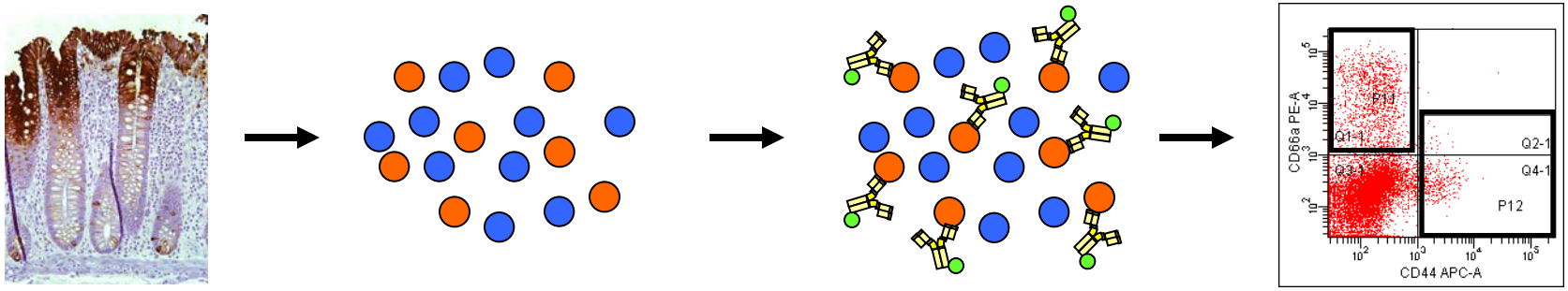
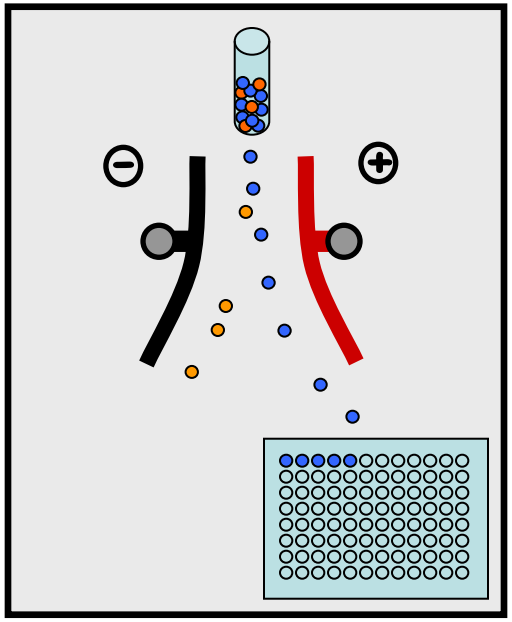
Supplementary Figure 1

Workflow schematic of the SINCE-PCR method.



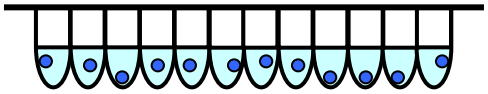
1) Primary tissues are collected from surgical specimens and dis-aggregated into single-cell suspensions.

2) Single-cell suspensions are stained with fluorochrome-conjugated monoclonal antibodies and analyzed by flow cytometry

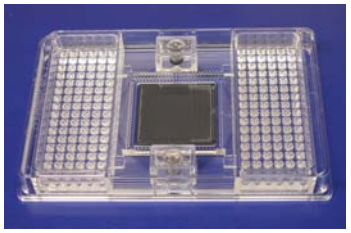


3) Single-cells from selected phenotypic populations are sorted into individual 96-well PCR plates

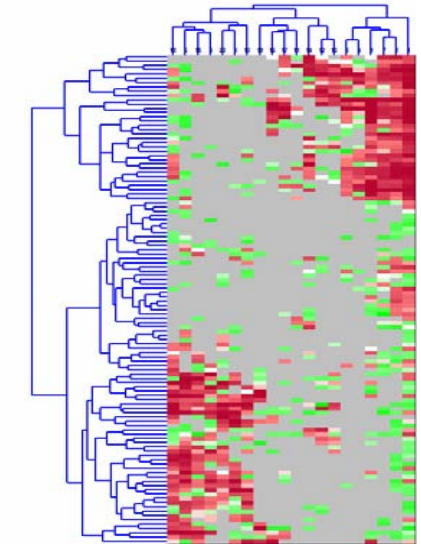
1 cell/well (96-well PCR plate)



4) RNA is reverse transcribed and loaded into M96 qPCR DynamicArray™ chips (Fluidigm®)



5) 9,216 (96x96) single-cell TaqMan PCR reactions are run in parallel using the BioMark™ real-time PCR reader (Fluidigm®)



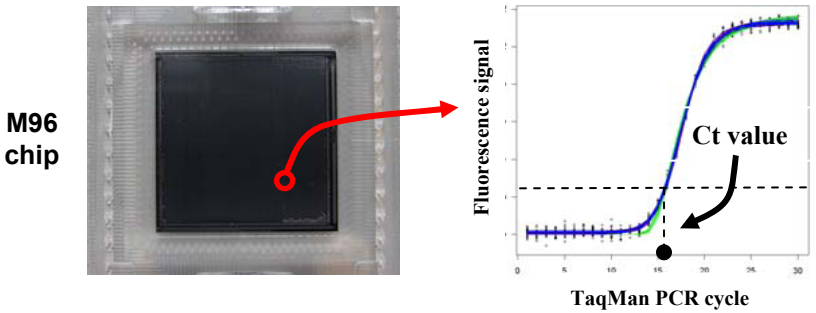
6) Individual cell real-time PCR curves are analyzed and converted into gene-expression levels. Individual cells are associated into distinct subsets using statistical clustering algorithms.

Supplementary Figure 1. **Workflow schematic of the SINCE-PCR method.**

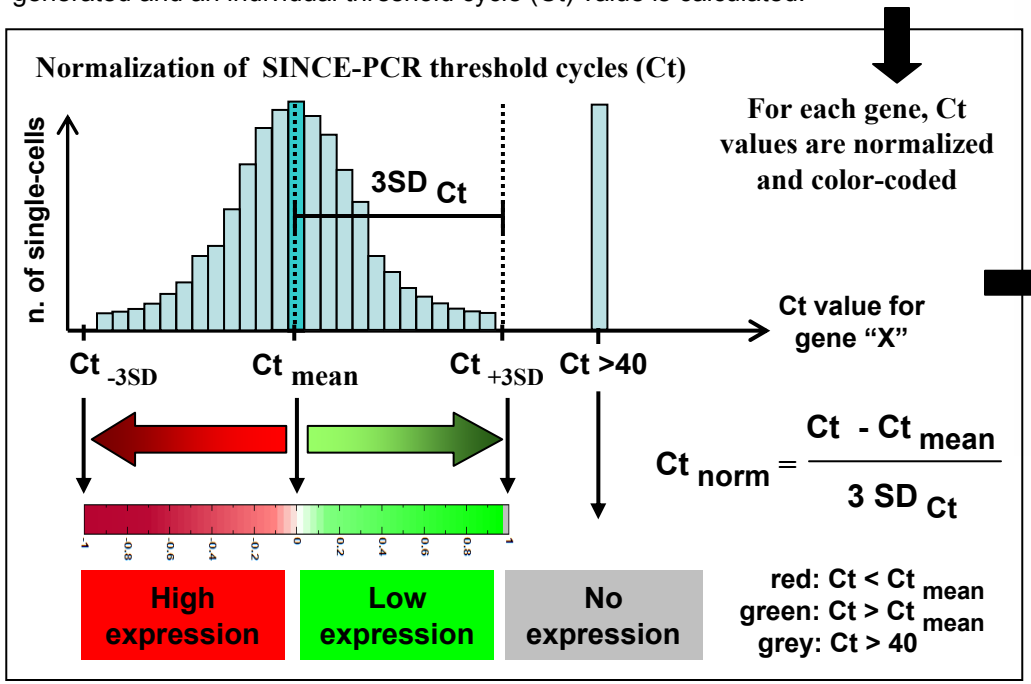
Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.,* **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Supplementary Figure 2

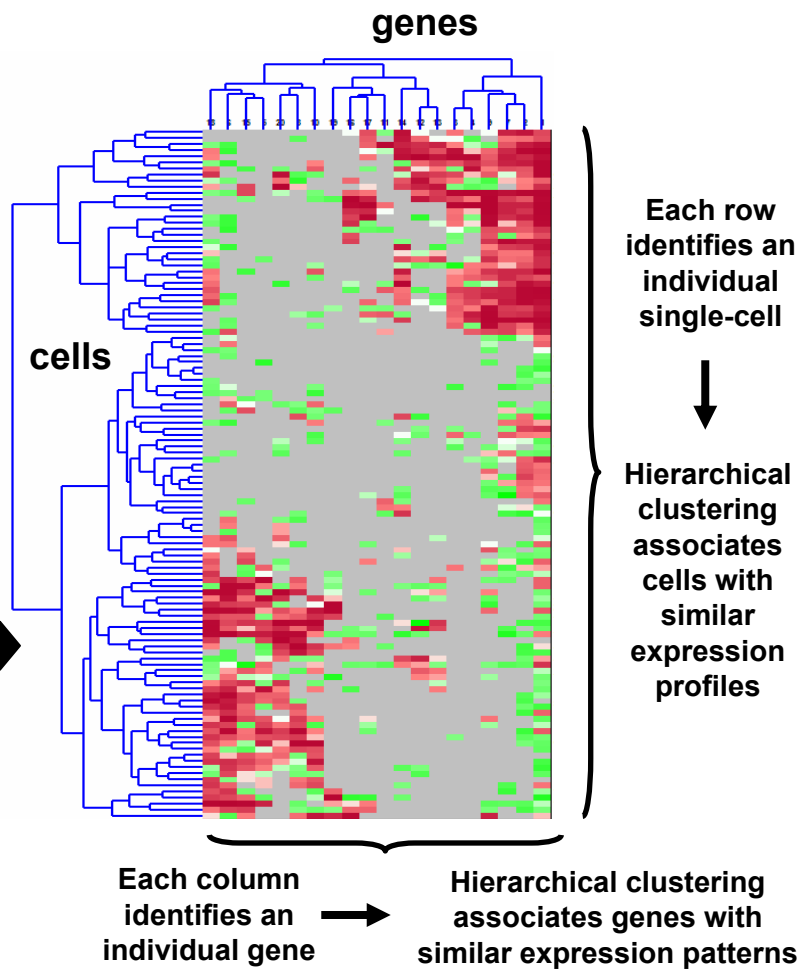
Analysis and graphic display of SINCE-PCR data



1) For each of the 9,216 (96x96) SINCE-PCR reactions, an amplification curve is generated and an individual threshold cycle (Ct) value is calculated.



2) Ct values are normalized and color-coded. Normalized Ct values (Ct_{norm}) are obtained by subtracting from the raw Ct value (Ct) the mean Ct value for the same gene on the whole sample (Ct_{mean}) and then by dividing by three times the standard deviation of the same gene's Ct values distribution ($3SD_{Ct}$). Results are color-coded using increasingly darker shades of red for high expression values ($Ct < Ct_{mean}$), increasingly darker shades of green for low expression values ($Ct > Ct_{mean}$) and grey for lack of expression ($Ct > 40$).



3) Gene-expression results are plotted using hierarchical clustering algorithms available in the MATLAB software (MathWorks Inc.). Hierarchical clustering is performed on both cells and genes, to visualize simultaneously cells with similar expression patterns and genes with similar expression profiles.

Supplementary Figure 2. **Workflow schematic of the method applied for analysis and graphic display of SINCE-PCR data.**

Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.,* **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Accuracy and precision of single-cell sorting by flow cytometry.

A

B

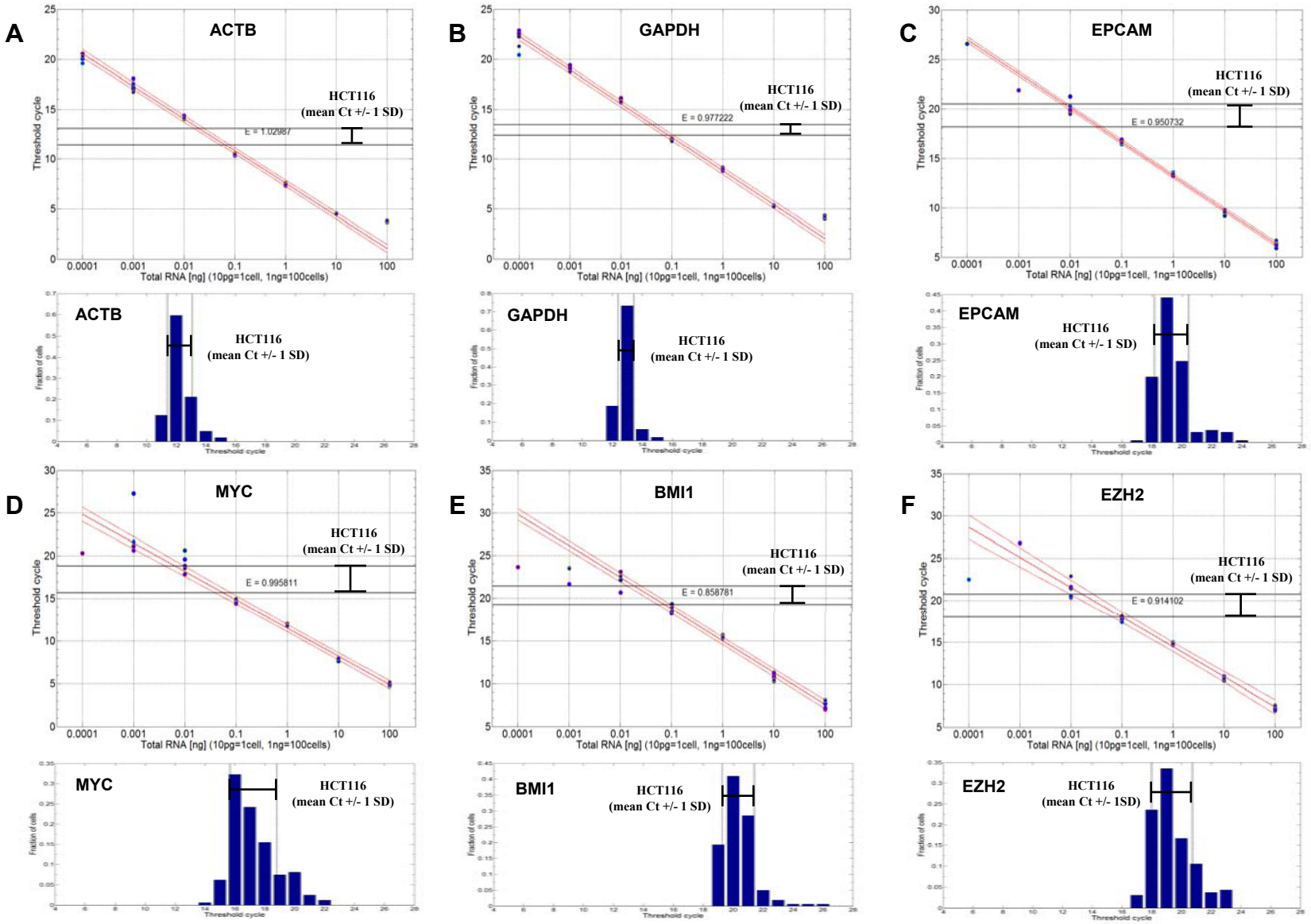
C

n. of cells/well	n. of wells	% of wells
0	8	4.5%
1	171	95.0%
>1	1	0.5%
Total	180	100%

Supplementary Figure 3. **Accuracy and precision of single-cell sorting by flow cytometry.** Accuracy and precision of single-cell sorting by flow cytometry were measured by sorting single-cells (n = 1) into individual microwells of three independent Terasaki microplates (total = 180 microwells). The design and small volume of Terasaki microwells allows direct visualization and counting of sorted cells by optical microscopy (A). To increase both the sensitivity and specificity of the assay, we sorted single-cells from a clone of the HCT116 human colon cancer cell line infected with a lentivirus encoding for the enhanced green fluorescent protein (EGFP; B). Results indicated that 95% of wells (n = 171) contained a single-cell, while 4.5% (n = 8) contained no cells and only 1 (0.5%) contained a doublet, therefore confirming that single-cell sorting by flow cytometry is both accurate and precise (C).

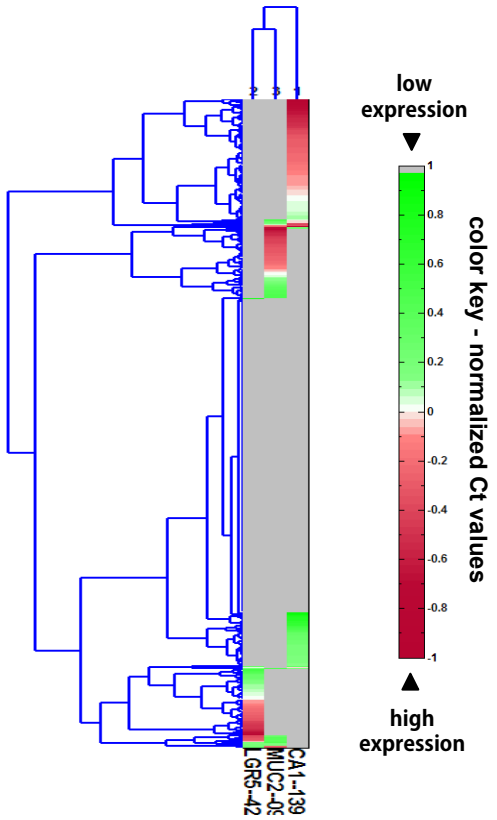
Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.,* **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Supplementary Figure 4



