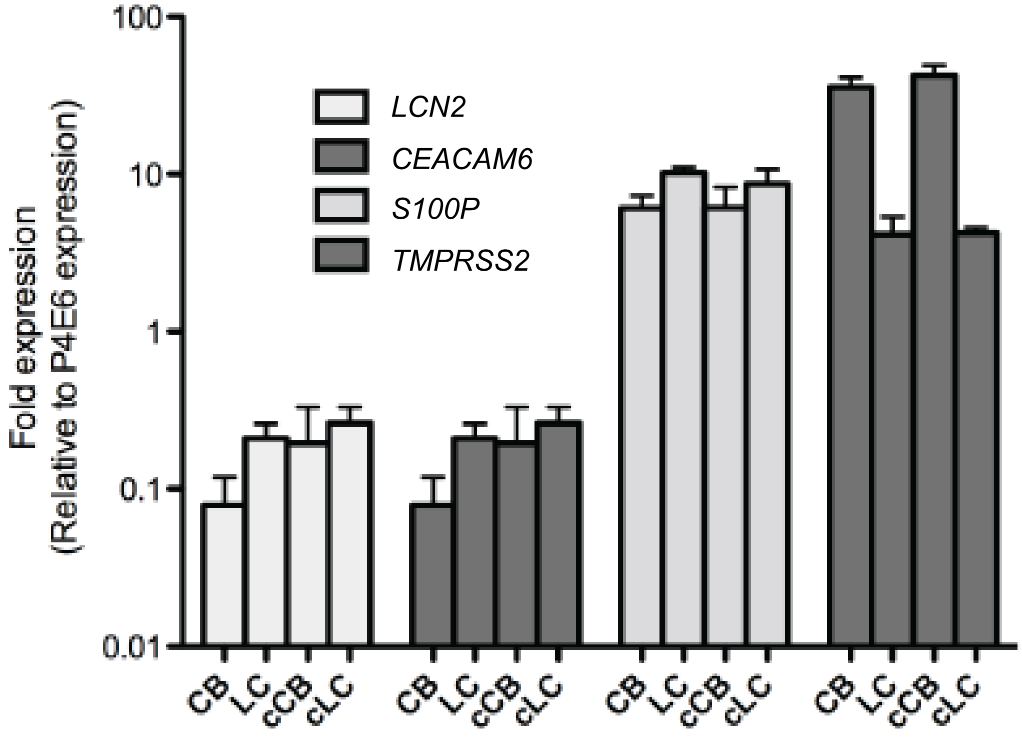

Figure S3: qRT-PCR analysis showing the expression status of *LCN2, CEACAM6, S100P*, and *TMPRSS2* in basal and luminal cells enriched from benign and malignant primary prostate epithelial cultures. CB: BPH-derived committed basal cells, LC: BPH-derived luminal cells, cCB: PCa-derived committed basal cells, cLC: PCa-derived luminal cells. N=4 each for BPH and PCa, all biological replicates.

