

Single-cell dissection of transcriptional heterogeneity in human colon tumors

Piero Dalerba^{1,2,9}, Tomer Kalisky^{3,9}, Debashis Sahoo^{1,9}, Pradeep S Rajendran¹, Michael E Rothenberg^{1,4}, Anne A Leyrat³, Sopheak Sim¹, Jennifer Okamoto^{3,5}, Darius M Johnston^{1,3,5}, Dalong Qian¹, Maider Zabala¹, Janet Bueno⁶, Norma F Neff³, Jianbin Wang³, Andrew A Shelton⁷, Brendan Visser⁷, Shigeo Hisamori¹, Yohei Shimonono¹, Marc van de Wetering⁸, Hans Clevers⁸, Michael F Clarke^{1,2,9} & Stephen R Quake^{3,5,9}

Cancer is often viewed as a caricature of normal developmental processes, but the extent to which its cellular heterogeneity truly recapitulates multilineage differentiation processes of normal tissues remains unknown. Here we implement single-cell PCR gene-expression analysis to dissect the cellular composition of primary human normal colon and colon cancer epithelia. We show that human colon cancer tissues contain distinct cell populations whose transcriptional identities mirror those of the different cellular lineages of normal colon. By creating monoclonal tumor xenografts from injection of a single ($n = 1$) cell, we demonstrate that the transcriptional diversity of cancer tissues is largely explained by *in vivo* multilineage differentiation and not only by clonal genetic heterogeneity. Finally, we show that the different gene-expression programs linked to multilineage differentiation are strongly associated with patient survival. We develop two-gene classifier systems (*KRT20* versus *CA1*, *MS4A12*, *CD177*, *SLC26A3*) that predict clinical outcomes with hazard ratios superior to those of pathological grade and comparable to those of microarray-derived multigene expression signatures.

The *in vivo* cellular composition of solid tissues is often difficult to investigate in a comprehensive and quantitative way. Techniques such as immunohistochemistry and flow cytometry are limited by the availability of antigen-specific monoclonal antibodies and by the small number of parallel measurements that can be performed on each individual cell. Traditional high-throughput assays, such as gene-expression arrays, when performed on whole tissues, provide information on average gene expression levels, and can be correlated only indirectly to quantitative modifications in cellular subpopulations. These limitations become particularly difficult to overcome when studying minority populations, such as stem cells, whose identification is made elusive by their low numbers and by the lack of exclusive markers. Moreover, in pathological states, such as cancer, it is usually impossible to determine whether perturbations in gene expression detected in whole tissues are due to modifications in the relative composition of different cell types or to aberrations in the gene-expression profile of mutated cells.

For example, although it has been postulated that multilineage differentiation can contribute to tumor heterogeneity^{1–3}, this issue remains controversial⁴. Many in the field view cancer heterogeneity mainly as the result of clonal evolution secondary to genomic instability^{5,6}. Previous studies addressed this question, but could rely only on *in vitro* cultured cell lines and on simple morphological evidence^{7–9}.

Moreover, recent evidence indicates that, in the absence of a molecular proof of monoclonal origin, results from *in vitro* experiments based on limiting dilution can be biased due to a dramatic increase in cell survival by cell hetero-doublings. This phenomenon is best exemplified in the case of the mouse small intestine, where growth and expansion of *LGR5*⁺ progenitor cells is dramatically enhanced by the presence of bystander epithelial feeder cells¹⁰. Based on these studies, it remained difficult to perform a quantitative measure of the degree of multilineage differentiation in cancer tissues and, above all, to investigate to what degree it actually translated into the differential activation of distinct transcriptional programs that would mirror and recapitulate the physiological processes observed in normal tissues. In this study we developed a method to dissect and investigate at the single-cell level the gene-expression profile of the distinct cell populations contained in primary human colon epithelia, both normal and neoplastic.

RESULTS

Description and technical validation of single-cell PCR

We combined fluorescence activated cell sorting (FACS) and single-cell PCR gene-expression analysis to perform a high-throughput transcriptional analysis of the distinct cellular populations contained in solid human tissues (**Supplementary Figs. 1 and 2**). This method exploits the capacity of modern flow cytometers to sort

¹Stanford Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA. ²Department of Medicine, Division of Oncology, Stanford University, Stanford, California, USA. ³Department of Bioengineering, Stanford University, Stanford, California, USA. ⁴Department of Medicine, Division of Gastroenterology and Hepatology, Stanford University, Stanford, California, USA. ⁵Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. ⁶Tissue Bank, Stanford University, Stanford, California, USA. ⁷Department of Surgery, Stanford University, Stanford, California, USA. ⁸Hubrecht Institute for Developmental Biology and Stem Cell Research, Utrecht, The Netherlands. ⁹These authors contributed equally to this work. Correspondence should be addressed to S.R.Q. (quake@stanford.edu) or M.F.C. (mfclarke@stanford.edu).