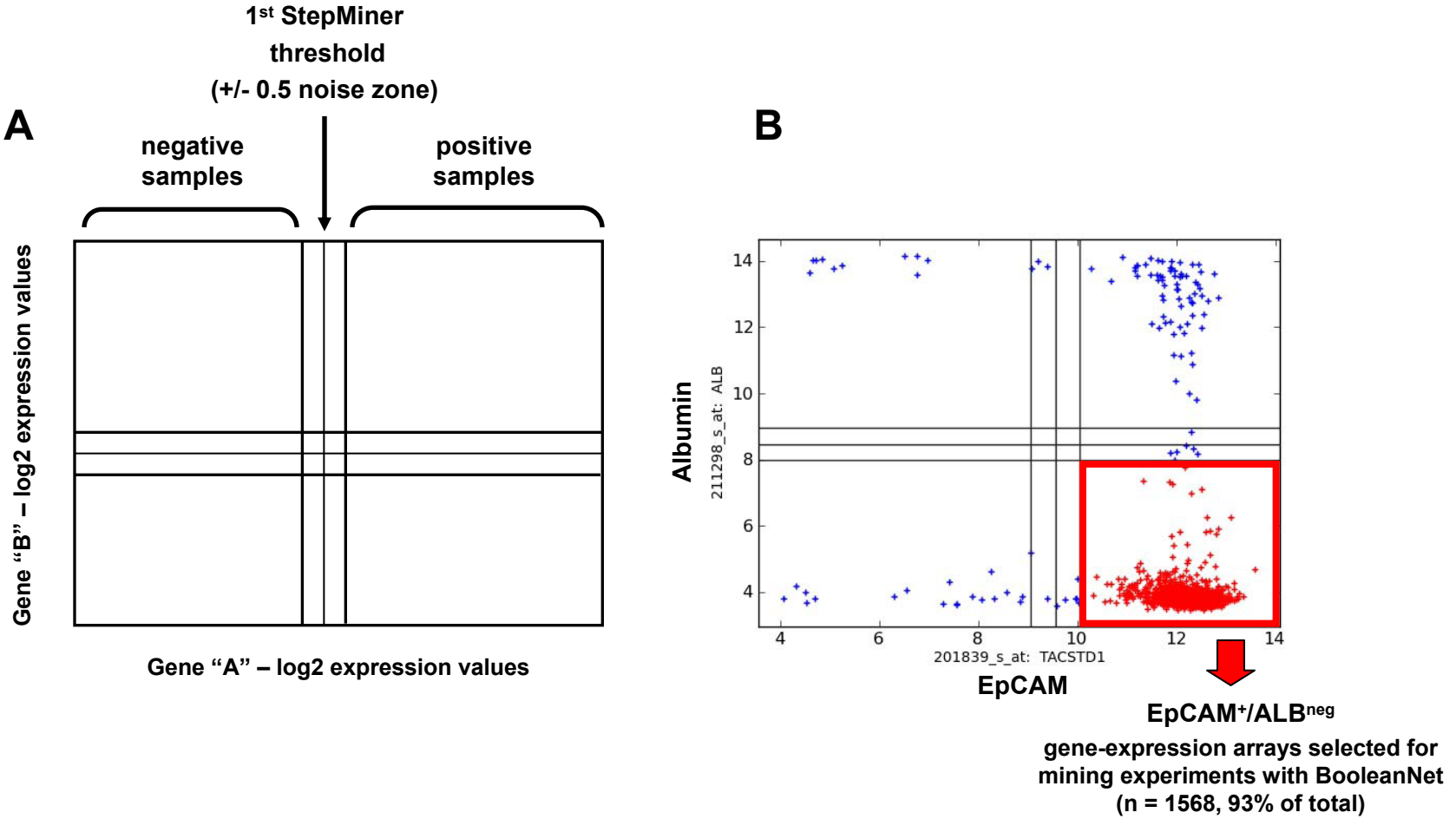
Supplementary Figure 4. **Measure of SINCE-PCR sensitivity.** SINCE-PCR sensitivity was measured on purified RNA standards, using a mixture of mRNA from normal human colon (Applied Biosystems # AM7986), normal human testis (Applied Biosystems # AM7972) and HeLa cells (Applied Biosystems # AM7852) in a 1:1:1 ratio to ensure for a wide repertoire of target mRNAs. Titration curves on 10-fold dilutions of RNA standard confirmed that SINCE-PCR is able to amplify multiple target mRNAs (A-F) on a wide dynamic range (100 ng – 0.001 ng total mRNA) and with high precision (red curves: mean Ct value +/- 1SD). A parallel analysis on single-cells from the HCT116 human colon cancer cell line (n = 168 independent single-cells, blue histograms) indicated that the average amount of target mRNA per cell is within SINCE-PCR's linear range of analysis across multiple genes, including housekeeping genes (A, ACTB; B, GAPDH), epithelial-lineage genes (C, EpCAM), oncogenes (D, MYC) and genes involved in stem cell self-renewal (E, BMI1, EZH2), as visualized by comparing the distribution of Ct values obtained on HCT116 single-cells to the dynamic range of SINCE-PCR on mRNA standards (horizontal black lines: mean Ct value +/- 1SD in HCT116 cells).

Supplementary Figure 5



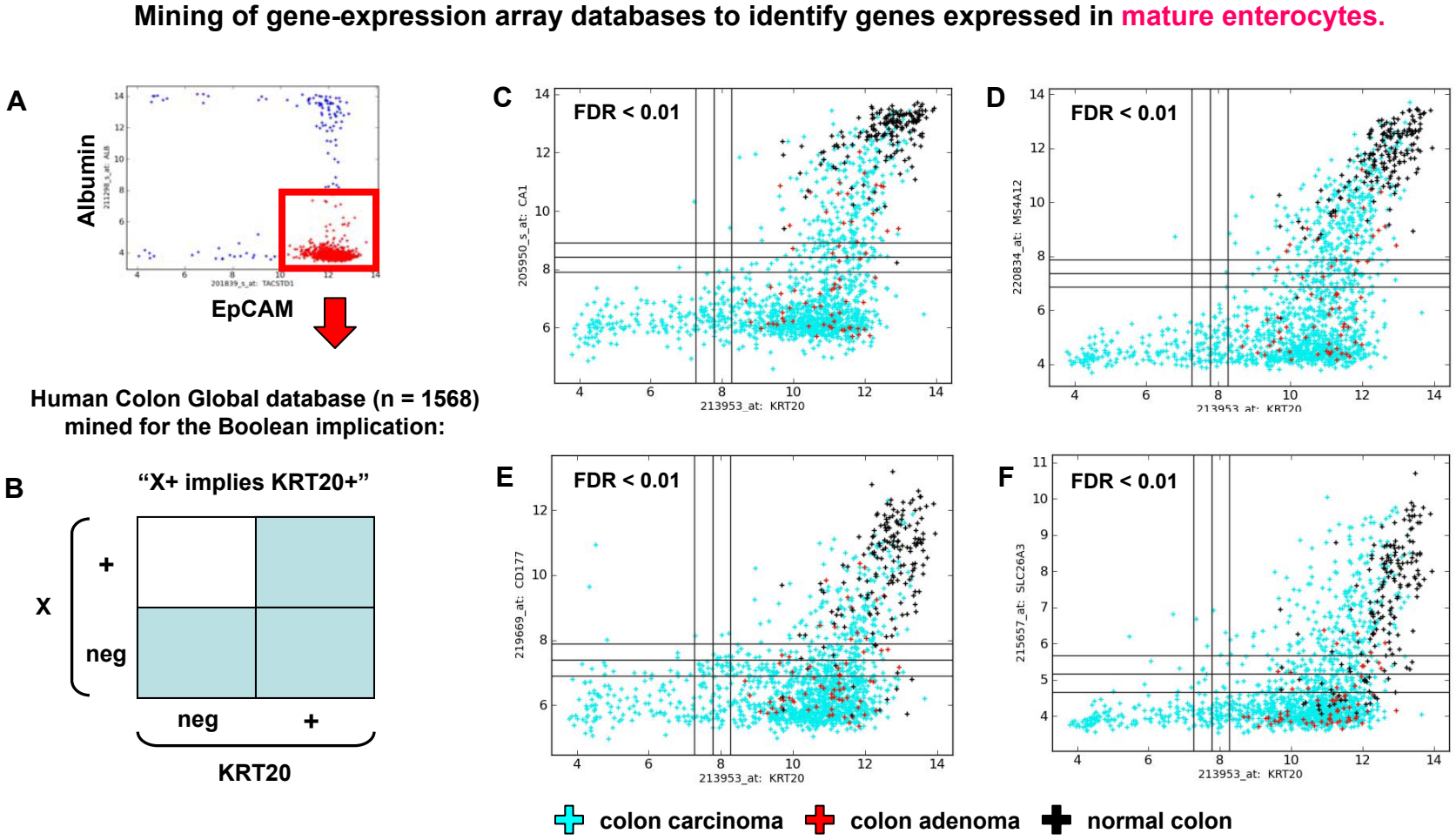
Supplementary Figure 5: **Pilot SINCE-PCR experiment using markers known to be specific to individual cell lineages of the human colonic epithelium.** In our first pilot experiments, we tested the method's feasibility using well established reference markers. We analyzed and clustered colon epithelial cells using three genes encoding for markers known to be exclusive to either one of the two major cell lineages (i.e. MUC2 for goblet cells and CA1 for enterocytes) or the immature compartment (i.e. LGR5) of the colon epithelium. This experiment showed that genes encoding for lineage-specific markers are frequently expressed in a mutually exclusive way, mirroring the expression pattern of corresponding proteins. Moreover, it suggests the existence of additional, uncharacterized populations, negative for the expression of CA1, MUC2 or LGR5 (i.e. $CA1^{neg}$, $MUC2^{neg}$, $LGR5^{neg}$).

Supplementary Figure 6



Supplementary Figure 6: **Definition of the human colon gene-expression array database to be used for mining experiments using Boolean implications.** The search for genes differentially expressed in the human colon epithelium was performed on a "*human colon global database*", obtained by pooling 1684 publicly available gene-expression arrays (Supplementary Table 1). To minimize the risk that results might be affected by poor quality samples or, in the case of hepatic metastases, by samples contaminated with significant amounts of normal liver tissue, bioinformatic analysis was restricted to the subset of arrays whose gene-expression profile could be defined as $EpCAM^{+}/Albumin^{neg}$. EpCAM (TACSTD1) was chosen as a positive marker for the presence of colon epithelial cells, Albumin (ALB) was chosen as a positive marker for the presence of hepatocytes. Gene-expression levels were assigned for each gene in each array, using the log2 of the expression values. The thresholds for definition of positive and negative samples were calculated using the StepMiner algorithm and an intermediate region was defined around each threshold with a width of 1 (i.e. threshold +/- 0.5), corresponding to a 2-fold change in expression, which is the minimum noise level in these type of datasets (Sahoo *et al*., Genome Biology, 9:R157, 2008). All the data below the intermediate region (< 1st StepMiner threshold - 0.5) were considered negative, and all above the intermediate region (> 1st StepMiner threshold + 0.5) were considered positive (A). Based on these rules, $EpCAM^{+}$ samples were defined as Affymetrix probe 201839_s_at >10.05, and ALB^{neg} samples were defined as Affymetrix probe 211298_s_at <7.97). The "purging" operation removed 116 arrays (7%) and left 1568 arrays (93%) for subsequent analysis (B).

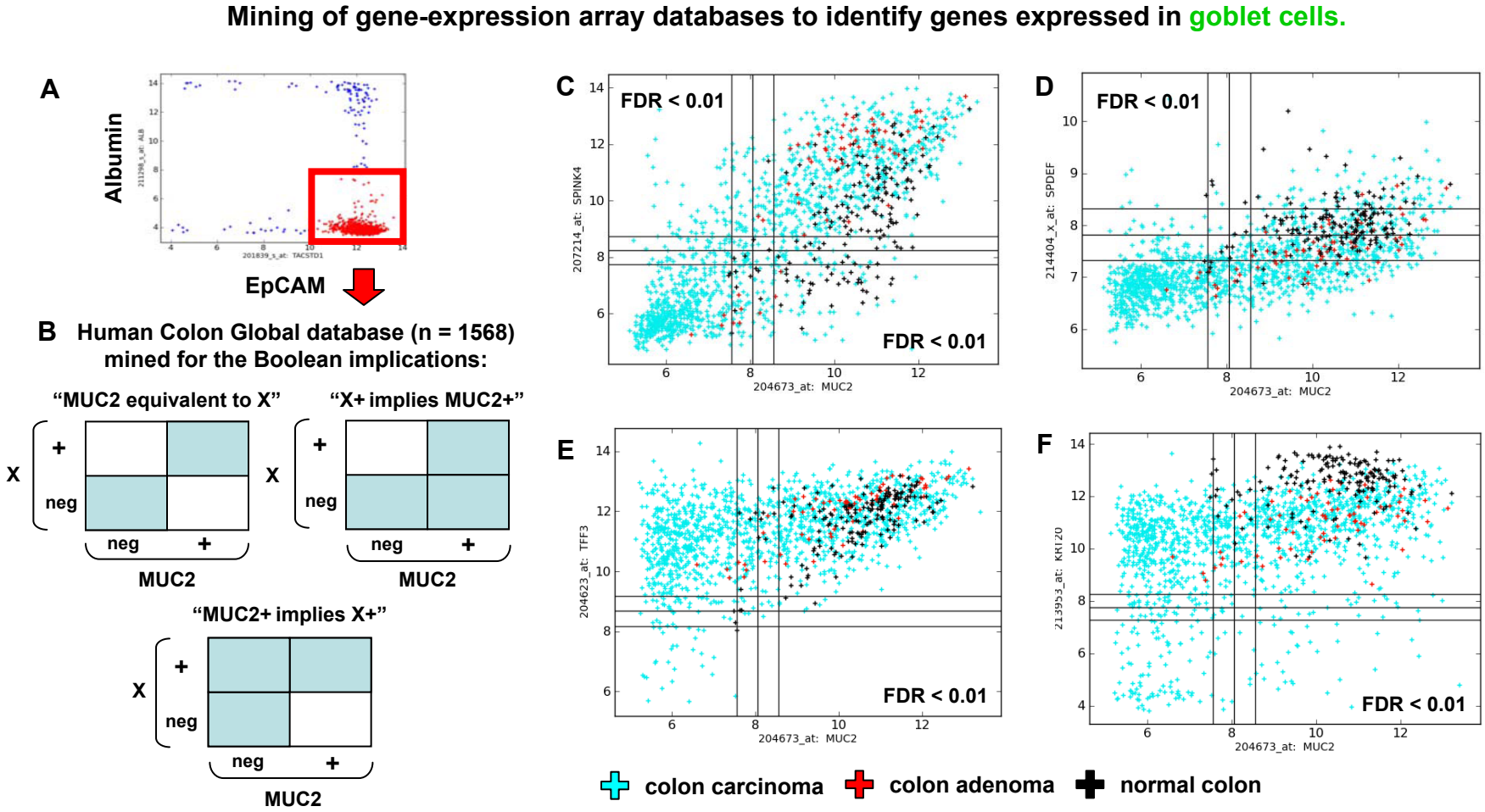
Supplementary Figure 7



Supplementary Figure 7. **Mining of publicly available human colon gene-expression array databases using a "Boolean implications" method (BooleanNet software): identification of genes expressed in mature enterocytes.** Using a software algorithm designed to identify pairs of genes whose expression is regulated by Boolean implications across multiple microarray datasets (BooleanNet; Sahoo *et al.*, Genome Biology, 9:R157, 2008), we performed a high-throughput screening on a pooled database of human colon gene-expression arrays, after selection for $EpCAM^{+}/ALB^{neg}$ samples (A, see also Supplementary Fig. 6). This database is composed of 1,568 samples, and includes 170 arrays from normal colon epithelium (black crosses), 68 arrays from colorectal adenomas (red crosses) and 1,330 arrays from colorectal carcinomas (blue crosses). The mining strategy aimed at the discovery of genes selectively expressed in mature enterocytes (B) was based on the fulfillment of the "X^{+} implies $KRT20^{+}$" Boolean implication (i.e. identification of genes selectively expressed in $KRT20^{+}$ samples). Threshold gene expression levels were calculated using the StepMiner algorithm, based on our total pool of 46,047 publicly available human gene-expression arrays. Gene-expression patterns were considered to fulfill the Boolean implication "X^{+} implies $KRT20^{+}$" when the false-discovery rate (FDR) of a sparsity test in the upper left quadrant was < 0.05. Among the genes fulfilling this Boolean implication were: CA1 (C), MS4A12 (D), CD177 (E) and SLC26A3 (F). Gene-expression levels were assigned for each gene in each array, using the log2 of the expression values.

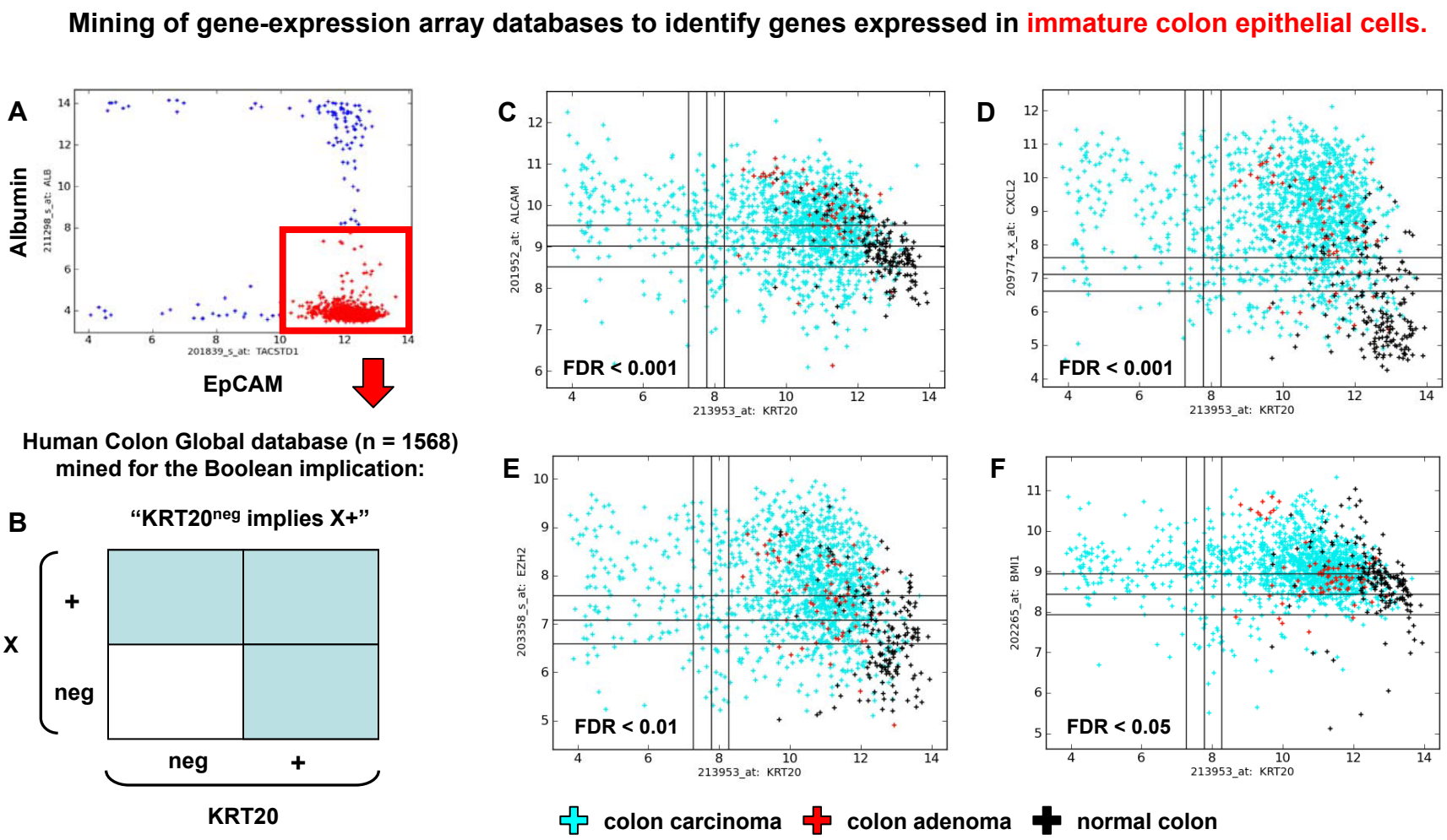
Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.*, **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Supplementary Figure 8



Supplementary Figure 8. **Mining of publicly available human colon gene-expression array databases using a "Boolean implications" method (BooleanNet software): identification of genes expressed in goblet cells.** Using a software algorithm designed to identify pairs of genes whose expression is regulated by Boolean implications across multiple microarray datasets (BooleanNet, Sahoo *et al*., Genome Biology, 9:R157, 2008), we performed a high-throughput screening on a pooled database of human colon gene-expression arrays, after selection for EpCAM^{+}/ALB^{neg} samples (A, see also Supplementary Fig. 6). This database is composed of 1,568 samples, and includes 170 arrays from normal colon epithelium (black crosses), 68 arrays from colorectal adenomas (red crosses) and 1,330 arrays from colorectal carcinomas (blue crosses). The mining strategy aimed at the discovery of genes expressed in goblet cells (B) was based on two sets of Boolean implications: a) "MUC2 is equivalent to X" or "X^{+} implies MUC2^{+}" (i.e. identification of genes selectively expressed in MUC2^{+} samples); b) "MUC2^{+} implies X^{+}" (i.e. identification of genes always expressed in MUC2^{+} samples). Threshold gene expression levels were calculated using the StepMiner algorithm, based on our total pool of 46,047 publicly available human gene-expression arrays. Gene-expression patterns were considered to fulfill the Boolean implications "MUC2 is equivalent to X" or "X^{+} implies MUC2^{+}" when the false-discovery rate (FDR) of a sparsity test in both the upper left and the lower right quadrant or in the upper left quadrant alone was < 0.05. Among the genes fulfilling these Boolean implications were: SPINK4 (C) and SPDEF (D). Gene-expression patterns were considered to fulfill the Boolean implication "MUC2^{+} implies X^{+}" when the false-discovery rate (FDR) of a sparsity test in the lower right quadrant was < 0.05. Among the genes fulfilling this Boolean implication were: TFF3 (E) and KRT20 (F). Gene-expression levels were assigned for each gene in each array, using the log2 of the expression values.