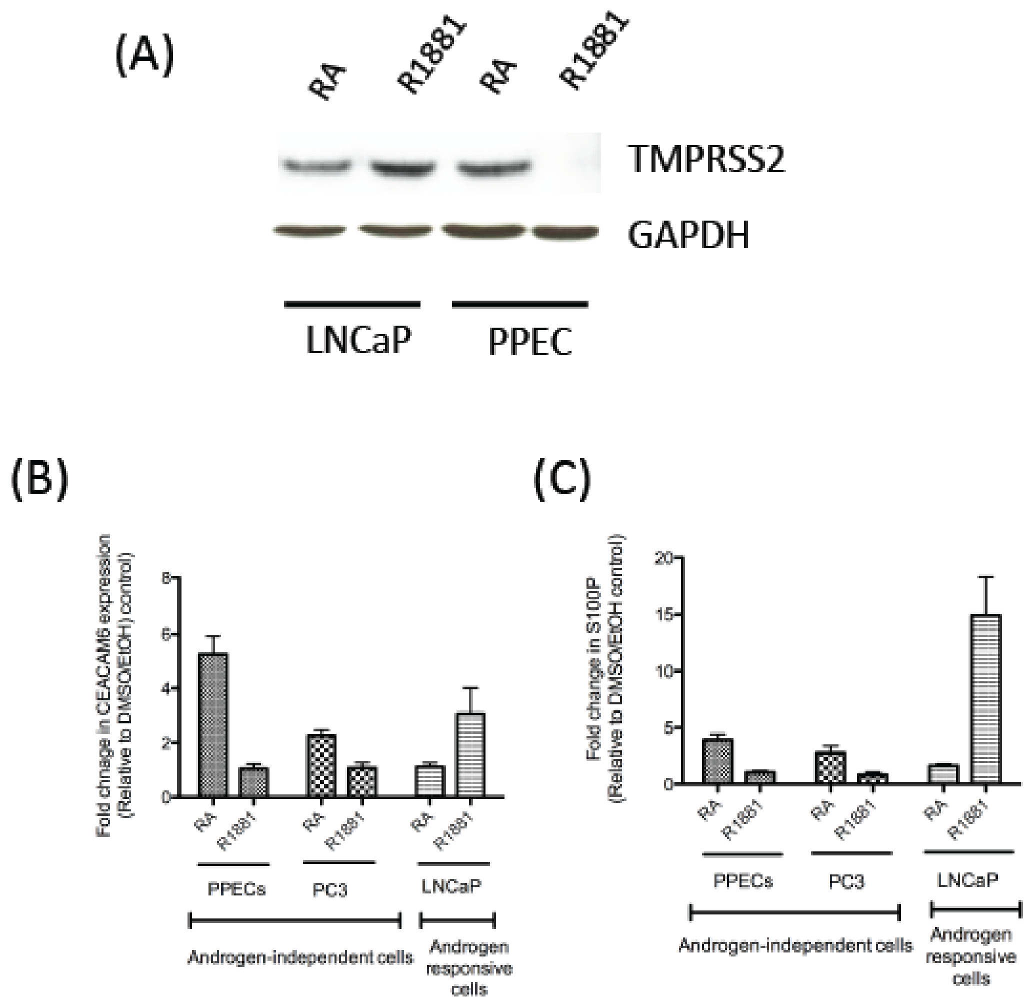

Figure S4: Context dependent changes in the expression of candidate genes. A: Western blot showing changes in TMPRSS2 protein expression in LNCaP and primary prostate epithelial cultures (PPEC) after 10nM R1881 or 100nM at-RA (RA) treatment for 72 hours (n=3, representative image). qRT-PCR analysis of *CEACAM6* (B) and *S100P* (C) expression after 48 hours treatment with at-RA (RA) and R1881 (n=3, biological replicates for PPECs and experimental replicates for PC3 and LNCaP).

GO Term	**n_t**	**n_s**	**p-value**
translational elongation	225	61	6.40e-111
Translation	448	55	3.30e-78
structural constituent of ribosome	327	48	2.33e-71
Ribosome	371	49	1.45e-70
cytosolic small ribosomal subunit	73	31	1.08e-61
cytosolic large ribosomal subunit	91	17	5.83e-27
ribonucleoprotein complex	259	19	5.88e-22
rRNA binding	40	11	5.90e-20
small ribosomal subunit	34	10	1.32e-18
response to calcium ion	125	11	3.95e-14
eukaryotic translation elongation factor complex	12	6	2.04e-13
ribosomal small subunit biogenesis	32	7	1.91e-12
structural constituent of cytoskeleton	183	11	2.68e-12
cellular component movement	253	12	4.52e-12
rRNA processing	199	11	6.68e-12

Table S1: Enrichment for GO terms associated with low variance genes. Enrichment p-values are calculated for the top or bottom 100 probes in the dataset (by coefficient of variation), against a total of 35,250 annotated probes. n_t indicates the total number unique probes with the given GO term annotation, whilst n_s gives the number of probes with the corresponding annotation in the set of 100 probes. The enrichment p-value is calculated using a 2-tailed hypergeometric test with mid-P-value correction. Very broad terms such as ‘intracellular’ are meaningless in this analysis and were removed from the low variation subset.