Received 2 May; accepted 12 October; published online 13 November 2011; doi:10.1038/nbt.2038

individual single cells with accuracy and precision (**Supplementary Fig. 3**), together with the use of microfluidic technologies to perform high-sensitivity multiplexed PCR from minute amounts of mRNA, thereby allowing parallel analysis of the expression of up to 96 genes for each individual cell. The large number of measurements per cell and the possibility to analyze several hundred cells in parallel from the same sample allow the use of statistical clustering algorithms to associate cells with similar gene expression profiles into well-defined subpopulations (**Supplementary Fig. 2**). Microfluidic platforms have been previously validated for single-cell gene-expression analysis^{11–13}. Consistent with those results, our control experiments with titrated mRNA standards as well as single-cell experiments on a cell line validated the sensitivity of this approach for high-throughput analysis across multiple genes (**Supplementary Fig. 4**).

Analysis of normal human colon epithelium

We first applied single-cell PCR to the study of normal human colon epithelial cells. Human colon epithelium is composed of heterogeneous populations of cells that express different protein markers based on their lineage, differentiation stage and functional status. Many of these cell subsets can be identified by immunohistochemistry against well-characterized markers, such as MUC2, expressed by goblet cells; MKI67, expressed by proliferating cells; KRT20 and CEACAM1 (also known as CD66a), preferentially expressed by cells at the top of the colonic crypt (**Fig. 1a–d**)¹⁴.

Under normal conditions, immature colon epithelial cells reside at the bottom of the colonic crypts (bottom-of-the-crypt cells) and express high levels of the surface marker CD44, whereas differentiated mature cells progressively migrate to the top (top-of-the-crypt cells) and progressively lose CD44 expression^{14,15}. We focused our analysis on the stem and progenitor cell compartments of the colonic epithelium by sorting the EpCAM^{high}/CD44⁺ population (**Fig. 1e,f**; P12) which, in normal tissues, corresponds to the bottom of the human colonic crypt¹⁴. To study the more mature, terminally differentiated cell populations, we sorted and analyzed an equal number of cells from the EpCAM⁺/CD44⁻/CD66a^{high} population, which corresponds to the top of the human colonic crypt (**Fig. 1e,f**; P11)¹⁶.

We first tested the ability of single-cell PCR gene-expression analysis to distinguish different cell populations using well-established reference markers. We analyzed and clustered colon epithelial cells using three genes encoding markers linked to either one of the two major cell lineages (that is, MUC2 for goblet cells and CA1 for enterocytes) or the immature compartment (that is, LGR5) of the colon epithelium^{14,17–19}. This experiment showed that genes encoding lineage-specific markers are frequently expressed in a mutually exclusive way, mirroring the expression pattern of corresponding proteins (**Supplementary Fig. 5**).

We then searched for gene-expression markers of the different cell populations, with a special focus on putative stem cell markers. We mined 1,568 publicly available gene-expression array data sets from human colon epithelia (**Supplementary Table 1**), using a bioinformatics approach designed to identify developmentally regulated genes based on Boolean implication logic (**Supplementary Fig. 6**)²⁰. The search yielded candidate genes whose expression was associated with that of other markers previously linked to individual colon epithelial cell lineages (**Supplementary Figs. 7–9**). Using an iterative approach, we screened >230 genes on eight independent samples of normal human colon epithelium by single-cell PCR gene-expression analysis. At each round, genes that were noninformative (that is, not differentially expressed in either positive or negative association with CA1, MUC2 or LGR5) were removed 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 (3 housekeeping, 3 proliferation-related and 47 differentially expressed genes; **Supplementary Table 2**) with high robustness (**Supplementary Fig. 10**). This allowed us to characterize multiple cell populations, using both hierarchical clustering (**Fig. 1g**) and principal component analysis (PCA; **Fig. 1h,i**).

Analysis of the EpCAM⁺/CD44⁻/CD66a^{high} population (enriched for top-of-the-crypt cells) revealed that this subset, although transcriptionally heterogeneous, was almost exclusively composed of cells expressing high levels of genes characteristic of mature enterocytes (e.g., CA1⁺, CA2⁺, KRT20⁺, SLC26A3⁺, AQP8⁺ and MS4A12⁺)^{14,21–23} and led to the discovery of at least two gene expression markers whose differential expression pattern—to our knowledge—has not been reported before (CD177 and GUCA2B) (**Fig. 1g**). To validate the reliability of single-cell PCR gene-expression analysis results, we evaluated the distribution of SLC26A3 and CD177 protein expression in tissue sections and we confirmed its preferential expression at the top of the human colonic crypts (**Supplementary Figs. 11 and 12**).

We could also distinguish different subsets of cells with different transcriptional profiles within the EpCAM⁺/CD44⁻/CD66a^{high} population (e.g., CA1⁺/SLC26A3⁺ versus GUCA2B⁺). At the present time, it is not clear whether they represent distinct stages of differentiation or distinct functional subsets of colonic enterocytes. Nonetheless, their clearly unique transcriptional programs identify them as part of a distinct cellular population.

Analysis of the EpCAM^{high}/CD44⁺ population (enriched for 'bottom-of-the-crypt' cells) revealed the presence of multiple populations, including: (i) a cell compartment characterized by the expression of genes linked to goblet cells (MUC2⁺, TFF3^{high}, SPDEF⁺, SPINK4⁺)^{24,25}, (ii) a cell compartment characterized by the co-expression of genes associated with immature cells as well as genes known to be expressed by enterocytes (OLFM4⁺, CA2^{high}) and (iii) a cell compartment whose gene-expression profile mirrors that of a stem and/or progenitor cell compartment in the mouse small intestine (LGR5⁺, ASCL2⁺, PTPRO⁺, RGMB⁺)^{17,26}. A synopsis of the key genes that define the gene-expression profile of the different populations is provided in **Supplementary Table 3**.