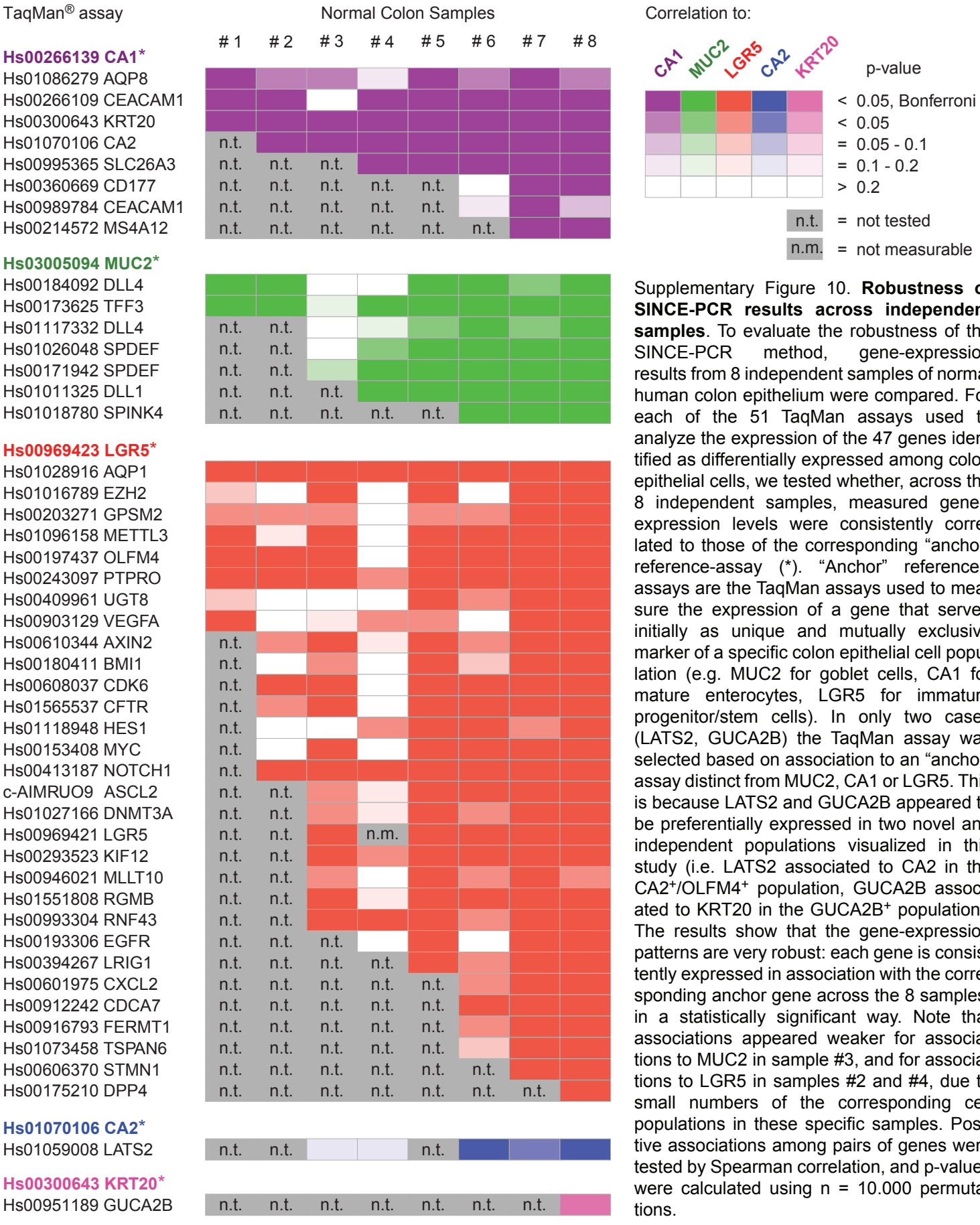
Supplementary Figure 9



Supplementary Figure 9. **Mining of publicly available human colon gene-expression array databases using a "Boolean implications" method (BooleanNet software): identification of genes expressed in immature cell populations.** Using a software algorithm designed to identify pairs of genes whose expression is regulated by Boolean implications across multiple microarray datasets (BooleanNet; Sahoo *et al*., Genome Biology, 9:R157, 2008), we performed a high-throughput screening on a pooled database of human colon gene-expression arrays, after selection for $EpCAM^{+}/ALB^{neg}$ samples (A, see also Supplementary Fig. 6). This database is composed of 1,568 samples, and includes 170 arrays from normal colon epithelium (black crosses), 68 arrays from colorectal adenomas (red crosses) and 1,330 arrays from colorectal carcinomas (blue crosses). The mining strategy aimed at the discovery of genes always expressed in immature colon epithelial cells (B) was based on the fulfillment of the "$KRT20^{neg}$ implies X^{+}" Boolean implication (i.e. identification of genes always expressed in $KRT20^{neg}$ samples). Threshold gene expression levels were calculated using the StepMiner algorithm, based on our total pool of 46,047 publicly available human gene-expression arrays. Gene-expression patterns were considered to fulfill the Boolean implication "$KRT20^{neg}$ implies X^{+}" when the false-discovery rate (FDR) of a sparsity test in the lower left quadrant was < 0.05. Among the genes fulfilling this Boolean implication were: ALCAM/CD166 (C), CXCL2 (D), EZH2 (E) and BMI1 (F). Gene-expression levels were assigned for each gene in each array, using the log2 of the expression values.

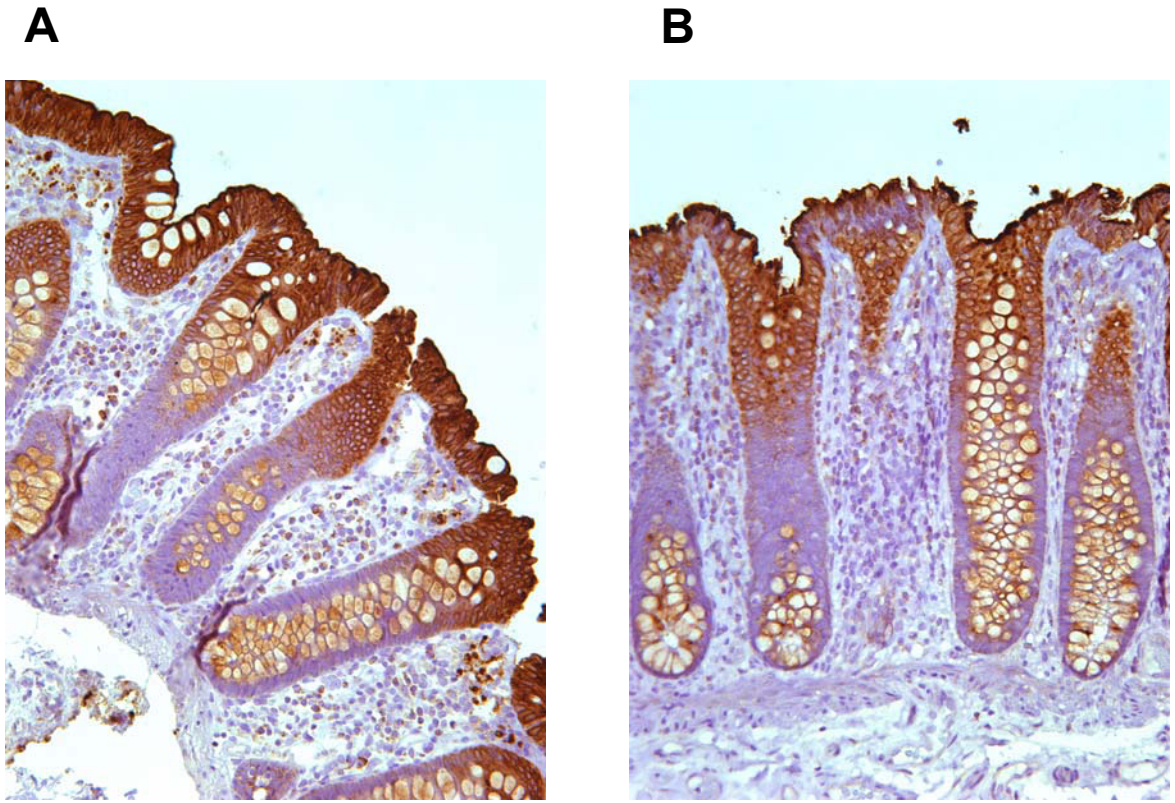
Supplemementary Figure 10

Robustness of SINCE-PCR results across independent samples



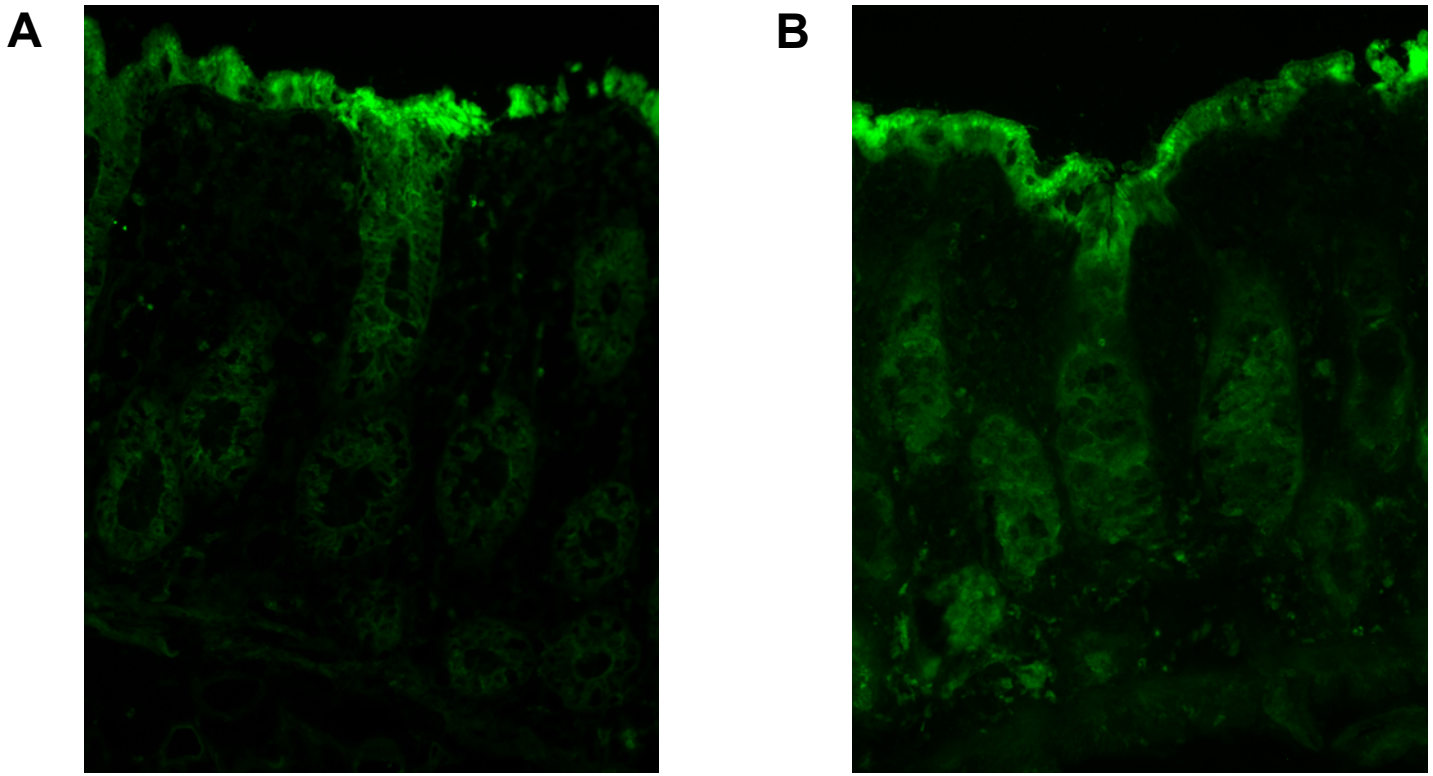
Supplementary Figure 10. **Robustness of SINCE-PCR results across independent samples**. To evaluate the robustness of the SINCE-PCR method, gene-expression results from 8 independent samples of normal human colon epithelium were compared. For each of the 51 TaqMan assays used to analyze the expression of the 47 genes identified as differentially expressed among colon epithelial cells, we tested whether, across the 8 independent samples, measured gene-expression levels were consistently correlated to those of the corresponding "anchor" reference-assay (*). "Anchor" reference-assays are the TaqMan assays used to measure the expression of a gene that served initially as unique and mutually exclusive marker of a specific colon epithelial cell population (e.g. MUC2 for goblet cells, CA1 for mature enterocytes, LGR5 for immature progenitor/stem cells). In only two cases (LATS2, GUCA2B) the TaqMan assay was selected based on association to an "anchor" assay distinct from MUC2, CA1 or LGR5. This is because LATS2 and GUCA2B appeared to be preferentially expressed in two novel and independent populations visualized in this study (i.e. LATS2 associated to CA2 in the $CA2^{+}$/$OLFM4^{+}$ population, GUCA2B associated to KRT20 in the $GUCA2B^{+}$ population). The results show that the gene-expression patterns are very robust: each gene is consistently expressed in association with the corresponding anchor gene across the 8 samples, in a statistically significant way. Note that associations appeared weaker for associations to MUC2 in sample #3, and for associations to LGR5 in samples #2 and #4, due to small numbers of the corresponding cell populations in these specific samples. Positive associations among pairs of genes were tested by Spearman correlation, and p-values were calculated using n = 10.000 permutations.

Analysis of SLC26A3 protein expression in human normal colon tissues.



Supplementary Figure 11. **SLC26A3 protein expression in human normal colon epithelium.** SINCE-PCR analysis of the human normal colorectal epithelium identified SLC26A3 as a gene preferentially expressed within the EpCAM^{+}/CD44^{neg}/CD66a^{high} ("top of the crypt") population (Fig. 1). Analysis by immunohistochemistry of SLC26A3 protein expression in tissue sections of human normal colorectal mucosa confirmed SINCE-PCR data, showing dramatic increase of SLC26A3 protein levels in the upper third of colon crypts, in two different patients (A-B). Immunohistochemistry was performed using a polyclonal affinity-purified rabbit anti-human SLC26A3 antibody preparation (Sigma Life Science – Atlas Antibodies; Lot #R32905).

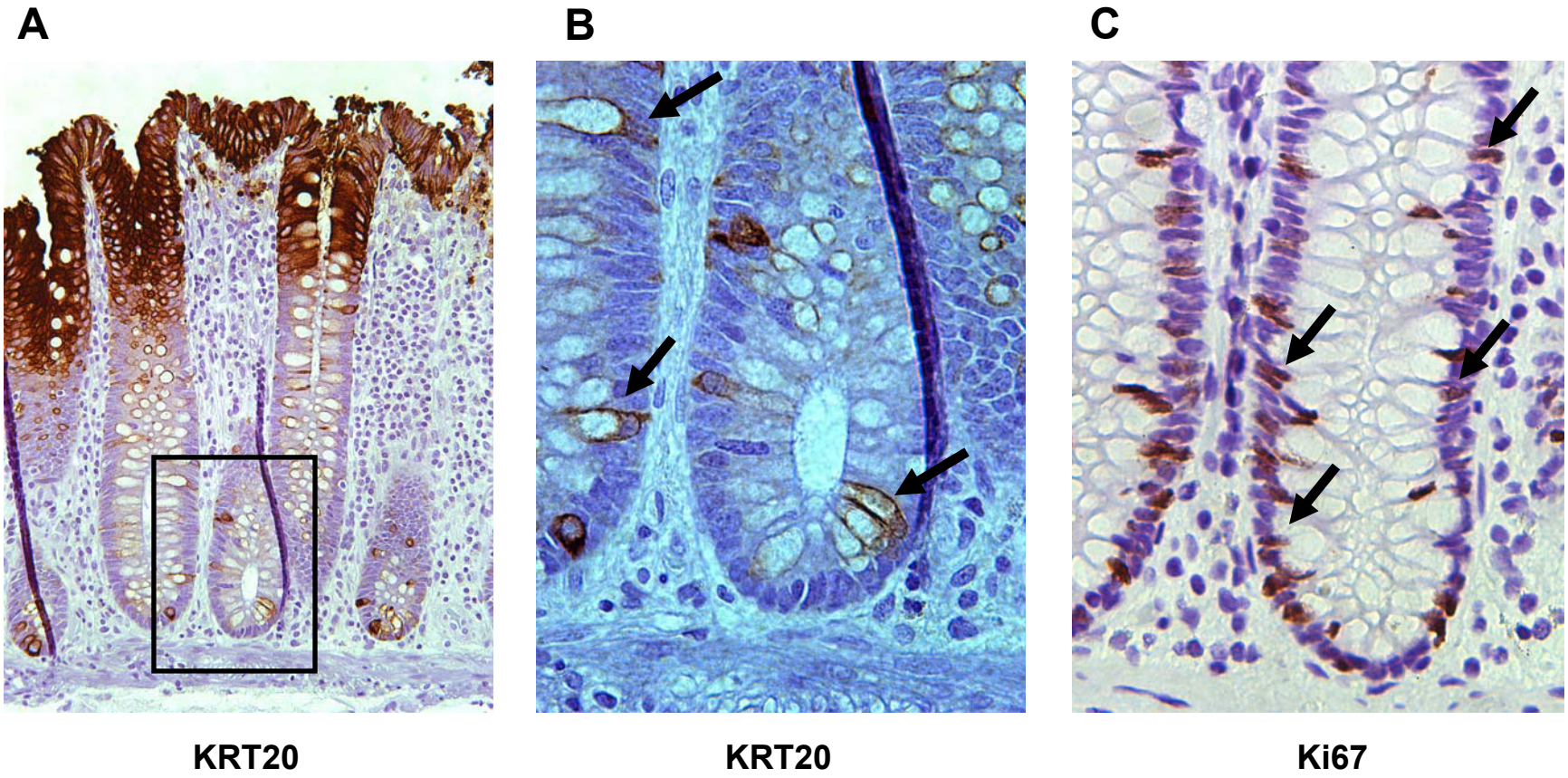
Analysis of CD177 protein expression in human normal colon tissues.