Supplementary experimental procedure

Definition of benign and cancer cultures

Apart from directed biopsy and pathological confirmation, we believe that cultures derived from cancer samples are indeed predominantly cancer cultures due to following evidence:

i. They express high levels of the classical prostate cancer marker AMCAR (D.P. unpublished data)
ii. They are more proliferative, migratory and invasive in-vitro (Collins et al., 2005)
iii. When implanted as an intact tissue, cancer samples can form sub-cutaneous and intra-prostatic xenografts, which can be serially passaged (Kroon et al., 2013; Maitland et al., 2011) and ATC & NJM Unpublished data
iv. About 50% of them express characteristic prostate cancer-associated fusion TMPRSS2-ERG (Polson et al., 2013)
v. They significantly overexpress telomerase compared to BPH derived cultures (JKR & NJM unpublished data)
vi. They have distinct miRNA profile (JKR & NJM unpublished data) and are radio-/chemoresistant compared to BPH-derived cultures (NJM unpublished data).

Compendium Dataset Generation & Processing

Data Selection

Although it is possible to combine data from multiple microarray types and species, there are many difficulties with processing and analysing such data. Limiting the input dataset to a single type of microarray, and only considering a single species greatly simplifies the processing and analysis. In this work, only data from Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays hybridised with material from *Homo sapiens* were utilised.

Input datasets were selected from the ArrayExpress database. A complete description of the ArrayExpress database was downloaded in XML format (on 20/11/2009). The raw (.CEL files) downloaded via FTP. As of the download date, 9,435 experiments were contained within the ArrayExpress database. Of these, 806 (8.54%) experiments were selected and the corresponding 820 raw data files downloaded and uncompressed. 18 invalid files were removed during decompression. After extraction, 24,536 (9.40% of the chips present in ArrayExpress) raw data files in .CEL format (each corresponding to a single microarray chip) were present.

Data Cleaning

Although the ArrayExpress database has strict data cleanliness criteria, the large numbers of data files used in this work means that some invalid data will be selected. A CEL file can be either *corrupt*, in which case it is unreadable; or it can have the right format, but the data contained within it can be *invalid*. Both of these types of file errors were eliminated. Corrupt files are found by simply trying to open the file. Invalid data is detected by outlier analysis; density distributions for the raw data PM chip value statistics are calculated, and any chip files with a statistic falling outside a set arbitrary threshold are rejected. This method removes those chips that show unexpected overall behaviour. Raw data validity testing was performed using R and bioconductor. The bioconductor affyio package was used to attempt to open each CEL file in turn. Any error or warning when opening the file indicates a corrupt file structure, and caused rejection of the file. Once the file was determined to be valid, summary statistics for the raw PM data were calculated. Distributions for chip PM minimum, median, maximum and variance were calculated, and the outlying 0.5% of chips for each distribution were identified and removed. 90 files were removed due to invalid formatting, and 417 files were removed by outlier detection.

Processing

After data cleaning, RMA processing was performed on the complete dataset. Expression density curves for each microarray in the resulting dataset are shown in figure SEP1. The RMA normalisation step makes the distribution of probe intensities identical across microarrays. Differences between the probe density distributions over all microarrays are expected due to the summarisation of several probes into a single probe set.