The OLMF4⁺/CA2^{high} and the LGR5⁺/ASCL2⁺ compartments shared expression of several genes of functional interest in both stem cell and cancer biology, such as genes involved in self-renewal and chromatin remodeling (EZH2, BMI1)^{27–29}, Wnt-pathway signaling (AXIN2)³⁰, cell growth and chemotaxis (CXCL2)³¹, stem cell quiescence (LRIG1)³² and oncogenes (MYC)³³. The expression of proliferation markers, such as MKI67, TOP2A, BIRC5 (also known as *Survivin*) appeared to be restricted to the EpCAM^{high}/CD44⁺ (bottom-of-the-crypt) population and particularly to the LGR5⁺/ASCL2⁺ and MUC2⁺/TFF3^{high} cells. This was partially expected based on both previously published data^{14,17,19} and our own immunohistochemistry results (**Supplementary Fig. 13c**).

We also observed that MUC2⁺/TFF3^{high} cells were characterized by high expression levels of several genes of interest, including DLL1 and DLL4, encoding for two Notch ligands, and KRT20. The expression of KRT20 at the bottom of the crypt appears contrary to the notion of KRT20 as a terminal differentiation marker. However, a more careful examination of immunohistochemical stainings identified scattered KRT20⁺ cells, which can be morphologically identified as goblet cells (**Supplementary Fig. 13a,b**). We also noticed that MUC2⁺/TFF3^{high} cells, for the most part, did not express CFTR, the gene mutated in cystic fibrosis. The differential expression of DLL4 is of potential relevance to the clinical development of novel anti-tumor therapeutic agents directed against this molecule³⁴.