Supplementary Figure 12. **CD177 protein expression in human normal colonic epithelium.** SINCE-PCR analysis of the human normal colorectal epithelium identified CD177 as a gene preferentially expressed within the $EpCAM^{+}/CD44^{neg}/CD66a^{high}$ ("top of the crypt") population (Fig. 1). Analysis by immunofluorescence of CD177 protein expression in the human normal colorectal mucosa confirmed SINCE-PCR data, showing a dramatic increase of CD177 protein levels in the upper third of colon crypts, in two different patients (A-B). Immunofluorescence was performed on frozen tissue sections, using a mouse anti-human CD177 monoclonal antibody (clone MEM-166, BD Biosciences).

Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.,* **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

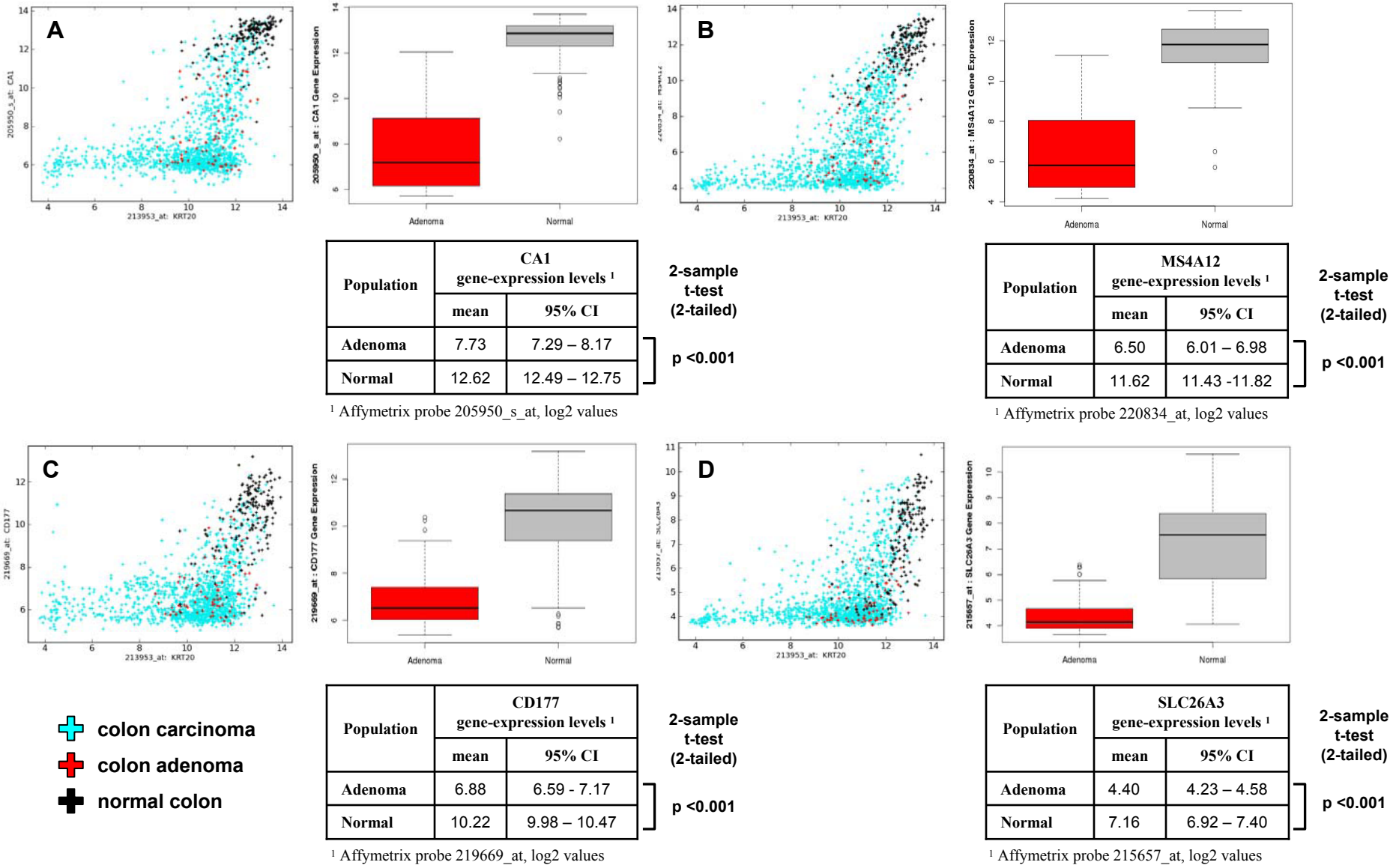
Supplementary Figure 13



Supplementary Figure 13. **KRT20 and Ki67 expression in goblet cells of the human normal colon epithelium.** SINCE-PCR analysis of cells with goblet-like transcriptional profiles (i.e. MUC2^{+}, TFF3^{high}, SPDEF^{+}, SPINK4^{+}) within the EpCAM^{+}/CD44^{+} ("bottom of the crypt") population revealed frequent expression of KRT20 mRNA (Fig. 1). This observation, at first, appeared contrary to the notion of KRT20 as a terminal differentiation marker. However, upon more careful examination of human colon tissue sections analyzed by immunohistochemistry with anti-KRT20 antibodies, we were able to identify scattered KRT20^{+} cells throughout the full length of the human colonic crypts (A). On close look, these KRT20^{+} cells could be morphologically identified as a subset of goblet cells (B, arrows). As indicated by SINCE-PCR data, a subset of goblet cells also expresses proliferation markers, such as Ki67 (C, arrows).

Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.,* **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Down-regulation in human benign colorectal adenomas of genes preferentially expressed in CA1^{+}/SLC26A3^{+} cells.



Population	CA1 gene-expression levels [1]		2-sample t-test (2-tailed)
	mean	95% CI	
Adenoma	7.73	7.29 – 8.17	p <0.001
Normal	12.62	12.49 – 12.75	

[1] Affymetrix probe 205950_s_at, log2 values

Population	MS4A12 gene-expression levels [1]		2-sample t-test (2-tailed)
	mean	95% CI	
Adenoma	6.50	6.01 – 6.98	p <0.001
Normal	11.62	11.43 -11.82	

[1] Affymetrix probe 220834_at, log2 values

Population	CD177 gene-expression levels [1]		2-sample t-test (2-tailed)
	mean	95% CI	
Adenoma	6.88	6.59 - 7.17	p <0.001
Normal	10.22	9.98 – 10.47	

[1] Affymetrix probe 219669_at, log2 values

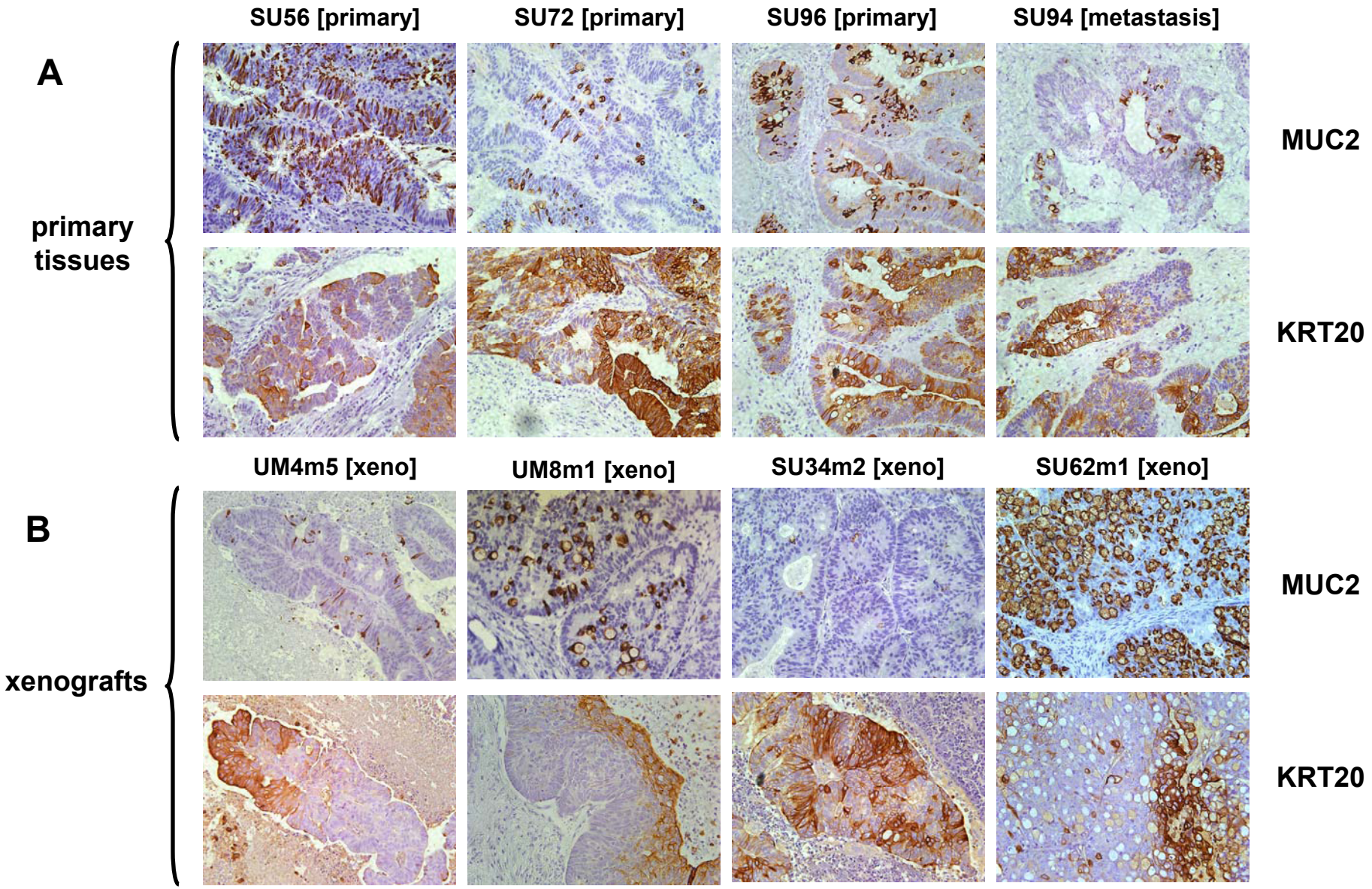
Population	SLC26A3 gene-expression levels [1]		2-sample t-test (2-tailed)
	mean	95% CI	
Adenoma	4.40	4.23 – 4.58	p <0.001
Normal	7.16	6.92 – 7.40	

[1] Affymetrix probe 215657_at, log2 values

Supplementary Figure 14. **Genes preferentially expressed in "top-of-the-crypt" CA1^{+}/SLC26A3^{+} enterocyte-type cells are down-regulated in colon adenomas.** A systematic comparison of gene-expression array results between normal colon epithelium and human colorectal adenomas indicates that expression levels of genes preferentially expressed by "top-of-the-crypt" CA1^{+}/SLC26A3^{+} enterocyte-like cells (i,.e. CA1, A; MS4A12, B; CD177, C; SLC26A3, D) are down-regulated in colorectal adenomas, in a statistically significant way ($p < 0.001$).

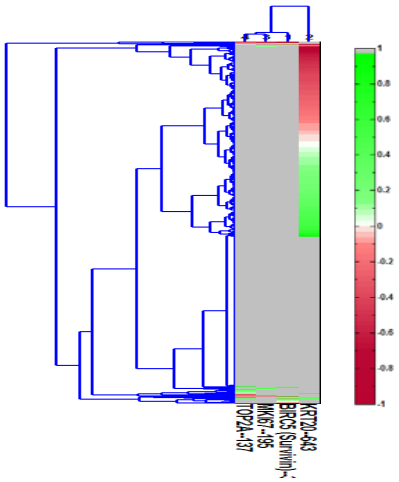
Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.,* **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Supplementary Figure 15



Supplementary Figure 15. **Histopathological analysis of differentiation markers (KRT20, MUC2) in human colorectal cancer tissues.** A systematic study of KRT20 and MUC2 protein expression in human colorectal cancer tissues reveals that both markers are frequently expressed heterogeneously, in patterns that mirror those observed in normal colorectal epithelium (MUC2 in mucus-secreting goblet cells, KRT20 in clusters of enterocyte-like cells and selected goblet cells). The percentage of both $KRT20^+$ and $MUC2^+$ cells is very variable from patient to patient and, in selected tumors, it can be lost almost completely, together with the corresponding cellular lineages and differentiation stages. Similar patterns can be observed both in primary tissues (A) and in solid tissue xenografts established in immunodeficient mice (B).

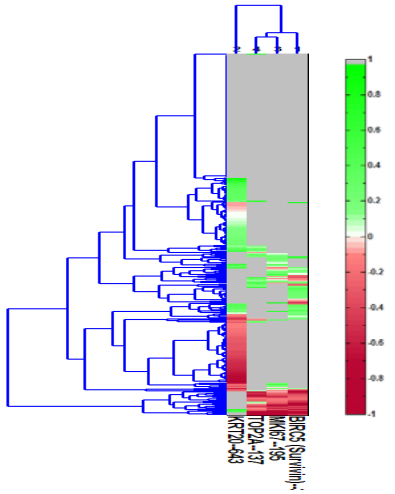
Human normal colon epithelium



Gene 1	Gene 2	Correlation coefficient (Spearman)	p-value (n=10,000)
MKI67	KRT20--643	- 0.07	0.018
TOP2A	"	- 0.14	< 0.001**
BIRC5/Survivin	"	- 0.08	0.006*

* = p < 0.01
** = p < 0.001

Colon Cancer xenograft (UM-COLON4, clone#8)

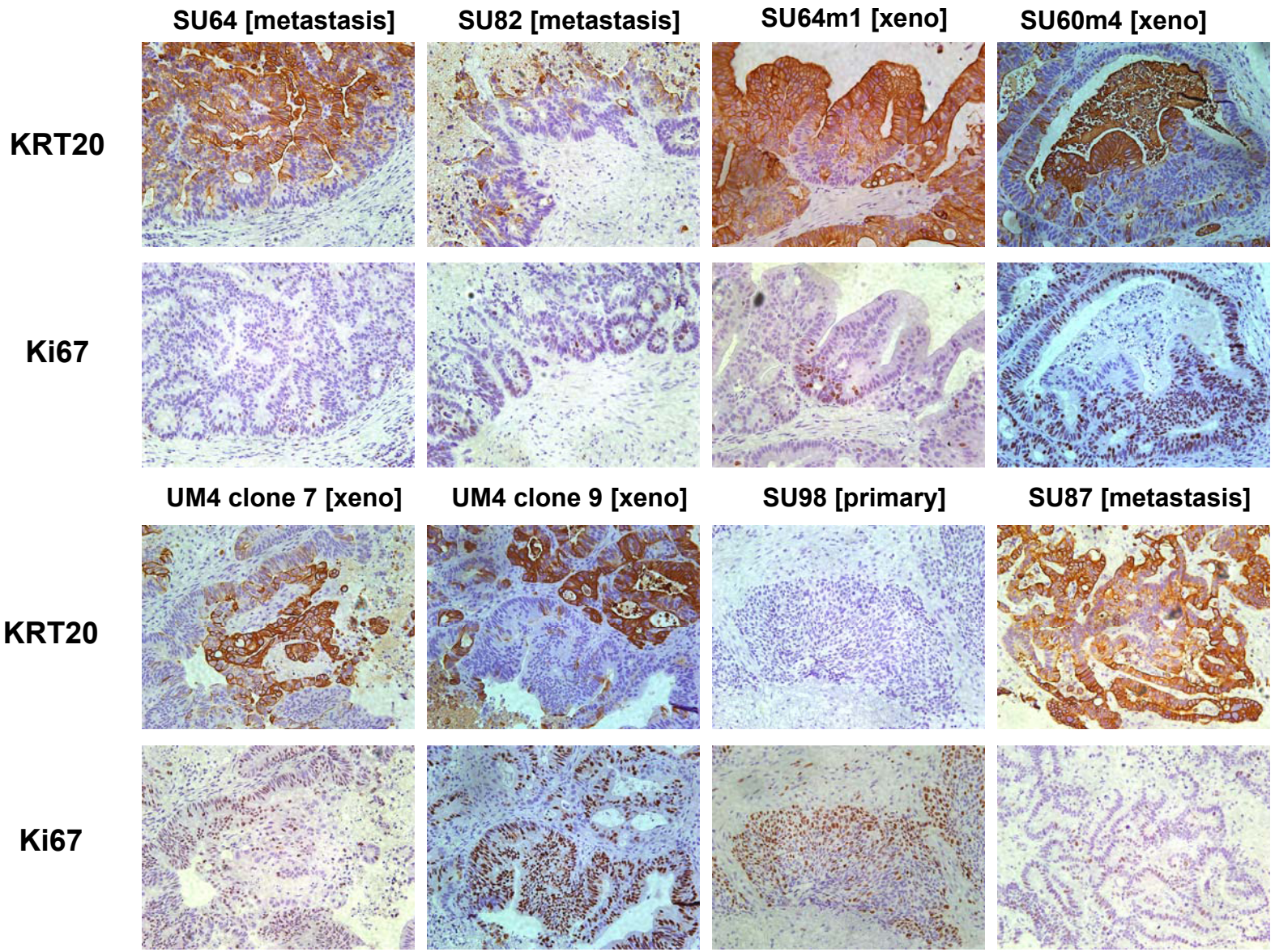


Gene 1	Gene 2	Correlation coefficient (Spearman)	p-value (n=10,000)
MKI67	KRT20--643	- 0.19	< 0.001**
TOP2A	"	- 0.17	< 0.001**
BIRC5/Survivin	"	- 0.21	< 0.001**