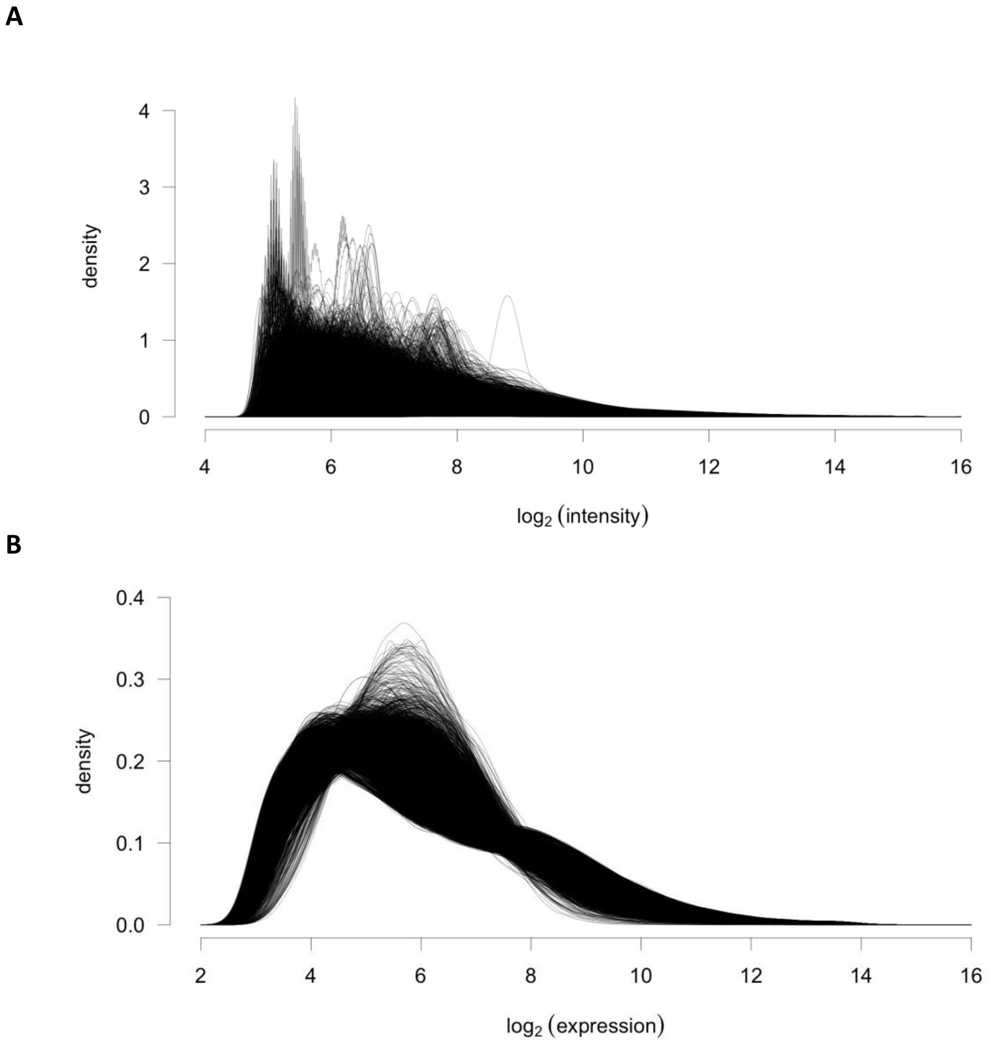

Figure SEP1: Density curves for the 24,029 chips in the complete dataset. (**A)** Probe mean intensity data before pre-processing (**B)** The probeset expression density after RMA preprocessing. During preprocessing, a median polish process is used to bring the density of all individual probes into alignment. Subsequent summarisation of multiple probes to generate a single value for each probeset introduces differences between individual chips. More rigorous outlier removal would result in a cleaner input dataset, and thus bring all chips into closer alignment at the expense of removing more chips from the dataset.

After RMA preprocessing, a data matrix of 54,675 $\log_2$ signal values across 24,029 microarrays from 806 experiments was obtained. No attempt was made to remove control probes from the dataset, as these are often necessary for internal validation.

Annotation

Metadata for each experiment was downloaded from ArrayExpress along with the raw data. Annotations are available at both the experiment level, and the chip level. Experiment level annotations are extracted from the experimental description files (.IDF) collected from

ArrayExpress. Chip level annotations are extracted from the .SDRF files collected from ArrayExpress. Although SDRF files exist for each experiment, mapping between metadata and individual arrays is not always possible. 805/806 (99.9%) of the experiments could be uniquely annotated by description. 20851/24029 (86.77%) of the individual arrays could be annotated.

Metadata for individual chips was taken from the “Description” column of the SDRF file; this is a free-text space for researchers to fill in a description of the chip. No ontologies or controlled vocabularies are available for this data. Because of the open nature of the description text, analysis of the chip-level annotations is very difficult. The “TM” package in R was used to mine the annotations.

Compendium Dataset Correlation Analysis

Correlation analyses were performed in R. Due to the very large size of the dataset, a standard correlation analysis using the cor() function was not feasible. Instead, we used the fact that the product of a normalized matrix (in which each row has a mean of 0 and a magnitude of 1) multiplied by its transpose is its correlation matrix. (If the matrix is not normalized, the result is a covariance matrix). A row-wise normalization is very fast, and simply scales the data. After normalization, the correlation analysis can be performed as a simple matrix multiplication. Although this is possible in R, it requires far too much memory to be computationally feasible. For this reason, the complete matrix was broken up into smaller sub-matrices, and the appropriate combinations of matrix multiplications were performed. The R code used to perform these steps is available from the authors on request.