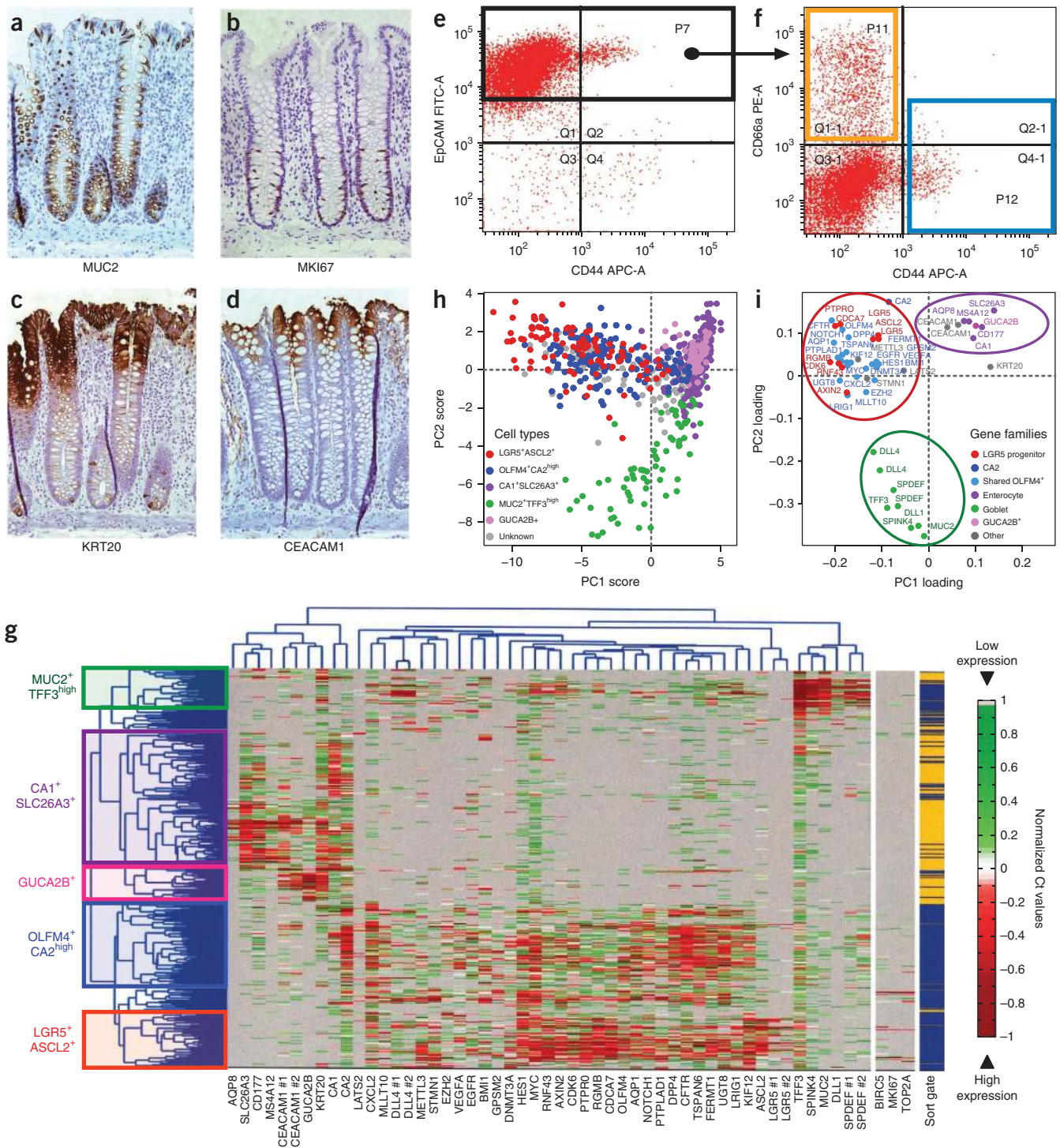


Figure 1 Single-cell PCR gene-expression analysis of human normal colon epithelium. (a–d) Immunohistochemistry of normal human colon epithelium, stained for MUC2 (a), labeling goblet cells, MKI67 (b), labeling proliferating cells, KRT20 (c) and CEACAM1 (d), preferentially labeling top-of-the-crypt cells. (e,f) Flow cytometry sorting strategy for top-of-the-crypt and bottom-of-the-crypt epithelial cells. (e) Colon epithelial cells, both CD44^{neg} and CD44⁺, were separated from stromal cells based on their EpCAM⁺ phenotype. (f) Bottom-of-the-crypt epithelial cells were defined as EpCAM^{high}/CD44⁺ (f, P12 blue sort gate) and top-of-the-crypt epithelial cells as EpCAM⁺/CD44⁻/CD66a^{high} (f, P11 orange sort gate). (g) Hierarchical clustering of single-cell PCR gene-expression analysis data visualized distinct cell populations, including enterocyte-like cells (CA1⁺/SLC26A3⁺ and GUCA2B⁺), goblet-like cells (MUC2⁺/TFF3^{high}) and two compartments defined by gene-expression profiles reminiscent of more immature progenitors (OLFM4⁺/CA2^{high} and LGR5⁺/ASCL2⁺). (h,i) Principal component analysis of single-cell PCR gene-expression data visualized different cell types and different gene families. Different cell types were characterized by different scores along the two main principal components (PC1 and PC2) (h). Different gene families were characterized by different contributions to the two main principal components. To allow comparisons between hierarchical clustering and PCA results, we displayed each cell or gene in PCA plots with the color corresponding to the cell type or gene family it was assigned to based on hierarchical clustering (i).

Analysis of a primary human colon adenoma

We then turned to cancer and investigated whether the cellular composition of the normal colonic epithelium is preserved in colorectal tumors, both benign and malignant. Analysis by single-cell PCR gene-expression analysis of EpCAM^{high}/CD44⁺ cells from a primary tubulo-villous adenoma (sample name: SU-COLON#76; **Supplementary Table 4**) revealed the presence

of at least two different cell populations (that is, *LGR5*⁺/*ASCL2*⁺ and *MUC2*⁺/*TFF3*^{high}) characterized by distinctive gene signatures, closely mirroring the subpopulations observed in corresponding EpCAM^{high}/CD44⁺ populations of normal tissues (**Fig. 2a–c**).

These observations were confirmed at the protein level by parallel immunohistochemical investigations for KRT20 and MUC2 (**Fig. 2d,e**) and are in agreement with the recent finding that