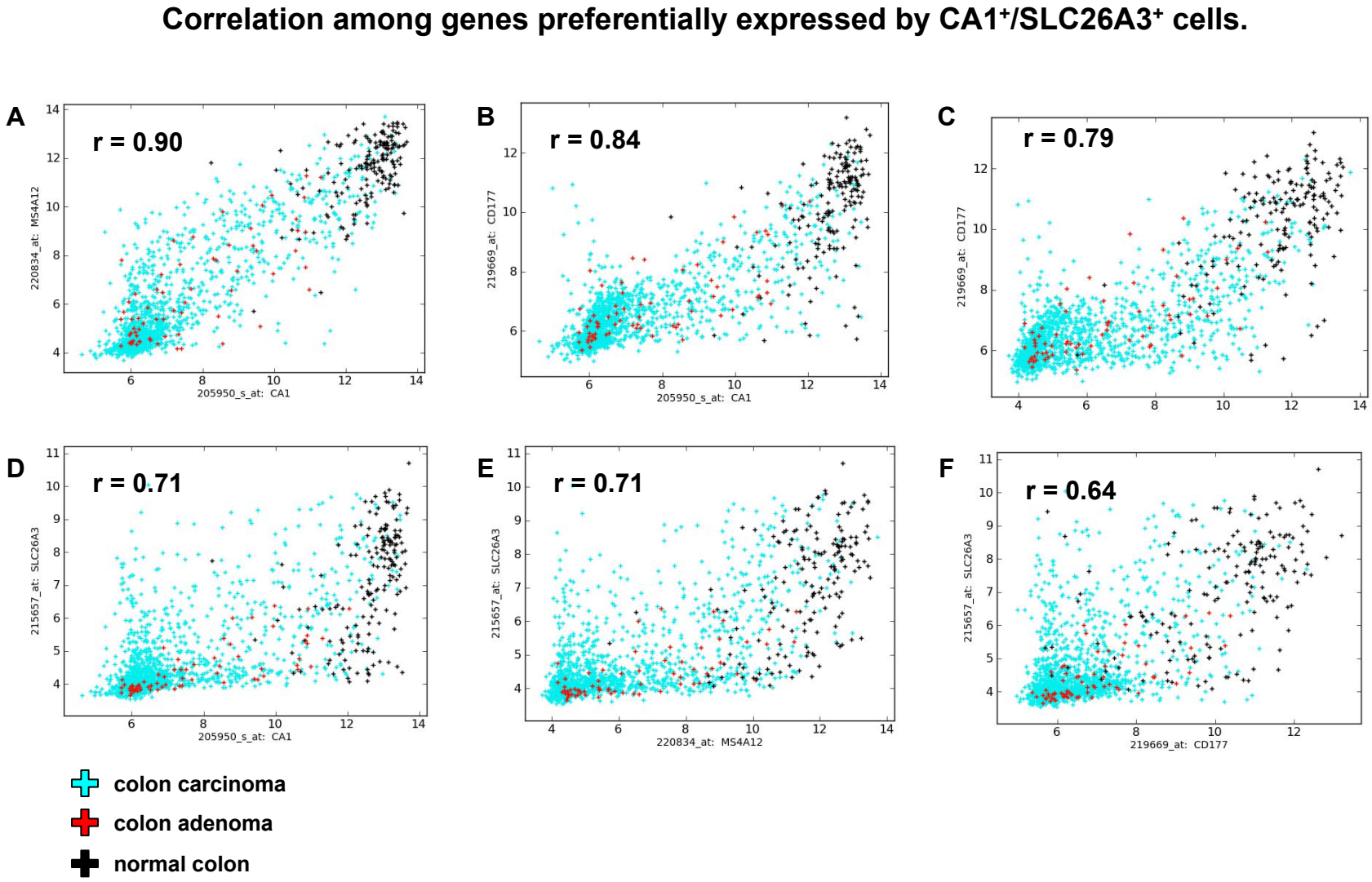
* = p < 0.01
** = p < 0.001

Supplementary Figure 16: **The gene expression levels of cell proliferation markers (MKI67, TOP2A, BIRC5/Survivin) are inversely correlated to those of the differentiation marker KRT20 in both human normal colon epithelium and human colorectal cancer tissues.** A correlation analysis of the gene-expression levels of the proliferation markers MKI67, TOP2A and BIRC5/Survivin in single-cells reveals that the expression of these genes is inversely associated with that of the differentiation marker KRT20, in both human normal colon epithelium and human colorectal cancer xenografts (UM-COLON4 clone#8). Statistically significant correlations were assessed by a Spearman correlation test, and p-values were calculated using n = 10.000 permutations**.**

Supplementary Figure 17

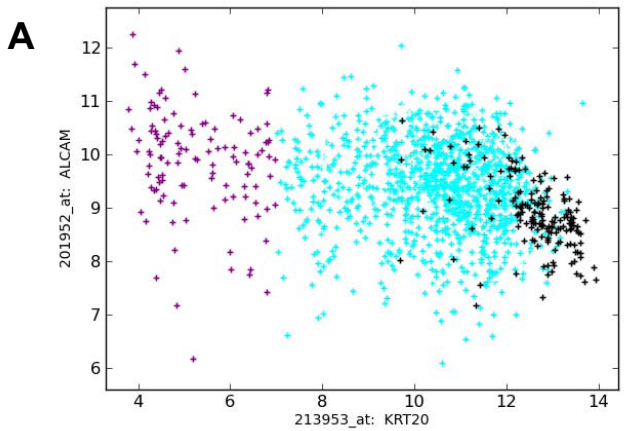


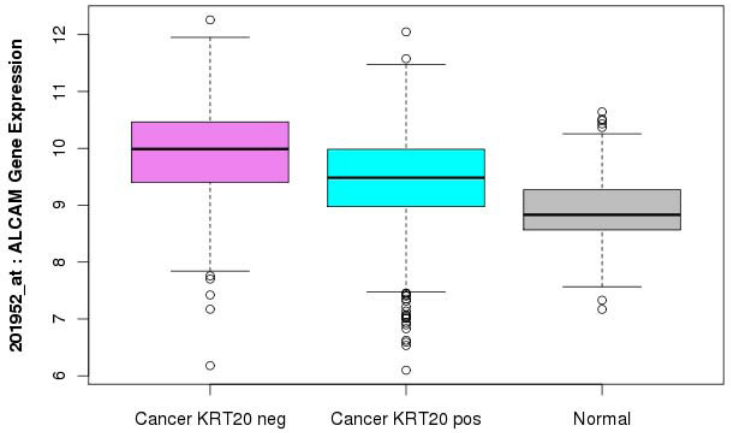
Supplementary Figure 17: **Histopathological analysis of KRT20 and Ki67 expression in human colorectal cancer tissues.** A systematic study of KRT20 and Ki67 protein expression in human colorectal cancer tissues reveals that KRT20 protein expression is, in many cases, inversely associated with that of Ki67, a known proliferation marker. This feature, however, is not absolute, as some tumors display KRT20 protein expression across the almost entirety of the cancer cell population (e.g. SU87, liver metastasis), while others are characterized by complete absence of it (SU98, primary tumor). Interestingly, tumors characterized by the complete absence of KRT20 expression were very poorly differentiated and contained high percentages of $Ki67^{+}$ cells (SU98, primary tumor).



Supplementary Figure 18. **Correlation of expression levels among genes characteristic of "top-of-the-crypt" CA1$^+$/SLC26A3$^+$ enterocyte-type cells.** A systematic analysis of the expression levels of genes preferentially expressed by "top-of-the-crypt" CA1$^+$/SLC26A3$^+$ enterocyte-type cells (i,.e. CA1, MS4A12, CD177, SLC26A3) indicates that they are all correlated between each-other, in normal colon samples, as well as in colorectal adenomas and carcinomas (r = Pearson correlation coefficient). Correlation values among CA1, MS4A12 and CD177 (A-C) appear stronger than those between each of those genes and SLC26A3 (D-F), probably due to a lower sensitivity of the SLC26A3 probe (see also Supplementary Methods).

Higher gene-expression levels of ALCAM/CD166 in KRT20^{neg} human colon carcinomas.





KRT20^{neg} colon carcinoma

KRT20^{pos} colon carcinoma

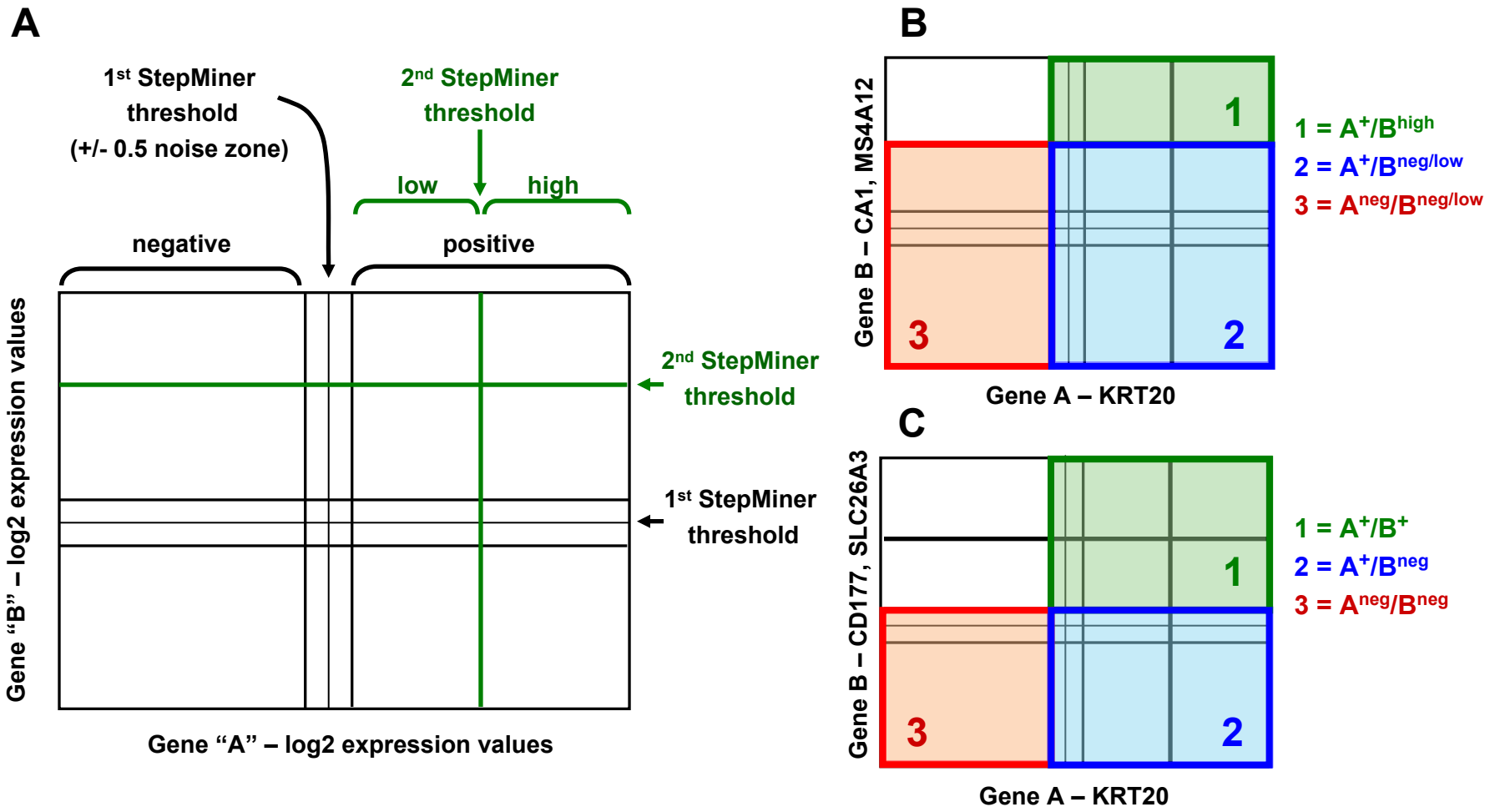
normal colon

C

Population	ALCAM/CD166 gene-expression levels [1]		2-sample t-test (2-tailed)
	mean	95% CI	
Cancer KRT20^{neg}	9.81	9.63 – 9.99	p <0.001
Cancer KRT20^{pos}	9.43	9.38 – 9.48	
Normal	8.92	8.81 – 9.02	p <0.001

[1] Affymetrix probe 201952_at, log2 of expression values

Supplementary Figure 19. **Comparison of ALCAM/CD166 gene-expression levels between human normal colorectal epithelium and human colorectal carcinomas (KRT20^{neg} vs KRT20^{pos}).** A systematic comparison of gene-expression results from publicly available human gene-expression arrays indicates that ALCAM/CD166 gene-expression levels are higher in KRT20^{neg} colorectal carcinomas as compared to KRT20^{pos} ones and to normal colorectal epithelium. The visual suggestion provided by the scatter-plot (A) is confirmed by box-plots (B). A 2-sample t-test to compare mean ALCAM/CD166 gene-expression levels in the three populations indicates that differences are statistically significant (C).



Supplementary Figure 20: **Definition of "gene-expression groups" for patient survival analysis.** To explore the possible correlations between gene-expression profiles and patient survival in human colon cancer, we stratified human colon cancer samples in different "gene-expression" groups using the StepMiner algorithm (Sahoo *et al.*, Genome Biology, 9:R157, 2008). In this case, we used the StepMiner algorithm to calculate two distinct thresholds: a 1st StepMiner threshold, to discriminate between "negative" and "positive" samples, and a 2nd StepMiner threshold, to discriminate between "low" and "high" expression samples (A). We stratified human colon cancer samples based on the gene-expression levels of KRT20 and each one of four different genes preferentially expressed by "top-of-the-crypt" $CA1^{+}/SLC26A3^{+}$ enterocyte-like cells (i.e. CA1, MS4A12, CD177, SLC26A3), which are related by a "top-crypt^{+} implies KRT20^{+}" ("B^{+} implies A^{+}") Boolean implication (Supplementary Fig. 7). We stratified human colon cancer samples into three "*gene-expression groups*": Group 1 ($KRT20^{+}$/top-crypt^{high}), Group 2 ($KRT20^{+}$/top-crypt$^{neg/low}$), Group 3 ($KRT20^{neg}$/top-crypt$^{neg/low}$). Given the variable sensitivity of the probes for the four different "top-crypt" genes, in order to maintain consistency in the selection of sample subsets with highest expression levels, we adopted a scaled approach, different for CA1 and MS4A12 (B) as compared to CD177 and SLC26A3 (C). In the case of CA1 and MS4A12 we defined colon tumors as "top-crypt^{high}" when they scored $CA1^{high}$ and $MS4A12^{high}$, respectively (i.e. > 2nd StepMiner threshold, B). In the case of CD177 and SLC26A3, we defined colon tumors as "top-crypt^{high}" when they scored $CD177^{+}$ and $SLC26A3^{+}$, respectively (i.e. > 1st StepMiner threshold + 0.5, C)

Supplementary Figure 21

"*Gene-expression groups*" vs pathological grading (grading database, n = 639)

Pathological Grade	KRT20/CA1			Enrichment of Group 1 in G1-G2 (Group 1 vs Group 2+3)			Enrichment of Group 3 in G3-G4 (Group 3 vs Group 1+2)		
	Group 1 KRT20^{+} CA1^{high}	Group 2 KRT20^{+} CA1$^{neg/low}$	Group 3 KRT20^{neg} CA1$^{neg/low}$	% Tumors Group 1	OR [1] (95%CI[2])	Pearson's χ^2	% Tumors Group 3	OR [1] (95%CI[2])	Pearson's χ^2
G1-G2 (n = 507)	34	433	40	6.7 % (34/507)	3.1 (0.9 – 10.2)	3.8 p = 0.05	7.9 % (40/507)	1	**25.8 p <0.001**
G3-G4 (n = 132)	3	98	31	2.3 % (3/132)	1		23.5 % (31/132)	3.6 (2.1 – 6.0)	

Pathological Grade	KRT20/MS4A12			% Tumors Group 1	OR [1] (95%CI[2])	Pearson's χ^2	% Tumors Group 3	OR [1] (95%CI[2])	Pearson's χ^2
	Group 1 KRT20^{+} MS4A12^{high}	Group 2 KRT20^{+} MS4A12$^{neg/low}$	Group 3 KRT20^{neg} MS4A12$^{neg/low}$						
G1-G2 (n = 507)	22	445	40	4.3 % (22/507)	2.0 (0.6 – 6.6)	1.2 p = 0.28	7.9 % (40/507)	1	**25.8 p <0.001**
G3-G4 (n = 132)	3	98	31	2.3 % (3/132)	1		23.5 % (31/132)	3.6 (2.1 – 6.0)	

Pathological Grade	KRT20/CD177			% Tumors Group 1	OR [1] (95%CI[2])	Pearson's χ^2	% Tumors Group 3	OR [1] (95%CI[2])	Pearson's χ^2
	Group 1 KRT20^{+} CD177^{+}	Group 2 KRT20^{+} CD177^{neg}	Group 3 KRT20^{neg} CD177^{neg}						
G1-G2 (n = 506)	70	397	39	13.8 % (70/506)	2.2 (1.1 – 4.5)	**4.7 p = 0.03**	7.7 % (39/506)	1	**26.7 p <0.001**
G3-G4 (n = 132)	9	92	31	6.8 % (9/132)	1		23.5 % (31/132)	3.7 (2.2 – 6.2)	

Pathological Grade	KRT20/SLC26A3			% Tumors Group 1	OR [1] (95%CI[2])	Pearson's χ^2	% Tumors Group 3	OR [1] (95%CI[2])	Pearson's χ^2
	Group 1 KRT20^{+} SLC26A3^{+}	Group 2 KRT20^{+} SLC26A3^{neg}	Group 3 KRT20^{neg} SLC26A3^{neg}						
G1-G2 (n = 507)	100	367	40	19.7 % (100/507)	1.6 (0.9 – 2.7)	2.6 p = 0.11	7.9 % (40/507)	1	**25.8 p <0.001**
G3-G4 (n = 132)	18	83	31	13.6 % (18/132)	1		23.5 % (31/132)	3.6 (2.1 – 6.0)	

[1] OR: Odds-ratio; [2] CI: confidence interval

Supplementary Figure 21: **Relationship between "gene-expression groups" and pathological grade.** The relationship between traditional pathological grade and "gene-expression groups" identified based on the mRNA expression levels of KRT20 and one of four genes characteristic of "top-of-the-crypt" CA1^{+}/SLC26A3^{+} enterocyte-type cells (i.e. CA1, MS4A12, CD177, SLC26A3) was analyzed on a pooled database of 639 independent microarrays annotated with grading information ("grading database", Supplementary Table 1). The analysis indicated that the two classification systems are largely non-overlapping, but positively correlated. An analysis of the distribution of low-grade (G1/G2) vs high-grade (G3/G4) tumors with respect to the different gene-expression groups, indicated that Group 3 tumors are enriched in high-grade tumors (Pearson's $\chi 2$ test, $p < 0.001$), while Group 1 tumors display a trend towards being enriched in low-grade tumors, although in most cases not reaching statistical significance (Pearson's $\chi 2$ test, $p = 0.03$-0.11).