Gene	Probe ID
LCN2	Hs01008571_m1
CEACAM6	Hs03645554_m1
S100P	Hs00195584_m1
TMPRSS2	Hs01120965_m1
NF-kB1	Hs00765730_m1
CSF2	Hs00929873_m1
WNT5A	Hs00998537_m1
PAP (ACPP)	Hs00173475_m1
PSA (KLK3)	Hs03047000_s1
AR	Hs00171172_m1
NKX3.1	Hs00171834_m1
RPLP0	Hs99999902_m1

TaqMan probes used for qRT-PCR analysis.

	Company	**Catalogue number**	**Experimental conditions**
Western blot			
LCN2	Abcam	ab23477	1:150 for overnight at 4^{o}C
CEACAM6	Abcam	ab56234	1:2000 for overnight at 4^{o}C
S100P	Cell Signaling	7677	1:1000 for overnight at 4^{o}C
TMPRSS2	Abcam	ab115265	1:200 for overnight at 4^{o}C
B-actin (ACTB)	Sigma	A5316	1:10000 for 1 hour RT
Immunofluorescence			
Cytokeratin 5	Vector Laboratories	VP-C400	1:200 for 1 hour RT
Cytokeratin 18	Sigma	C8541	1:200 for 1 hour RT
RXRA	Santa Cruz Biotechnology	sc-553	1:200 for 1 hour RT
RARA	Santa Cruz Biotechnology	sc-551	1:200 for 1 hour RT
RARG	Santa Cruz Biotechnology	sc-550	1:200 for 1 hour RT
FACS			
CD49f (ITGA6)	eBioscience	11-0495-80	1:200 for 20 minutes RT
CD49b (ITG2)	Serotec	MCA743F	1:200 for 20 minutes RT

Antibodies used for expression analysis.

Dataset	**Cell/Sample Name**	**Cell type**	**Treatment Compound**	**Concentration**	**Time**
EMEXP3577	HUES-1	Human ESCs	at-RA	200nM	7 Days
GSE9169	SK-N-SH	Human neuroblastoma cells	at-RA	10 μM	72h
GSE10434	Skin	Skin biopsies	13-cis RA	0.5-0.67 mg/kg/d	7Days
GSE22298	keratinocytes	primary human epidermal keratinocytes	at-RA	1 μM	72h

Published microarray datasets used to assess the effect of retinoic acid stimulation on group C

Dataset	**Cell/Sample Name**	**Cell type**	**Treatment Compound**	**Concentration**	**Time**
GSE17044	LNCaP	Prostate Cancer Cells	R1881	1nM	48h
GSE22606	LNCaP	Prostate Cancer Cells	R1881	5nM	48h
GSE29232	RWPE-1-AR	Prostate Epithelial Cells overexpressing AR	R1881	1nM	24h

Published microarray datasets used to assess the effect of androgen stimulation on group C

References:

Collins, A.T., Berry, P.A., Hyde, C., Stower, M.J., and Maitland, N.J. (2005). Prospective identification of tumorigenic prostate cancer stem cells. Cancer Res *65*, 10946-10951.

Kroon, P., Berry, P.A., Stower, M.J., Rodrigues, G., Mann, V.M., Simms, M., Bhasin, D., Chettiar, S., Li, C., Li, P.K.*, et al.* (2013). JAK-STAT Blockade Inhibits Tumor Initiation and Clonogenic Recovery of Prostate Cancer Stem-like Cells. Cancer Res *73*, 5288-5298.

Maitland, N., Frame, F., Polson, E., Lewis, J., and Collins, A. (2011). Prostate Cancer Stem Cells: Do They Have a Basal or Luminal Phenotype? Hormones and Cancer *2*, 47-61.

Polson, E.S., Lewis, J.L., Celik, H., Mann, V.M., Stower, M.J., Simms, M.S., Rodrigues, G., Collins, A.T., and Maitland, N.J. (2013). Monoallelic expression of TMPRSS2/ERG in prostate cancer stem cells. Nat Commun *4*, 1623.