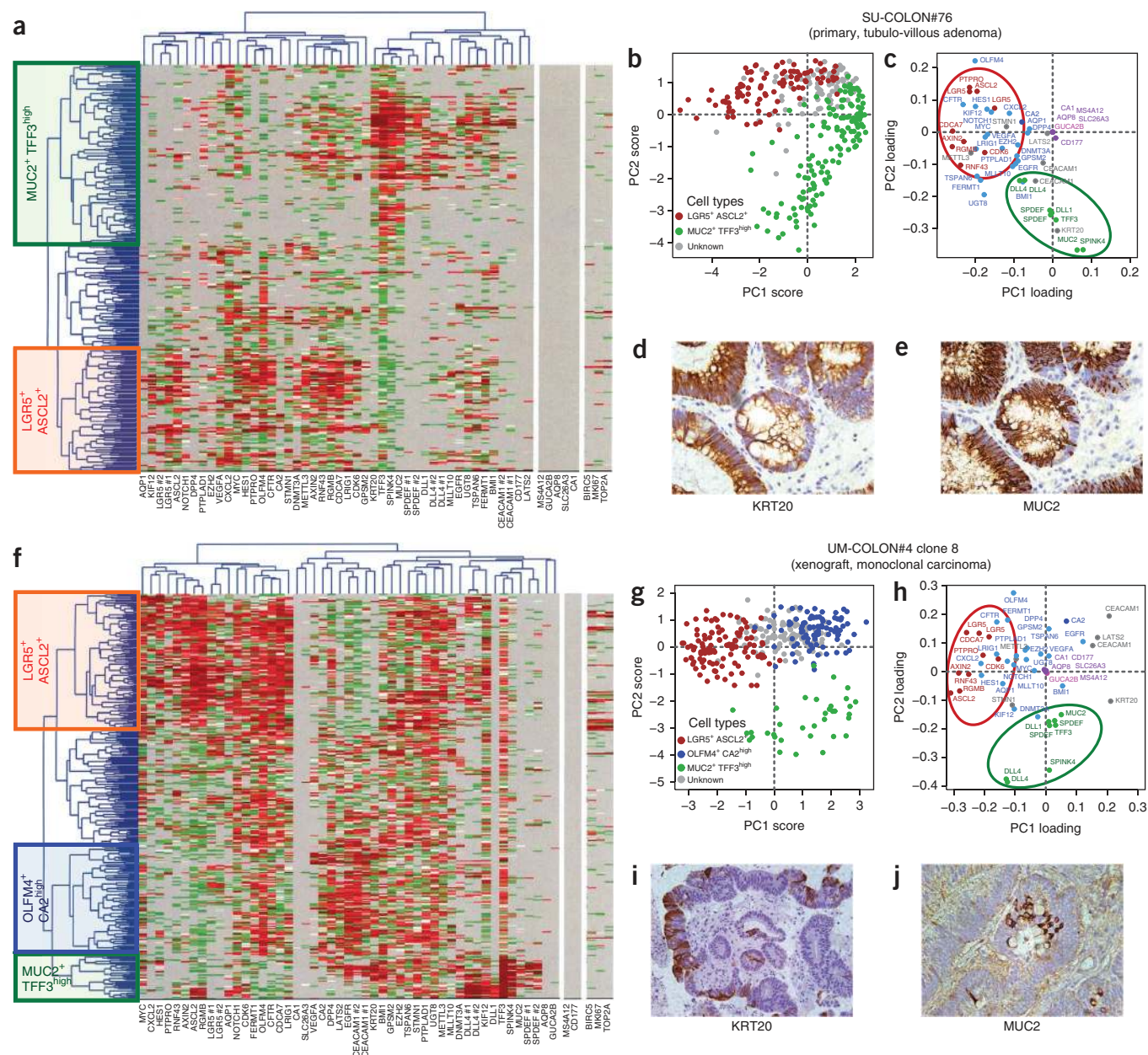
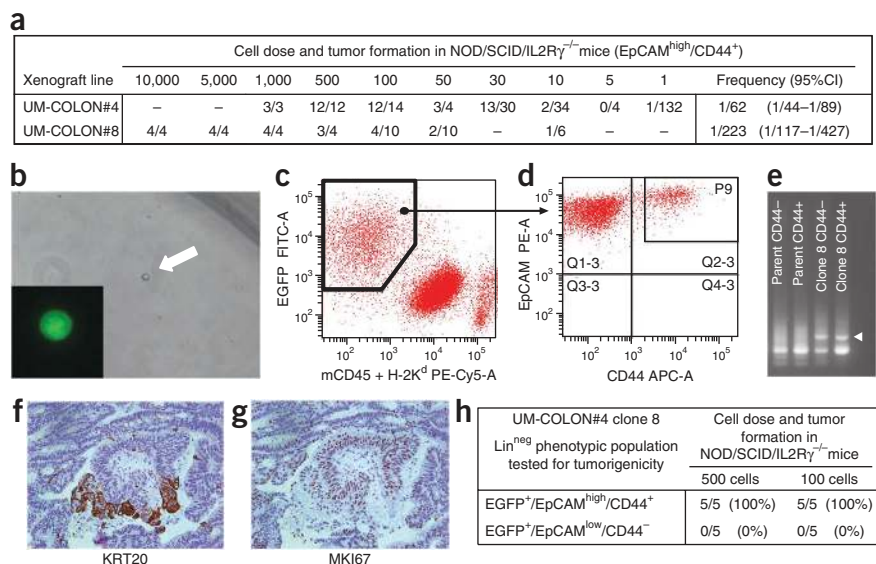


Figure 2 Single-cell PCR gene-expression analysis of human colon tumor tissues. **(a)** Hierarchical clustering of single-cell PCR gene-expression data from the EpCAM⁺/CD44⁺ population of a large primary benign adenoma (sample: SU-COLON#76; see **Supplementary Table 4**). The analysis revealed the presence of multiple cell populations characterized by distinct gene signatures, closely mirroring lineages and differentiation stages observed in the EpCAM⁺/CD44⁺ population from the normal colon epithelium. **(b,c)** Principal component analysis (PCA) of single-cell PCR gene-expression analysis data confirmed hierarchical clustering results, visualizing cell types **(b)** and gene families **(c)** similar to those identified in normal tissues. **(d,e)** Gene-expression data were confirmed at the protein level by immunohistochemistry, testing for expression of KRT20 **(d)** and MUC2 **(e)** on corresponding tissue sections. **(f–j)** A similar study on a monoclonal colon cancer xenograft obtained from injection of a single ($n = 1$) cell in a NOD/SCID/IL2R $\gamma^{-/-}$ mouse (UM-COLON#4 clone 8) produced similar results in terms of hierarchical clustering **(f)**, cell types identified by PCA **(g)**, gene families identified by PCA **(h)**, immunohistochemistry results for KRT20 **(i)** and immunohistochemistry results for MUC2 **(j)**. Results from the monoclonal tumor xenograft indicated that the distinct cell populations visualized by single-cell PCR did not arise as the result of the coexistence within the tumor tissue of independent genetic subclones, but as the result of multilineage differentiation processes during tumor growth. Color coding of normalized threshold cycle (Ct) values in hierarchical clustering plots and of gene families in PC loading plots are identical to those of **Figure 1**.