"*Gene-expression groups*" vs MSS/MSI status (MSS/MSI database, n = 229)

				Enrichment of Group 1 in MSS (Group 1 vs Group 2+3)			Enrichment of Group 3 in MSI (Group 3 vs Group 1+2)		
Pathological Grade	**KRT20/CA1**			**% Tumors Group 1**	**OR[1] (95%CI[2])**	**Pearson's χ^2**	**% Tumors Group 3**	**OR[1] (95%CI[2])**	**Pearson's χ^2**
	Group 1 KRT20$^+$ CA1^{high}	**Group 2 KRT20$^+$ CA1$^{neg/low}$**	**Group 3 KRT20^{neg} CA1$^{neg/low}$**						
MSS (n = 140)	21	115	4	15.0 % (21/140)	2.4 (0.9 – 6.3)	3.6 p = 0.06	2.9 % (4/140)	1	**18.9 p <0.001**
MSI (n = 89)	6	65	18	6.7 % (18/89)	1		20.2 % (18/89)	8.6 (2.8 – 26.4)	

Pathological Grade	**KRT20/MS4A12**			**% Tumors Group 1**	**OR [1] (95%CI[2])**	**Pearson's χ^2**	**% Tumors Group 3**	**OR [1] (95%CI[2])**	**Pearson's χ^2**
	Group 1 KRT20$^+$ MS4A12^{high}	**Group 2 KRT20$^+$ MS4A12$^{neg/low}$**	**Group 3 KRT20^{neg} MS4A12$^{neg/low}$**						
MSS (n = 140)	20	116	4	14.3 % (20/140)	2.8 (1.0 – 7.8)	**4.2 p = 0.04**	2.9 % (4/140)	1	**18.9 p <0.001**
MSI (n = 89)	5	66	18	5.6 % (5/89)	1		20.2 % (18/89)	8.6 (2.8 – 26.4)	

Pathological Grade	**KRT20/CD177**			**% Tumors Group 1**	**OR [1] (95%CI[2])**	**Pearson's χ^2**	**% Tumors Group 3**	**OR [1] (95%CI[2])**	**Pearson's χ^2**
	Group 1 KRT20$^+$ CD177$^+$	**Group 2 KRT20$^+$ CD177^{neg}**	**Group 3 KRT20^{neg} CD177^{neg}**						
MSS (n = 140)	16	120	4	11.4 % (16/140)	0.6 (0.3 – 1.2)	2.0 p = 0.15	2.9 % (4/140)	1	**17.5 p <0.001**
MSI (n = 88)	16	55	17	18.2 % (16/89)	1		19.3 % (17/88)	8.1 (2.6 – 25.1)	

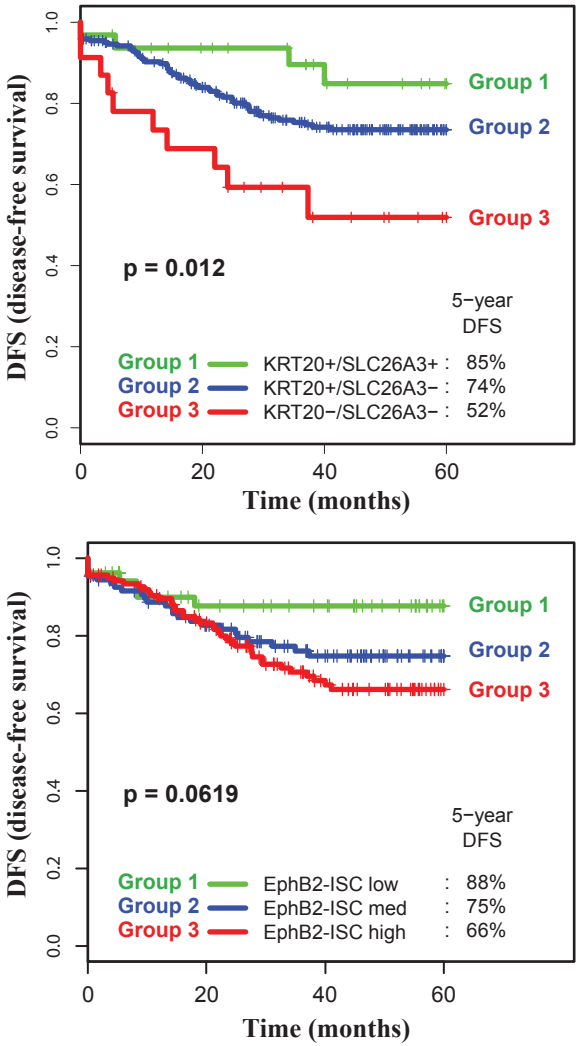
Pathological Grade	**KRT20/SLC26A3**			**% Tumors Group 1**	**OR [1] (95%CI[2])**	**Pearson's χ^2**	**% Tumors Group 3**	**OR [1] (95%CI[2])**	**Pearson's χ^2**
	Group 1 KRT20$^+$ SLC26A3$^+$	**Group 2 KRT20$^+$ SLC26A3^{neg}**	**Group 3 KRT20^{neg} SLC26A3^{neg}**						
MSS (n = 140)	37	99	4	26.4 % (37/140)	2.5 (1.2 – 5.3)	**6.5 p = 0.01**	2.9 % (4/140)	1	**18.9 p <0.001**
MSI (n = 89)	11	60	18	12.4 % (11/89)	1		20.2 % (18/89)	8.6 (2.8 – 26.4)	

[1] OR: Odds-ratio; [2] CI: confidence interval

Supplementary Figure 22: **Relationship between "gene-expression groups" and MSS/MSI status.** The relationship between microsatellite stability (MSS) or instability (MSI) status and "gene-expression groups" defined based on the mRNA expression levels of KRT20 and one of four genes characteristic of "top-of-the-crypt" CA1$^+$/SLC26A3$^+$ enterocyte-type cells (i.e. CA1, MS4A12, CD177, SLC26A3) was analyzed on a pooled database of 229 independent microarrays annotated with MSS/MSI information ("MSS/MSI database", Supplementary Table 1). The analysis indicated that the two variables are largely non-overlapping, but positively correlated. An analysis of the distribution of microsatellite stable (MSS) vs unstable (MSI) tumors with respect to the different gene-expression groups, indicated that Group 3 tumors are enriched in MSI tumors (Pearson's $\chi 2$ test, $p < 0.001$), while Group 1 tumors display a trend towards being enriched in MSS ones, although in most cases not reaching statistical significance (Pearson's $\chi 2$ test, $p = 0.01$-0.06). Based on this observation, the hypothesis that the prognostic effect of gene-expression groups might be caused by an enrichment of MSI tumors in Group 1 vs Group 3 tumors can be safely rejected.

A

Kaplan-Meier survival curves
Log-rank test
(Jorissen & Smith database, n = 299)



B

Multivariate analysis
Cox proportional hazards model

(Jorissen & Smith database, n = 299)

Prognostic variable	HR [1]	95% CI [2]	p-value
Groups (1-3, KRT20/SLC26A3)	**2.45**	**1.37 - 4.38**	**0.0026***
EphB2-ISC (low, medium, high) [3]	**1.47**	**1.03 – 2.09**	**0.0318***
Stage (I-IV)	**3.37**	**2.39 – 4.74**	**< 0.001****
Age[4]	0.89	0.74 – 1.07	0.21
Sex (M/F)[5]	1.18	0.92 – 1.51	0.20

(Smith database, n = 181)

Prognostic variable	HR [1]	95% CI [2]	p-value
Groups (1-3, KRT20/SLC26A3)	**2.37**	**1.13 - 4.96**	**0.022***
EphB2-ISC (low, medium, high) [3]	**1.66**	**1.05 – 2.63**	**0.029***
Stage (I-IV)	**3.35**	**2.14 – 5.24**	**< 0.001****
Grade (G1-G3)	1.21	0.63 – 2.33	0.57
Age[4]	0.91	0.72 – 1.15	0.41
Sex (M/F)[5]	1.30	0.94 – 1.79	0.11

[1] HR: hazard ratio
[2] CI: confidence interval
[3] Merlos-Suarez *et al.*, *Cell Stem Cell*, 8:511-524, 2011
[4] Age modeled as a continuous variable
[5] M/F: male vs female
*** p < 0.05**
**** p < 0.001**

Supplementary Figure 23. **The prognostic effect of "gene-expression groups" based on KRT20/SLC26A3 expression levels is independent of the EphB2-ISC signature.** Both KRT20/SLC26A3 gene-expression groups and the EphB2-ISC gene-expression signature can be used to stratify colon cancer patients in different groups characterized by different disease-free survival outcomes (A). A multivariate analysis comparing the prognostic effect of KRT20/SLC26A3 "gene-expression groups" with that of the EphB2-ISC signature indicated that the two prognostic systems do not confound each other, and that both are not confounded by stage or pathological grade (B; * p-value < 0.05, ** p-value < 0.001).

Supplementary Table 1. List of publicly available NCBI - GEO[1] datasets used for gene-discovery, gene-correlation and patient survival experiments.

NCBI - GEO dataset	number of samples	Affymetrix® Platform	PubMed ID	Reference
Human Colon - global database				
GSE2109 (only colorectal cancer patients)	n = 427	HG U133 Plus 2.0	n.a.	Expression Project for Oncology (expO)[2]
GSE2361 (only one normal colon sample)	n = 1	HG U133A	PMID 15950434	Ge *et al.*, *Genomics*, 86:127-141, 2005
GSE4045	n = 37	HG U133A	PMID 16819509	Laiho *et al.*, *Oncogene*, 26:312-320, 2007
GSE4107	n = 22	HG U133 Plus 2.0	PMID 17317818	Hong *et al.*, *Clin Cancer Res*, 13:1107-1114, 2007
GSE4183 (excluding inflammatory bowel disease)	n = 38	HG U133 Plus 2.0	PMID 19461970	Gyorffy *et al.*, *PLoS One*, 4:e5645, 2009
GSE5851	n = 80	HG U133A 2.0	PMID 17664471	Khambata-Ford *et al.*, *J. Clin. Oncol.*, 25:3230-3237, 2007
GSE8671	n = 64	HG U133 Plus 2.0	PMID 18171984	Sabates-Bellver *et al.*, *Mol Cancer Res*, 5:1263-1275, 2007
GSE9254	n = 19	HG U133 Plus 2.0	PMID 18056783	La Pointe *et al.*, *Physiol Genomics*, 33:50-64, 2008
GSE9348	n = 82	HG U133 Plus 2.0	PMID 20143136	Hong *et al.*, *Clin. Exp. Metastasis*, 27:83-90, 2010
GSE10714 (excluding inflammatory bowel disease)	n = 15	HG U133 Plus 2.0	PMID 20087348	Galamb *et al.*, *Br. J. Cancer*, 102:765-773, 2010
GSE10961	n = 18	HG U133 Plus 2.0	PMID 18827815	Pantaleo *et al.*, *Br J Cancer*, 99:1729-1734, 2008
GSE11831	n = 17	HG U133 Plus 2.0	PMID 19603079	Nielsen *et al.*, *PLoS One*, 4:e6210, 2009
GSE12945	n = 62	HG U133A	PMID 19399471	Staub *et al.*, *J. Mol. Med.*, 87:633-644, 2009
GSE13067	n = 74	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin Cancer Res*, 14:8061-8069, 2008
GSE13294	n = 155	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin Cancer Res*, 14:8061-8069, 2008
GSE13471 (only colon samples)	n = 8	HG U133A	PMID 19151715	Irizarry *et al.*, *Nat. Genet.*, 41:178-186, 2009
GSE14333 (samples non-redundant with GSE13067)	n = 226	HG U133 Plus 2.0	PMID 19996206	Jorissen *et al.*, *Clin. Cancer Res.*, 15:7642-7651, 2009
GSE15960	n = 18	HG U133 Plus 2.0	PMID 20087348	Galamb *et al.*, *Br. J. Cancer*, 102:765-773, 2010
GSE17538 (samples non-redundant with GSE14333)	n = 65[3]	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968, 2010
GSE18105	n = 111	HG U133 Plus 2.0	PMID 20162577	Matsuyama *et al.*, *Int. J. Cancer*, 127:2292-2299, 2010
GSE20916	n = 145	HG U133 Plus 2.0	PMID 20957034	Skrzypczak *et al.*, *PLoS One*, 5:e13091, 2010
Total number of samples	n = 1684			
Total number of samples after "purging"[4]	n = 1568			
Colon Cancer - pathological grading database				
GSE2109 (only samples with grading information)	n = 367	HG U133 Plus 2.0	n.a.	Expression Project for Oncology (expO)[2]
GSE4045 (only samples with grading information)	n = 23	HG U133A	PMID 16819509	Laiho *et al.*, *Oncogene*, 26:312-320, 2007
GSE12945	n = 62	HG U133A	PMID 19399471	Staub *et al.*, *J. Mol. Med.*, 87:633-644, 2009
GSE17538 (only samples with grading information)	n = 213	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968, 2010
Total number of samples	n = 665			
Total number of samples after "purging"[4]	n = 639			
Colon Cancer - disease-free survival (DFS) database				
GSE17538 (DFS data, VMC + MCC)[5]	n = 200	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968, 2010
GSE14333 (DFS data, Melbourne + MCC)[6]	n = 99	HG U133 Plus 2.0	PMID 19996206	Jorissen *et al.*, *Clin. Cancer Res.*, 15:7642-7651, 2009
Total number of samples[7]	n = 299			
Colon Cancer - multivariate analysis vs. grading				
GSE17538 (patients with both DFS and grading data)	n = 181	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968, 2010
Total number of samples[7]	n = 181			
Colon Cancer - MSI/MSS database				
GSE13067	n = 74	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin Cancer Res*, 14:8061-8069, 2008
GSE13294	n = 155	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin Cancer Res*, 14:8061-8069, 2008
Total number of samples[7]	n = 229			

[1] National Center for Biotechnology Information (NCBI) - Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo

[2] International Genomic Consortium (IGC) - Expression Project for Oncology (expO), https://expo.intgen.org/geo/

[3] Six additional patients without DFS data from the VMC were recently added to the GSE17538 database: they are not included here in the global database.

[4] After removal of samples that do not fulfill the $EpCAM^{+}/ALB^{neg}$ condition (see also Supplementary Fig. 6)

[5] Only patients with DFS data: Vanderbilt Medical Center (n = 55, VMC) and Moffit Cancer Center (n = 145, MCC).

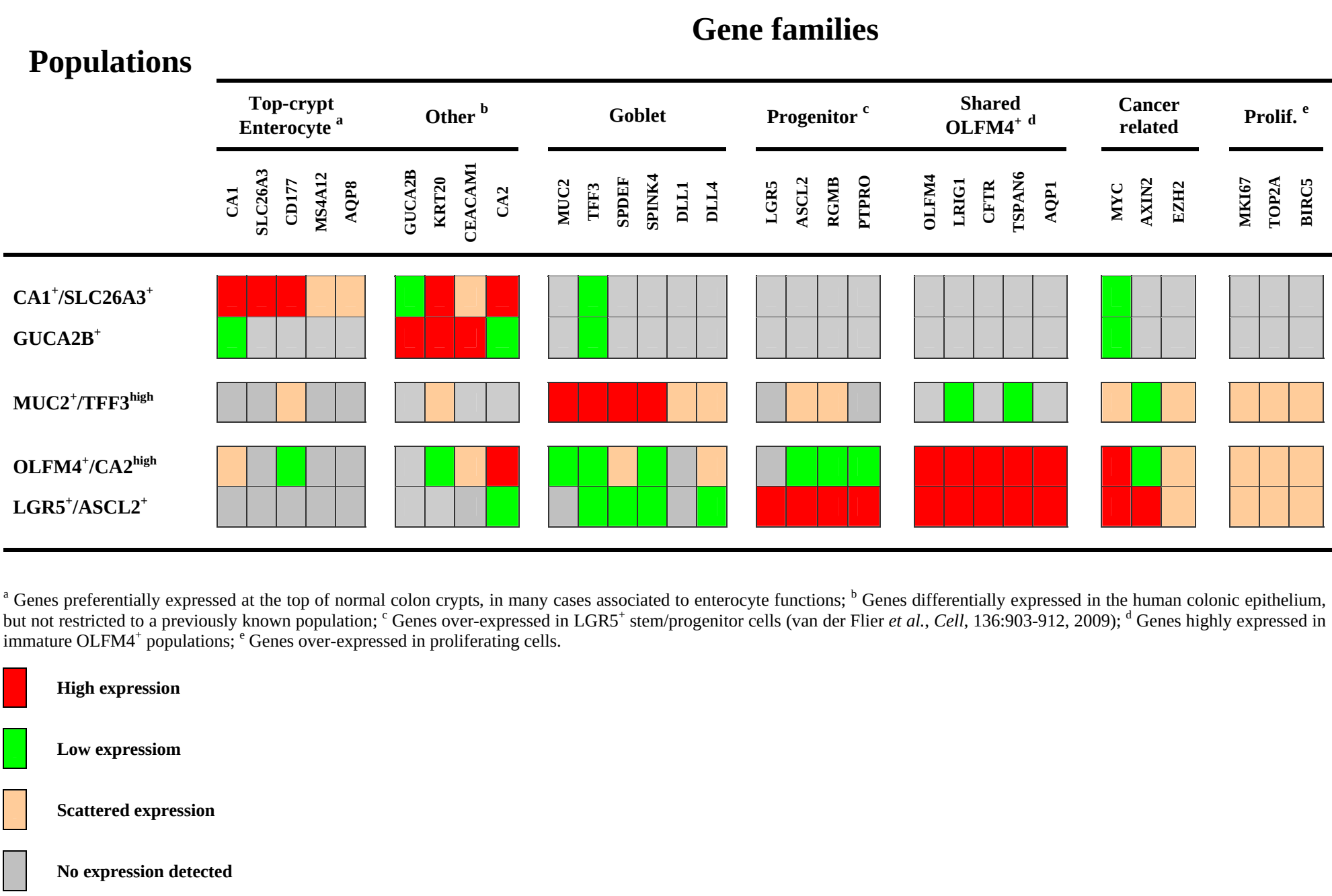
[6] Only patients with DFS data, non-duplicated between GSE14333 and GSE17538: Melbourne Royal Hospital (n = 80, Melbourne) and Moffit Cancer Center (n = 19, MCC).

[7] No purging required as only 1 sample did not fulfill the $EpCAM^{+}/ALB^{neg}$ condition

Supplementary Table 2. List of TaqMan® gene-expression assays (Applied Biosystems) used for SINCE-PCR experiments on human colon epithelial cells.

TaqMan® assay ID (Applied Biosystems)	Gene Symbol	Gene Name
Positive controls		
Hs00357333_g1	**ACTB**	actin, beta
Hs99999905_m1	**GAPDH**	glyceraldehyde-3-phosphate dehydrogenase
Hs00158980_m1	**EPCAM (TACSTD1)**	tumor-associated calcium signal transducer 1
Proliferation-related genes		
Hs00153353_m1	**BIRC5 (Survivin)**	baculoviral IAP repeat-containing 5 (survivin)
Hs00267195_m1	**MKI67 (Ki67)**	antigen identified by monoclonal antibody Ki-67
Hs01032137_m1	**TOP2A**	topoisomerase (DNA) II alpha 170kDa
Colon - differentially expressed genes*		
Hs01028916_m1	**AQP1**	aquaporin 1 (Colton blood group)
Hs01086279_m1	**AQP8**	aquaporin 8
custom designed - AIMRUO9	**ASCL2**	achaete schute-like 2, achaete-scute complex homolog 2 (Drosophila)
Hs00610344_m1	**AXIN2**	axin 2 (conductin, axil)
Hs00180411_m1	**BMI1**	BMI1 polycomb ring finger oncogene
Hs00266139_m1	**CA1**	carbonic anhydrase I
Hs01070106_m1	**CA2**	carbonic anhydrase II
Hs00360669_m1	**CD177**	CD177 molecule
Hs00912242_g1	**CDCA7**	cell division cycle associated 7
Hs00608037_m1	**CDK6**	cyclin-dependent kinase 6
Hs00266109_m1	**CEACAM1 #1**	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)
Hs00989784_m1	**CEACAM1 #2**	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)
Hs01565537_m1	**CFTR**	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
Hs00601975_m1	**CXCL2**	chemokine (C-X-C motif) ligand 2
Hs01011325_g1	**DLL1**	delta-like 1 (Drosophila)
Hs01117332_g1	**DLL4 #1**	delta-like 4 (Drosophila)
Hs00184092_m1	**DLL4 #2**	delta-like 4 (Drosophila)
Hs01027166_m1	**DNMT3A**	DNA (cytosine-5-)-methyltransferase 3 alpha
Hs00175210_m1	**DPP4 (CD26)**	dipeptidyl-peptidase 4
Hs00193306_m1	**EGFR**	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)
Hs01016789_m1	**EZH2**	enhancer of zeste homolog 2 (Drosophila)
Hs00916793_m1	**FERMT1**	fermitin family homolog 1 (Drosophila)
Hs00203271_m1	**GPSM2**	G-protein signaling modulator 2 (AGS3-like, C. elegans)
Hs00951189_m1	**GUCA2B**	guanylate cyclase activator 2B (uroguanylin)
Hs01118948_g1	**HES1**	hairy and enhancer of split 1, (Drosophila)
Hs00293523_m1	**KIF12**	kinesin family member 12
Hs00300643_m1	**KRT20**	keratin 20
Hs01059008_m1	**LATS2**	LATS, large tumor suppressor, homolog 2 (Drosophila)
Hs00969421_m1	**LGR5 #1**	leucine-rich repeat-containing G protein-coupled receptor 5
Hs00969423_m1	**LGR5 #2**	leucine-rich repeat-containing G protein-coupled receptor 5
Hs00394267_m1	**LRIG1**	leucine-rich repeats and immunoglobulin-like domains 1
Hs01096158_m1	**METTL3**	methyltransferase like 3
Hs00946021_m1	**MLLT10**	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10
Hs00214572_m1	**MS4A12**	membrane-spanning 4-domains, subfamily A, member 12
Hs03005094_m1	**MUC2**	mucin 2, oligomeric mucus/gel-forming
Hs00153408_m1	**MYC (c-Myc)**	v-myc myelocytomatosis viral oncogene homolog (avian)
Hs00413187_m1	**NOTCH1**	Notch homolog 1, translocation-associated (Drosophila)
Hs00197437_m1	**OLFM4**	olfactomedin 4
Hs01012905_m1	**PTPLAD1**	protein tyrosine phosphatase-like A domain containing 1
Hs00243097_m1	**PTPRO**	protein tyrosine phosphatase, receptor type, O
Hs01551808_m1	**RGMB**	RGM domain family, member B
Hs00993304_m1	**RNF43**	ring finger protein 43
Hs00995365_m1	**SLC26A3 (DRA)**	solute carrier family 26, member 3
Hs01026048_m1	**SPDEF #1**	SAM pointed domain containing ets transcription factor
Hs00171942_m1	**SPDEF #2**	SAM pointed domain containing ets transcription factor
Hs01018780_m1	**SPINK4**	serine peptidase inhibitor, Kazal type 4
Hs00606370_m1	**STMN1**	stathmin 1/oncoprotein 18
Hs00173625_m1	**TFF3**	trefoil factor 3 (intestinal)
Hs01073458_m1	**TSPAN6**	tetraspanin 6
Hs00409961_m1	**UGT8**	UDP glycosyltransferase 8
Hs00903129_m1	**VEGFA**	vascular endothelial growth factor A
Positive controls	**n = 3**	
Proliferation-related genes	**n = 3**	
Colon - differentially expressed genes	**n = 51***	* This is the assay-set used for hierarchical clustering and PCA analysis
Total number of assays	**n = 57**	

Supplementary Table 3. Gene-expression profile of colon epithelium cellular subpopulations identified by SINCE-PCR.