Figure 3 Analysis of a monoclonal human colon cancer xenograft obtained from injection of a single ($n = 1$) cell in NOD/SCID/IL2R $\gamma^{-/-}$ mice. **(a)** In human colon cancer, the frequency of EpCAM^{high}/CD44⁺ cells capable to establish a tumor upon xenotransplantation in NOD/SCID/IL2R $\gamma^{-/-}$ mice varies based on the xenograft line, as shown by comparative limiting-dilution experiments. **(b)** Single ($n = 1$) lentivirus-infected EGFP⁺/EpCAM^{high}/CD44⁺ cancer cells can be sorted by flow cytometry for injection in mice. **(c,d)** Analysis by flow cytometry of a monoclonal tumor derived from injection of a single ($n = 1$), lentivirus-tagged, EGFP⁺/EpCAM^{high}/CD44⁺ cancer cell from the human colon cancer xenograft UM-COLON#4 (clone 8) confirmed that human cells expressed EGFP **(c)** and contained both EpCAM^{low}/CD44⁻ and EpCAM^{high}/CD44⁺ populations **(d)**. **(e)** The monoclonal origin of the UM-COLON#4 clone 8 tumor was confirmed by LM-PCR, showing the presence of a unique lentivirus integration site in both EGFP⁺/EpCAM^{low}/CD44⁻ and EGFP⁺/EpCAM^{high}/CD44⁺ populations, contrary to what was observed in its polyclonal parent tumor. A larger image of the LM-PCR gel is provided in **Supplementary Figure 24**. **(f,g)** Immunohistochemistry of monoclonal tumor tissues revealed heterogeneous and mutually exclusive expression patterns of KRT20 **(f)** and MKI67 **(g)**. **(h)** Similar to what is observed in parent tumors, EpCAM^{high}/CD44⁺ and EpCAM^{low}/CD44⁻ populations from UM-COLON#4 clone 8 were characterized by different tumorigenic capacity, as evaluated by tumorigenicity experiments in NOD/SCID/IL2R $\gamma^{-/-}$ mice.



KRT20 is frequently expressed in a mutually exclusive pattern with respect to *LGR5* (ref. 19). This primary adenoma appeared depleted in *CA1⁺/SLC26A3⁺*, *GUCA2B⁺* and *OLFM4⁺/CA2^{high}* cell populations. A careful examination of public gene-expression array databases indicated that this unexpected feature is likely common to many benign adenomas (**Supplementary Fig. 14**).

Analysis of a human colon cancer xenograft derived from a single cancer cell

Tumor tissues, both benign and malignant, are known to undergo perturbations of normal differentiation processes, but it is unclear to what extent those perturbations reflect quantitative changes in cell composition or qualitative changes in gene-expression programs. This topic has historically been controversial^{4–9,35}. Our own systematic study of KRT20 and MUC2 protein expression in human malignant colorectal cancer tissues, for instance, revealed that both markers are frequently expressed heterogeneously, in patterns that mirror those observed in normal colorectal epithelium (**Supplementary Fig. 15**). It remained unclear, however, to what extent cancer transcriptional heterogeneity is the result of clonal genetic heterogeneity³⁶ or epigenetic heterogeneity due to multilineage differentiation processes⁹.

To address this question from a functional perspective, we investigated whether a single ($n = 1$) human colorectal cancer cell can recreate the heterogeneous cell composition of parent tumor tissues, including the subpopulations that we discovered in this study. We injected NOD/SCID/IL2R $\gamma^{-/-}$ mice with single ($n = 1$) EpCAM^{high}/CD44⁺ cancer cells purified by flow cytometry from one of our well-characterized solid xenograft lines³⁷, following infection with a lentivirus vector encoding enhanced green fluorescence protein (EGFP; **Fig. 3a,b**).

Notably, the single cell-derived, lentivirus-tagged, EGFP⁺ xenograft line generated in this experiment (UM-COLON#4 clone 8) closely reproduced the phenotypic diversity of its parent tumor both in terms of tissue histology (**Figs. 2i,j** and **3f,g**) and surface-marker phenotypic repertoire of cellular populations (**Fig. 3c,d**). The line's monoclonal origin was confirmed by identification of a unique lentivirus integration site in all cancer cells (**Fig. 3e**).

Tumorigenicity experiments done in NOD/SCID/IL2R $\gamma^{-/-}$ mice revealed that, as observed in the parent tumors³⁷, EGFP⁺/EpCAM^{high}/CD44⁺ and EGFP⁺/EpCAM^{low}/CD44^{-/low} cell populations were endowed with different tumorigenic capacity (**Fig. 3h**). A single-cell PCR gene-expression analysis of the EpCAM^{high}/CD44⁺ population from these monoclonal tumors demonstrated its heterogeneous lineage composition, showing the presence of three distinct compartments (that is, *LGR5⁺/ASCL2⁺*, *OLFM4⁺/CA2^{high}*, *MUC2⁺/TFF3^{high}*), again characterized by distinctive gene signatures, closely mirroring those observed in corresponding immature populations of normal tissues (**Fig. 2f–h**).

Taken together, these data formally prove that, in a subset of tumors, transcriptional heterogeneity is, at least partly, explained by multilineage differentiation processes that tend to recapitulate those observed in normal tissues.

Prognostic role of biomarkers identified by single-cell PCR

To gain further insight into the potential functional implications of these observations, we compared the gene-expression pattern of genes associated with cell proliferation (that is, *MKI67*, *TOP2A* and *BIRC5*) in normal and cancer tissues. In this case too, we observed that the expression pattern observed in malignant tissues frequently mirrored that of normal ones.

Both in the normal tissue and in the monoclonal human colon cancer xenograft, for instance, all three proliferation markers were frequently expressed in a mutually exclusive way as compared to the differentiation marker *KRT20* (**Supplementary Fig. 16**). This observation was subsequently confirmed at the protein level by a systematic study of MKI67 and KRT20 expression in serial sections from seven human colorectal cancer tissues, where MKI67 expression was often inversely associated with KRT20 (**Supplementary Fig. 17**).

These observations suggest that, in at least some cases, bulk short-term tumor growth is principally driven by a specific subset of the cancer cell population, characterized by a gene-expression repertoire characteristic of more immature cell compartments. This concept has important implications for the modeling of tumor growth kinetics

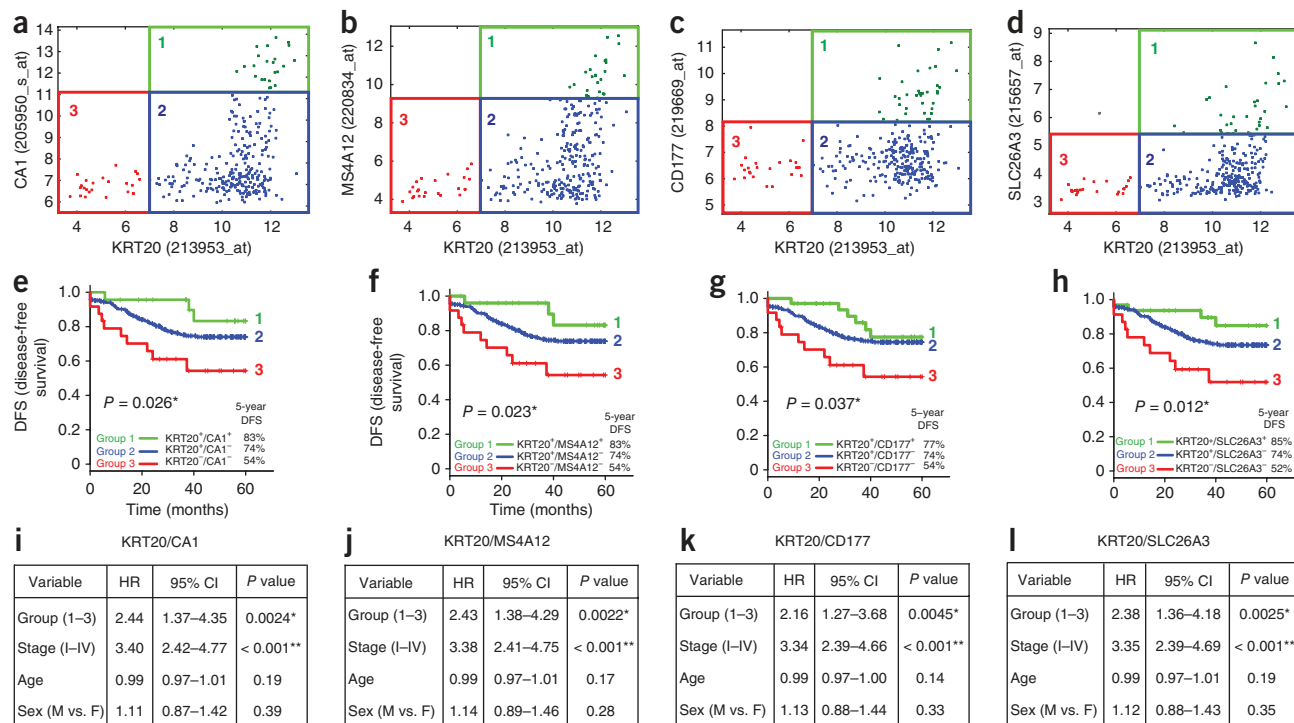


Figure 4 *KRT20* and *top-crypt* genes can be used as prognostic markers in colorectal cancer patients. (a–d) We used the Hegemon software to graph individual arrays according to the expression levels of *KRT20* and one of four genes characteristic of top-of-the-crypt $CA1^+/SLC26A3^+$ enterocyte-like cells: *KRT20* versus *CA1* (a), *KRT20* versus *MS4A12* (b), *KRT20* versus *CD177* (c), *KRT20* versus *SLC26A3* (d). We used the StepMiner algorithm to define gene-expression thresholds and identify three distinct gene-expression groups: Group 1 (green), defined as $KRT20^+/CA1^{high}$, $KRT20^+/MS4A12^{high}$, $KRT20^+/CD177^+$ or $KRT20^+/SLC26A3^+$, respectively; Group 2 (blue), defined as $KRT20^+/CA1^{low}$, $KRT20^+/MS4A12^{low}$, $KRT20^+/CD177^-$ or $KRT20^+/SLC26A3^-$, respectively; Group 3 (red), defined as $KRT20^-/CA1^{low}$, $KRT20^-/MS4A12^{low}$, $KRT20^-/CD177^-$ or $KRT20^-/SLC26A3^-$, respectively. (e–h) Survival analysis using Kaplan-Meier curves showed that, in all four cases, an increasingly immature gene-expression profile corresponded to a progressively worse prognosis. (i–l) Multivariate analysis of survival data based on the Cox proportional hazards model indicated that the prognostic effect of these two-gene classifiers was not confounded by clinical stage, age or sex. The analysis was performed on a pooled database of 299 primary colon cancer gene-expression arrays annotated with disease-free survival (DFS) data^{41,42} (Supplementary Table 1). * $P < 0.05$, ** $P < 0.001$. Age modeled as a continuous variable. HR, hazard ratio; CI, confidence interval; M, male; F, female.