[a] Genes preferentially expressed at the top of normal colon crypts, in many cases associated to enterocyte functions; [b] Genes differentially expressed in the human colonic epithelium, but not restricted to a previously known population; [c] Genes over-expressed in LGR5^{+} stem/progenitor cells (van der Flier *et al.*, *Cell*, 136:903-912, 2009); [d] Genes highly expressed in immature OLFM4^{+} populations; [e] Genes over-expressed in proliferating cells.

Piero Dalerba, Tomer Kalisky, Debashis Sahoo *et al.*, **"Single-cell dissection of transcriptional heterogeneity in human colon tumors."** *Nature Biotechnology*

Supplementary Table 4. Origin, morphology and differentiation marker expression of colorectal tumors included in this study.

Tumor	Origin	Patient data		Stage			Tissue source[b]	Pathology and differentiation markers		
		sex	age	TNM[a]	Dukes	AJCC[a]		diagnosis	CK20[c]	MUC2[c]
UM-COLON-#4	Right Colon, primary tumor	♂	62	T3N0	B	IIa	xeno.	adenocarcinoma	+	+
UM-COLON-#8	Sigmoid Colon, primary tumor	♂	49	T3N0	B	IIa	xeno.	adenocarcinoma	++	++
SU-COLON-#34	Rectum, primary tumor	♂	44	T2N0	A	I	xeno.	adenocarcinoma	++	+
SU-COLON-#56	Right Colon, primary tumor	♂	61	T3N0M1a	D	IVa	prim.	adenocarcinoma	++	+++
SU-COLON-#60	Right Colon, primary tumor	♂	58	T4aN2bM1b	D	IVb	xeno.	adenocarcinoma	+	neg
SU-COLON-#62	Right Colon, primary tumor	♂	44	T3N0	B	IIa	prim/xeno.	adenocarcinoma	++	++
SU-COLON-#64	Lymphnode, metastasis	♂	62	T3N1b	C	IIIb	lymph met./xeno.	adenocarcinoma	+++	neg
SU-COLON-#72	Cecum, primary tumor	♂	59	T3N2aM1b	D	IVb	prim.	adenocarcinoma	++	++
SU-COLON-#76	Left colon, adenoma	♀	77	n.a.[d]	n.a.[d]	n.a.[d]	prim.	tubulovillous adenoma	+++	+++
SU-COLON-#82	Liver, metastasis	♂	83	T3N2aM1a	D	IVa	liver met.	adenocarcinoma	+	neg
SU-COLON-#87	Liver, metastasis	♂	78	M1 (liver)	D	IV	liver met.	adenocarcinoma	++++	neg
SU-COLON-#94	Liver, metastasis	♀	55	T3N1aM1a	D	IVa	liver met.	adenocarcinoma	++	++
SU-COLON-#96	Sigmoid Colon, primary tumor	♀	73	T2N1bM1a	D	IVa	prim.	adenocarcinoma	+++	+++
SU-COLON-#98	Right Colon, primary tumor	♂	53	T3N0	B	IIa	prim.	adenocarcinoma (poorly diff.)	neg	neg

[a]According to the 7th edition of the American Joint Committee on Cancer (AJCC) staging system for colorectal cancer (2009). [b]Tumor tissues utilized for experimental studies included surgical specimens from primary tumors (prim.), lymphnode metasatases (lymph. met), liver metastases (liver met.), as well as solid xenografts (xeno.) established in NOD-SCID and/or NOD/SCID/IL2Rγ$^{-/-}$ immunodeficient mice. [c]Cytokeratin-20 (KRT20) and MUC2 expression was assessed by immunohistochemistry (Supplementary Fig. 15 and 17). [d] n.a.: not applicable.

Supplementary Methods

"Single-cell dissection of transcriptional heterogeneity in human colon tumors."

Piero Dalerba, Tomer Kalisky, Debashis Sahoo et al., Nature Biotechnology

Primary human tissues and human colon cancer xenograft lines. All primary human tissues, both normal and cancerous, were collected under protocols approved by Stanford University's institutional review board between 2006 and 2010. Informed consent was obtained from all patients included in the study. A list of all human colorectal cancer tissues used in this study, either from primary samples or xenograft lines, is provided in Supplementary Table 4, together with clinical information related to corresponding patients. Colon cancer xenograft lines were established by subcutaneous (s.c.) implantation of solid tissue fragments in 6- to 8-week-old NOD/SCID or NOD/SCID/IL2R$\gamma^{-/-}$ (NSG) mice (Charles River Laboratories, Wilmington, MA; The Jackson Laboratory, Bar Harbor, ME), as previously described [1]. Briefly, primary human colorectal cancer tissue specimens were minced with scissors into small (2 mm^3) fragments and implanted s.c. using a 10-gauge Trochar needle, through a small incision on the animal's right dorsal flank. Recipient mice were briefly anesthetized by isofluorane inhalation (AErrane®, Baxter Healthcare Corporation, Deerfield, IL) using a standard vaporizer (5% for induction, 2% for maintenance) [2]. Once established, solid tumor xenografts were serially passaged by using the same technique. Of the 6 xenograft lines used in this study, two (UM-COLON#4, #8) originated from the University of Michigan and four (SU-COLON#34, #60, #62, #64) from Stanford University. Xenograft lines established at the University of Michigan and Stanford University are part of a collection

of 31 independent lines originating from 47 distinct primary colon carcinoma specimens, for an estimated comprehensive success rate of 66% (n = 31 of 47).

Solid Tissue Disaggregation. Solid tissues, both normal and neoplastic, collected from primary surgical specimens or mouse xenografts, were mechanically and enzymatically disaggregated into single-cell suspensions and analyzed by flow cytometry, as described by Dalerba *et al.* [1]. Briefly, solid tissues were minced with scissors into small (2 mm^3) fragments, rinsed once with Hank's balanced salt solution (HBSS), finely chopped with a razor blade into minute (0.2 – 0.5 mm^3) aggregates, resuspended in serum-free RPMI medium 1640 (2 mM L-glutamine, 120 μg/ml penicillin, 100 μg/ml streptomycin, 50 μg/ml ceftazidime, 0.25 μg/ml amphotericin-B, 20 mM Hepes, 1mM Sodium Pyruvate) with 200 units/ml Collagenase type III (Worthington, Lakewood, NJ) and 100 units/ml DNase I (Worthington), and incubated for 2 h at 37°C to obtain enzymatic disaggregation. Cells were then resuspended by pipetting and serially filtered by using sterile gauze and 70-μm and 40-μm nylon meshes. Contaminating red blood cells were removed by osmotic lysis with ACK hypotonic buffer (i.e. incubation in 150 mM NH_4Cl, 1 mM $KHCO_3$ for 5 min. on ice).

Cell lines. Calibration experiments to measure accuracy and precision of single-cell sorting by flow cytometry, as well as to measure the single-cell sensitivity of the SINCE-PCR method, were performed on a clone of the HCT116 human colon cancer cell line infected with the pLentiLox 3.7 lentivirus (pLL3.7, Addgene plasmid #11795, http://www.addgene.org), which encodes for the enhanced green fluorescent protein

(EGFP). HCT116 cells are available from the American Tissue-type Culture Collection (ATCC; catalog number CCL-247, http://www.atcc.org). Cell cultures were maintained in RPMI-1640 medium, supplemented with 10% heat-inactivated fetal bovine serum (FBS), 2 mM L-glutamine, 120 μg/ml penicillin, 100 μg/ml streptomycin, 20 mM Hepes and 1mM Sodium Pyruvate, as previously described [3]. A detailed description of the lentivirus infection protocol is provided below, under the paragraph "*Lentivirus infection and LM-PCR characterization of lentivirus integration sites*."

Flow Cytometry and single-cell sorting experiments. To minimize experimental variability and loss of cell viability, all experiments were performed on fresh tumor cell suspensions prepared shortly before flow cytometry. Antibody staining was performed in HBSS supplemented with 2% heat-inactivated calf serum, 120 μg/ml penicillin, 100 μg/ml streptomycin, 50 μg/ml ceftazidime, 0.25 μg/ml amphotericin-B, 20 mM Hepes, 1mM Sodium Pyruvate and 5 mM EDTA. To minimize unspecific binding of antibodies, cells were first incubated with 0.6% human immunoglobulins (Gammagard Liquid; Baxter, Westlake Village, CA) for 10 min on ice at a concentration of 3-5 × 10^5 cells/100 μl. Cells were subsequently washed and stained with antibodies at dilutions determined by titration experiments on each xenograft line. Antibodies used in this study include: anti-human EpCAM-FITC (clone 9C4; BioLegend, San Diego, CA), anti-human CD44-APC (clone G44–26; BD Biosciences, San Diego, CA), anti-human CD166-PE (clone 105902; R&D Systems, Minneapolis, MN), anti-human CD66a-PE (clone 283340; R&D Systems). Cells positive for expression of non-epithelial lineage markers (Lin^+) were excluded by staining with PE.Cy5-labeled antibodies using two different strategies for

primary tissues and mouse xenografts. In experiments on primary human tissues, stromal cells were excluded by simultaneous staining with anti-human CD3-biotin (clone UCHT1; BD Biosciences), CD16-biotin (clone 3G8; BD Biosciences), CD45-biotin (clone HI30; BD Biosciences), and CD64-biotin (clone 10.1; BD Biosciences) + streptavidin-PE/Cy5 (BD Biosciences). In experiments on human colon cancer xenografts, mouse cells were excluded by simultaneous staining with anti-mouse CD45-PE/Cy5 (clone 30-F11; BD Biosciences) and anti-mouse H-2Kd-biotin (clone SF1–1.1; BD Biosciences) + streptavidin-PE/Cy5 (BD Biosciences). After 15 min on ice, stained cells were washed of excess unbound antibodies and resuspended in HBSS supplemented with 2% heat-inactivated calf serum, 20 mM Hepes, 5 mM EDTA, 1mM Sodium Pyruvate, and 1.1 μM DAPI dilactate (Molecular Probes, Eugene, OR) as viability dye. Flow-cytometry analysis was performed using a BD FACSAriaII cell sorter (Becton Dickinson, San Jose, CA). Forward-scatter height versus forward-scatter width (FSC-H vs FSC-W) and side-scatter height vs side-scatter width (SSC-H vs. SSC-W) profiles were used to eliminate cell doublets. Dead cells were eliminated by excluding $DAPI^+$ cells, whereas contaminating human or mouse Lin^+ cells were eliminated by excluding $PE/Cy5^+$ cells. In single-cell sorting experiments, each single (n = 1) cell was individually sorted into a different well of a 96-well PCR plate, using a protocol already built-in within the FACSAriaII flow cytometer software package (FACSDiva), with appropriate adjustements (device: 96-well plate; precision: single-cell; nozzle: 130 μm).

SINCE-PCR. Single cell gene-expression experiments were performed using Fluidigm's M48 or M96 quantitative PCR (qPCR) DynamicArray™ microfluidic chips.

Single cells were sorted by FACS into individual wells of 96-well PCR plates using a FACSAriaII flow cytometer (Becton Dickinson). Each well was pre-loaded with 5 μl of CellsDirect PCR mix (Invitrogen, Carlsbad, CA) and 0.1 μl (2 U) of SuperaseIn RNase Inhibitor (Invitrogen), promptly frozen and stored at -20C. On the day of analysis, 96-well plates were thawed and each well was supplemented with 1 μl of SuperScript III RT/Platinum Taq (Invitrogen), 1.5 μl of Tris-EDTA (TE) buffer and 2.5 μl of a mixture of 96 pooled TaqMan® assays (Applied Biosystems, Foster City, CA), containing each assay at 1:100 dilution. A list of the 57 gene-specific TaqMan® assays used in this study and their identification codes can be found in Supplementary Table 2. The mRNA from the cell lysates was then reverse-transcribed into cDNA (50°C for 15 min., 95°C for 2 min.) and pre-amplified for 20 PCR cycles (each cycle: 95°C for 15 sec, 60°C for 4 min.). As a positive control for each TaqMan® assay, we used a 1:1:1 mixture of RNA from human normal colon (Applied Biosystems, AM7986), human normal testes (Applied Biosystems, AM7972) and HeLa cells (Applied Biosystems, AM7852). The resulting amplified cDNA from each one of the cells was diluted 1:3 with TE buffer. A 2.25 μl aliquot of amplified cDNA was then mixed with 2.5 μl of TaqMan qPCR mix (Applied Biosystems) and 0.25 μl of Fluidigm "sample loading agent" (Fluidigm) and finally inserted into one of the chip "sample" inlets. Individual gene-specific TaqMan® assays were diluted at 1:1 ratios with TE. A 2.5 μl aliquot of each diluted TaqMan® assay was then mixed with 2.5 μl of Fluidigm "assay loading agent" (Fluidigm) and individually inserted into the chip "assay" inlets. Samples and probes were loaded into M96 chips using an HX IFC Controller (Fluidigm) and then transferred to a BioMark™ real-time PCR reader (Fluidigm) following the manufacturer's protocols and instructions.

Measure of SINCE-PCR sensitivity. To ensure that gene-expression measurements performed on single-cells were within the range of qPCR sensitivity, we performed a calibration experiment, comparing threshold cycle (Ct) measurements on single-cells from the HCT116 cell line with Ct measurements on 10-fold serial dilutions of the RNA standard mixture used as positive control (Supplementary Fig. 4). Titration curves obtained from 10-fold serial dilutions of the RNA standards confirmed that the SINCE-PCR method was able to robustly amplify multiple target mRNAs from a wide range of starting materials (100ng-1pg total RNA). Most importantly, parallel results obtained from HCT116 single-cells indicated that the average amount of target mRNA per cell was within SINCE-PCR's linear range of analysis across multiple genes (Supplementary Fig. 4).

Analysis and graphic display of SINCE-PCR data. SINCE-PCR data were analyzed and displayed using MATLAB® (MathWorks Inc., Natick, MA), as schematically summarized in Supplementary Figure 2. In each experiment, a minimum of 336 cells was analyzed for each phenotypic population, corresponding to 4 PCR plates, each containing 84 single-cells (84 x 4 = 336), 8 positive controls and 4 negative controls. Cells not expressing the housekeeping genes ACTB (β-actin) and GAPDH (Glyceraldehyde 3-phosphate dehydrogenase), or expressing them at extremely low values (Ct >35), were removed from the analysis, on the assumption that cells were absent, dead or damaged. The percentage of cells removed from the analysis due to failure to amplify housekeeping genes ranged from 5% to 15% of the total. All cells

included in the analysis scored positive for expression of EpCAM (Epithelial Cell Adhesion Molecule), as a confirmation of their epithelial cell lineage of origin.

Gene-expression results were normalized gene-by-gene, by mean-centering and dividing by 3 times the standard deviation (3 SD) of expressing cells (Supplementary Fig. 2). Hierarchical clustering was performed on both cells and genes, with a Euclidean or correlation distance metric and complete linkage. Hierarchical clustering was based on the results for 47 differentially expressed genes (51 assays), and excluded results from housekeeping genes (3 assays; ACTB, GAPDH, EpCAM) and proliferation-related genes (3 assays; MKI67, TOP2A, BIRC5/Survivin) to avoid noise based on proliferation status. Positive or negative associations among pairs of genes were tested by Spearman correlation, and p-values were calculated using n = 10.000 permutations.

Screening and selection of TaqMan® assays for SINCE-PCR. Using an iterative approach, we screened more than 250 TaqMan® assays (Applied Biosystems) to test for the differential expression of more than 230 genes in single cells from 8 independent samples of normal human colon epithelium (7 samples analyzed for 96 genes and 1 sample analyzed for 48 genes in parallel). At each round, genes that were non-informative for the previous sample were removed (i.e. not differentially expressed in either positive or negative association with CA1, MUC2 or LGR5) and replaced with new candidate genes. Thereby, we progressively built a list of 57 TaqMan assays that allowed us to analyze the expression pattern of 53 distinct genes, and robustly visualize and characterize multiple cell populations (Supplementary Fig. 10). A list of the 57 gene-specific TaqMan® assays used in this study and their identification codes can be found in

Supplementary Table 2. Among them, 24 assays (42%) for 24 of the genes (45%) were tested across all 8 normal colon samples, 42 assays (74%) for 39 of the genes (74%) were validated across at least 6 samples, 53 assays (93%) for 49 of the genes (92%) were validated across at least 3 samples, and only 2 assays (4%) for 2 of the genes (4%; DPP4/CD26, GUCA2B) were added in the last round. This demonstrated that the subpopulations were robust to small changes in the gene list and could be reproducibly visualized across independent samples.

Principal component analysis (PCA) of SINCE-PCR data. PCA is a technique used to identify the major sources of variation within a set of data characterized by many variables [4]. In essence, PCA is a mathematical process that reduces the number of variables that contribute to the diversity of a specific set of data by identifying novel, compounded variables, called principal components (PC), along which the variability of the data is highest. PCA was performed on normalized Ct values from SINCE-PCR experiments, as previously described by Guo *et al.* [5]. To allow comparison between PCA and hierarchical clustering results, cell populations visualized by hierarchical clustering and biologically annotated based on their gene-expression profiles were labeled with different colors on PC1 vs PC2 plots (Fig. 1, G; Fig. 2, D, I). Similarly, genes whose expression patterns appeared to be positively and coordinately associated with individual cell populations in hierarchical clustering were labeled with similar colors when evaluated in their individual contributions to major principal components (PC loading; Fig. 1, H; Fig. 2, E, J). Similar to hierarchical clustering, PCA was based on the results for 47 differentially expressed genes (51 assays), and excluded results from housekeeping

genes (3 assays; ACTB, GAPDH, EpCAM) and proliferation-related genes (3 assays; MKI67, TOP2A, BIRC5/Survivin) to avoid noise based on proliferation status.

Immunohistochemistry. Immunohistochemical analysis of tumor tissues was performed on formalin-fixed, paraffin-embedded tissue sections. Tissue sections were stained with anti-human CK20 (clone Ks20.8, DakoCytomation), anti-human MUC2 (clone Ccp58, Fitzgerald Industries), anti-human Ki67 (clone MIB-1, DakoCytomation) and anti-human CEACAM1/CD66a (clone 283340; R&D Systems) monoclonal antibodies, according to manufacturer instructions. In the case of SLC26A3, tissue sections were stained with an affinity-isolated rabbit anti-human polyclonal antibody preparation (Lot #R32905; Sigma Life Science – Atlas Antibodies), again following the manufacturer's instructions.

Immunofluorescence. Analysis of CD177 protein expression in normal human colon epithelia was performed by immunofluorescence on frozen tissue sections, cut from fresh primary surgical samples embedded in O.C.T. (optimal cutting temperature) compound (Sakura Finetek, Torrance, CA). Tissue sections were fixed in Acetone at -20 C for 5 minutes and air-dried at room-temperature for 10 minutes, then re-hydrated, permeabilized and blocked by incubation at room temperature for 30 minutes with PBS supplemented with 0.1% Triton X100, 5% heat-inactivated horse serum and 0.6% human immunoglobulins (hIgG). Primary antibody staining was performed using a mouse anti-human CD177 monoclonal antibody (clone MEM-166, BD Biosciences), followed by three washes with PBS and a secondary staining with an affinity-purified goat anti-mouse

IgG (H+L) polyclonal antibody preparation conjugated to the Alexa-488 fluorochrome (Invitrogen).

Lentivirus infection and LM-PCR characterization of lentivirus integration sites. Human EpCAM^{high}/CD44^{+} colon cancer cells, freshly purified from an UM-COLON4 xenograft, were infected with the pLentiLox 3.7 lentivirus (pLL3.7), which carries the enhanced green fluorescence protein (EGFP) as a green fluorescent selection marker (Addgene plasmid #11795, http://www.addgene.org). Cells were infected by spin-inoculation for 4 hours [6] and injected in bulk into the s.c. tissue of a NOD/SCID/IL2Rγ$^{-/-}$ mice. The resulting tumors were analyzed to evaluate infection efficiency, and EGFP^{+}/EpCAM^{high}/CD44^{+} were re-sorted and injected as single-cells, again into NOD/SCID/IL2Rγ$^{-/-}$ mice. The monoclonal origin of resulting tumors was confirmed by detection of a unique lentivirus integration site in cancer cells using a ligation-mediated PCR (LM-PCR) technique previously described by Wang *et al.* [7] and Mitchell *et al.* [8]. In the case of UM-COLON#4 Clone 8, DNA sequencing of LM-PCR amplification products revealed that the provirus was inserted on the long arm of human chromosome 19 (19q13.3), in proximity of the AP3D1 gene (adaptor-related protein complex 3, delta 1 subunit).

Tumorigenicity Experiments. The *in vivo* tumorigenic potential of human colorectal cancer cells was assessed according to previously published protocols [1], using NOD/SCID/IL2Rγ$^{-/-}$ immunodeficient mice [9, 10]. Sorted cells were spun down by low-speed centrifugation (850 × g for 5 min) and resuspended in RPMI 1640 supplemented

with 10% FBS, 2 mM l-glutamine, 20 mM Hepes and 1mM Sodium Pyruvate. In all experiments, a small aliquot of cells was set aside to confirm cell counts and viability using conventional techniques (i.e. trypan blue exclusion test). Once cell counts and viability were confirmed, cells were diluted to appropriate injection doses, mixed with BD Matrigel (BD Biosciences) at 1:1 ratio, and injected s.c. in NOD/SCID/IL2R$\gamma^{-/-}$ mice on the ventral side of each flank. To minimize experimental variability due to individual differences in recipient mice, cell populations subjected to comparison were injected on opposite flanks of the same animals. Injected mice were monitored weekly for tumor engraftment up to a maximum of 5 months, and euthanized once engrafted tumors reached a maximum diameter of 15 mm. All experiments involving the use of animals were performed in accordance with Stanford University's institutional animal welfare guidelines. Calculation of tumorigenic cell frequencies by limiting dilution assay (LDA) was performed using the L-Calc software (StemCell Technologies Inc., Vancouver, Canada, www.stemcell.com) [11].

Bioinformatic data collection and generation of a "*human colon global database*". All bioinformatic analyses were performed starting from a collection of 46,047 publicly available human gene-expression arrays, including 25,721 arrays on the human Affymetrix U133 Plus 2.0 platform, 16,357 arrays on the human Affymetrix U133A platform and 3,969 arrays on the human Affymetix U133A 2.0 platform. All gene-expression arrays were downloaded from NCBI's GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo) database and normalized using the RMA (Robust Multi-chip Average) algorithm. Normalization was performed either independently for

each of the different Affymetrix platforms or on the whole array collection, using a modified CDF (chip description file) reduced to contain only shared probes. From this general collection, composed of gene-expression arrays from all types of human samples, we extracted a subset database of 1,684 unique gene-expression arrays from human colon tissues, either normal or cancerous. We used this subset database as the "*human colon global database*", and we annotated all samples contained in it as normal colon mucosa (n = 173), benign colonic adenoma (n = 68) or colorectal cancer (n = 1443). To avoid redundancies (i.e. identical samples deposited two or more times in independent GEO datasets) we cross-checked all samples contained in our collection and removed duplicates. When available, we also collected all available clinical, pathological and molecular information related to the corresponding patients, with a special focus on the larger human colorectal cancer datasets [12-14]. Since not all arrays were annotated for all variables, individual hypotheses were tested on different subsets of the "*human colon global database*". A detailed listing of all GEO datasets used in this study, and their contribution to different analyses, is provided in Supplementary Table 1.

Computer-assisted data mining of gene-expression arrays using Boolean implications. To determine gene-expression thresholds between positive and negative samples, we used the StepMiner algorithm (Supplementary Fig. 6) [15]. Briefly, for each gene the expression values of individual samples were ordered from low-to-high, and a rising step function was fit to the data, trying to minimize the differences between the fitted and measured values. This approach identifies the step at the point of largest jump from low values to high values (but only if there are sufficiently many expression values

on each side of the jump to exclude a random oscillation due to noise) and sets the threshold at the expression value corresponding to the step[15]. An intermediate region can be defined around the threshold using a width of 1 (0.5 below and 0.5 above the threshold), corresponding to a 2-fold change in expression, which is the minimum noise level in these large datasets [15, 16]. All the data below the intermediate region (< 1st StepMiner threshold - 0.5) are considered negative, and all above the intermediate region (> 1st StepMiner threshold + 0.5) are considered positive. When gene-expression levels display a large dynamic range, the StepMiner algorithm can be used to calculate two distinct thresholds: a first threshold to discriminate between "negative" and "positive" samples (1st StepMiner threshold) and a second threshold to split "positive" samples into two subgroups with "low" and "high" gene expression levels (2nd StepMiner threshold) (Supplementary Fig. 20).

We started our search for developmentally regulated genes on our annotated "*human colon global database*", containing 1684 samples (Supplementary Table 1). To minimize the risk that results might be affected by samples containing significant contaminations from tissues other than colorectal epithelium (e.g. normal liver tissue in hepatic metastases), we restricted our investigation on the subset of arrays whose gene-expression profile could be defined as $EpCAM^{+}/Albumin^{neg}$ (Supplementary Fig. 6). Threshold gene expression levels were calculated using the StepMiner algorithm, based on the full set of 1684 arrays of the "*human colon global database*" ($EpCAM^{+}$ defined as Affymetrix probe 201839_s_at >10.05; $Albumin^{neg}$ defined as Affymetrix probe 211298_s_at <7.97). This operation removed 116 arrays (6.9%) and left 1568 arrays (93.1%) for subsequent analysis (normal colon mucosa: n = 170; colorectal adenoma: n =

68; colorectal carcinoma: n = 1330). We then systematically computed Boolean implication relationships between pairs of genes, using the BooleanNet software [16]. Mature enterocyte genes were predicted based on genes highly expressed in the $KRT20^{+}$ group of arrays and filtered based on their fulfillment of the "X^{+} implies $KRT20^{+}$" Boolean implication (Supplementary Fig. 7). Goblet genes were predicted based on genes highly expressed in the $MUC2^{+}$ group of arrays and filtered based on their fulfillment of at least one of three independent Boolean implications: a) "MUC2 is equivalent to X", b) "X^{+} implies $MUC2^{+}$", c) "$MUC2^{+}$ implies X^{+}" (Supplementary Fig. 8). Immature genes were predicted based on genes highly expressed in the $KRT20^{neg}$ group of arrays, and additionally filtered based on their fulfillment of the "$KRT20^{neg}$ implies X^{+}" Boolean implication (Supplementary Fig. 9). Threshold gene expression levels were calculated using the StepMiner algorithm, based on our total pool of 46,047 publicly available human gene-expression arrays, obtained from three distinct platforms: Affymetrix U133 Plus 2.0, Affymetrix U133A and Affymetix U133A 2.0. Gene-expression patterns were considered to fulfill a specific Boolean implication when the false-discovery rate (FDR) of a sparsity test in the relevant quadrant was < 0.05 [16].

Differences in the expression levels of individual genes among different sample subgroups (i.e. normal vs adenoma, $KRT20^{neg}$ vs $KRT20^{+}$ carcinomas) were evaluated using box-plots [17] and tested for statistical significance using a 2-sample t-test (2-tailed). Correlation between the gene-expression levels of two genes (Supplementary Fig. 18) was measured using Pearson correlation coefficients.

Stratification of human colon cancer patients in distinct gene-expression groups and survival analysis using the "Hegemon" software. To evaluate whether genes identified by SINCE-PCR as differentially expressed during normal colon differentiation (e.g. KRT20, CA1, MS4A12, CD177, SLC26A3) could be used as novel prognostic markers, we developed a novel bioinformatic tool to explore gene-expression datasets annotated with patient survival data. We named this tool "Hegemon" as an acronym for "hierarchical exploration of gene expression microarrays on-line". The Hegemon software is an upgrade of the BooleanNet software, where individual gene-expression arrays, after being plotted on a two-axis chart based on the expression levels of two given genes [16], can now be automatically compared for survival outcomes using Kaplan-Meier survival curves. The hypothesis behind this approach is that, on average, a tumor's overall gene expression profile would most closely resemble that of the most abundant cellular population, and that tumors highly enriched in more mature, terminally differentiated cell types would be characterized by a lower proliferation rate and/or a lower content of long-term self-renewing cells, thus being associated to a better prognosis as compared to tumors predominantly composed by immature, progenitor-like cells.

Survival analysis was performed on a gene-expression database which contains disease-free survival (DFS) information on 299 patients of different clinical stages (AJCC Stage I-IV/Duke's Stage A-D) from three independent institutions: H. Lee Moffit Cancer Center (n = 164), Vanderbilt Medical Center (n = 55) and Royal Melbourne Hospital (n = 80). This database was created by pooling information from two publicly available GEO datasets (GSE14333, GSE17538; see also Supplementary Table 1) [12, 14]. All samples contained in these three datasets were analyzed using the Affymetrix U133

Plus 2.0 platform and were carefully annotated with disease-free survival (DFS) information. To avoid bias due to redundancies (i.e. identical samples deposited in both GEO datasets) we cross-checked all samples and removed duplicates.

Based on SINCE-PCR data, we selected four genes whose expression is largely restricted to "top-of-the-crypt" $CA1^{+}/SLC26A3^{+}$ cells (i.e. CA1, MS4A12, CD177, SLC26A3) as markers of terminal differentiation, and KRT20, whose expression is observed in both "top-of-the-crypt" $CA1^{+}/SLC26A3^{+}$ cells and a subset of $MUC2^{+}/TFF3^{high}$ goblet-type cells, as a more promiscuous marker of both intermediate and terminal differentiation. Threshold gene expression levels were calculated using the StepMiner algorithm, based on the 25,576 arrays on the human Affymetrix U133 Plus 2.0 platform. KRT20 expression (Affymetrix probe 213953_at) was tested as a marker to separate poorly differentiated tumors ($KRT20^{neg}$) from differentiated ones ($KRT20^{+}$). Based on our previous experience with the StepMiner algorithm [15], we defined as $KRT20^{neg}$ all tumors whose KRT20 expression values were < 1st StepMiner threshold – 0.5 (Affymetrix probe 213953_at < 7.00). Genes expressed in "top-of-the-crypt" $CA1^{+}/SLC26A3^{+}$ cells (CA1, MS4A12, CD177, SLC26A3) were tested as markers to separate terminally differentiated tumors ($top\text{-}crypt^{high}$) from moderately differentiated ones ($top\text{-}crypt^{neg/low}$). In the case of CD177 (Affymetrix probe 219669_at) and SLC26A3 (Affymetrix probes 215657_at), the sensitivity of the probe appeared lower and its dynamic range narrower as compared to CA1 (Affymetrix probe 205950_s_at) or MS4A12 (Affymetrix probe 220834_at) (Supplementary Fig. 7). In order to maintain consistency in the selection of sample subsets with highest expression levels, we adopted a scaled approach based to match the different sensitivity of the individual gene-

expression probes (Supplementary Fig. 20). In the case of CD177 and SLC26A3 we chose to simply separate negative samples from positive ones ($CD177^{neg}$ vs $CD177^{+}$, $SLC26A3^{neg}$ vs $SLC26A3^{+}$, respectively), while in the case of CA1 and MS4A12 we chose to separate high-expression samples from low-to-negative expression ones ($CA1^{neg/low}$ vs $CA1^{high}$, $MS4A12^{neg/low}$ vs $MS4A12^{high}$, respectively). As a result, when we tested CD177 or SLC26A3 we defined as top-$crypt^{high}$ all tumors that scored as $CD177^{+}$ or $SLC26A3^{+}$ (defined as expression values > 1st StepMiner threshold + 0.5; CD177: Affymetrix probe 219669_at > 8.14; SLC26A3: Affymetrix probe 215657_at > 5.43), and when we tested CA1 or MS4A12 we defined as mature top-$crypt^{high}$ all tumors that scored as $CA1^{high}$ or $MS4A12^{high}$ (defined as expression values > 2nd StepMiner threshold; CA1: Affymetrix probe 205950_s_at > 11.14; MS4A12: Affymetrix probe 220834_at > 9.27).

Based on these definitions, we stratified human colon cancer samples into three "*gene-expression groups*": Group 1 ($KRT20^{+}$/top-$crypt^{high}$), Group 2 ($KRT20^{+}$/top-$crypt^{neg/low}$), Group 3 ($KRT20^{neg}$/top-$crypt^{neg/low}$). As predicted by the strong Boolean relationship linking KRT20 to all mature enterocyte genes (Supplementary Fig. 7), no tumors were observed that corresponded to the theoretical fourth group ($KRT20^{neg}$/top-$crypt^{high}$) with only the exception of one isolated single sample in the KRT20/SLC26A3 experiment. Once grouped based on gene-expression thresholds, patient subsets were compared for survival outcomes, using both Kaplan-Meier survival curves and multivariate analysis based on the Cox proportional hazards method. Differences in Kaplan-Meier curves were tested for statistical significance using the Log-rank test. In experiments involving comparions to the EphB2 "intestinal stem cell" (ISC) signature (Supplementary Fig. 23),

colon cancer patients were grouped in three categories (ISC^{low}, ISC^{medium}, ISC^{high}) as described in Merlos-Suarez *et al.* [18].

Studies on the association between gene-expression groups and other pathological or molecular variables (e.g. pathological grading, MSI/MSS status) were performed on appropriately selected subsets of the "*human colon global database*" (Supplementary Table 1). Enrichment of selected pathological or molecular features, such as high pathological grade (G3-G4) or microsatellite instability (MSI), in groups characterized by immature gene-expression patterns (e.g. Group 3, $KRT20^{neg}$/top-$crypt^{neg/low}$) was measured using odds-ratios (OR) and tested for significance using Pearson's χ^2 test.

SUPPLEMENTARY REFERENCES

1. Dalerba, P. *et al.* Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. USA,* **104**:10158-10163 (2007).
2. Szczesny, G., Veihelmann, A., Massberg, S., Nolte, D. & Messmer, K. Long-term anaesthesia using inhalatory isoflurane in different strains of mice-the haemodynamic effects. *Laboratory Animals,* **38**:64-69 (2004).
3. Dalerba, P. *et al.* Reconstitution of human telomerase reverse transcriptase expression rescues colorectal carcinoma cells from in vitro senescence: evidence against immortality as a constitutive trait of tumor cells. *Cancer Res.,* **65**:2321-2329 (2005).
4. Ringner, M. What is principal component analysis? *Nat. Biotechnol.,* **26**:303-304 (2008).
5. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell,* **18**:675-685 (2010).
6. O'Doherty, U., Swiggard, W.J. & Malim, M.H. Human immunodeficiency virus type 1 spinoculation enhances infection through virus binding. *Journal of Virology,* **74**:10074-10080 (2000).
7. Wang, G.P. *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.,* **36**:e49 (2008).
8. Mitchell, R.S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.,* **2**:E234 (2004).
9. Ishizawa, K. *et al.* Tumor-Initiating Cells Are Rare in Many Human Tumors. *Cell Stem Cell,* **7**:279-282 (2010).
10. Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature,* **456**:593-598 (2008).
11. Sutherland, H.J., Lansdorp, P.M., Henkelman, D.H., Eaves, A.C. & Eaves, C.J. Functional characterization of individual human hematopoietic stem cells cultured at limiting dilution on supportive marrow stromal layers. *Proc. Natl. Acad. Sci. USA,* **87**:3584-3588 (1990).
12. Jorissen, R.N. *et al.* Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clin. Cancer Res.,* **15**:7642-7651 (2009).
13. Khambata-Ford, S. *et al.* Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J. Clin. Oncol.,* **25**:3230-3237 (2007).

14. Smith, J.J. *et al.* Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, **138**:958-968 (2010).

15. Sahoo, D., Dill, D.L., Tibshirani, R. & Plevritis, S.K. Extracting binary signals from microarray time-course data. *Nucleic. Acids Res.*, **35**:3705-3712 (2007).

16. Sahoo, D., Dill, D.L., Gentles, A.J., Tibshirani, R. & Plevritis, S.K. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biology*, **9**:R157 (2008).

17. Williamson, D.F., Parker, R.A. & Kendrick, J.S. The box plot: a simple visual method to interpret data. *Annals of Internal Medicine*, **110**:916-921 (1989).

18. Merlos-Suarez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511-524 (2011).