and the response to anti-tumor drugs in different experimental settings. Although very common, this feature is not absolute, as we have observed exceptions characterized either by homogenous expression of *KRT20* in almost the entirety of the malignant epithelium or by complete absence of it in selected human tumors (Supplementary Fig. 17, samples SU87 and SU98, respectively). In accordance with our model, tumors characterized by the complete absence of *KRT20* expression were very poorly differentiated and contained high percentages of $MKI67^+$ cells (Supplementary Fig. 17, SU98).

We next tested whether these insights in the functional anatomy of the colon epithelium could have clinically useful applications. We evaluated whether quantitative expression levels of genes associated with differentiation processes could be used as a substitute measure for the cellular composition of the corresponding tumors and thereby serve to stratify colon cancer patients and predict clinical outcome. Our single-cell PCR gene-expression analysis data identified a set of sensitive and exclusive markers of top-of-the-crypt $CA1^+/SLC26A3^+$ cells (that is, *CA1*, *MS4A12*, *CD177*, *SLC26A3*). It also implicated *KRT20* as a more promiscuous differentiation marker, whose expression is high in $CA1^+/SLC26A3^+$ cells and a subset of $MUC2^+/TFF3^{high}$ cells, is absent in $LGR5^+/ASCL2^+$ cells, and is inversely associated with that of proliferation markers (*MKI67*, *TOP2A*, *BIRC5*). In addition, *KRT20* expression can be easily detected by immunohistochemistry and is commonly used

as a diagnostic marker in surgical pathology³⁸, thus representing an attractive candidate for further clinical applications³⁹.

Our first analysis of a pool of 1,568 independent human colon gene-expression arrays revealed that expression levels of genes characteristic for the $CA1^+/SLC26A3^+$ cell population are strongly correlated (Supplementary Fig. 18). The relationship between the expression of these top-of-the-crypt genes and *KRT20* was described by a Boolean implication: tumors expressing high levels of top-of-the-crypt genes (*top-crypt*^{high}) were always $KRT20^+$, whereas tumors expressing low-to-negative levels of top-of-the-crypt genes (*top-crypt*^{low}) could be clearly separated into two groups: $KRT20^+$ and $KRT20^-$ (Supplementary Fig. 7). Importantly, $KRT20^-$ tumors expressed high levels of *ALCAM/CD166* (Supplementary Fig. 19), a gene encoding for a surface marker characteristic of colon cancer cells with high tumorigenic potential in mouse xenotransplantation experiments³⁷.

We developed software ('hierarchical exploration of gene expression microarrays on-line', or Hegemon) to analyze the survival outcomes of human colon cancer patients after stratification into distinct gene-expression subsets, based on the expression of *KRT20* and one of the marker genes of $CA1^+/SLC26A3^+$ top-of-the-crypt cells (Fig. 4a–d). These subsets, or gene-expression groups, were numbered from more to less mature (group 1, $KRT20^+/top-crypt^{high}$; group 2, $KRT20^+/top-crypt^{low}$; group 3, $KRT20^-/top-crypt^{low}$). We used a computer-assisted method to determine the threshold level

Table 1 The prognostic effect of *KRT20/top-crypt* gene-expression groups

	HR ^a	95% CI ^b	P value
KRT20/CA1			
Prognostic variable			
Group (1–3) <i>KRT20/CA1</i>	2.93	1.37–6.27	0.0056*
Grade (G1–G4)	1.09	0.58–2.04	0.80
Stage (I–IV)	3.43	2.20–5.34	< 0.001**
Age ^c	0.99	0.97–1.01	0.43
Sex (M/F) ^d	1.18	0.86–1.61	0.31
KRT20/MS4A12			
Prognostic variable			
Group (1–3) <i>KRT20/MS4A12</i>	2.93	1.37–6.28	0.0057*
Grade (G1–G4)	1.07	0.57–2.00	0.84
Stage (I–IV)	3.41	2.19–5.31	<0.001**
Age ^c	0.99	0.97–1.01	0.41
Sex (M/F) ^d	1.19	0.87–1.63	0.28
KRT20/CD177			
Prognostic variable			
Group (1–3) <i>KRT20/CD177</i>	1.94	0.97–3.90	0.062
Grade (G1–G4)	1.19	0.63–2.22	0.59
Stage (I–IV)	3.21	3.03–7.06	<0.001**
Age ^c	0.99	0.97–1.01	0.39
Sex (M/F) ^d	1.20	0.87–1.64	0.26
KRT20/SLC26A3			
Prognostic variable			
Group (1–3) <i>KRT20/SLC26A3</i>	2.36	1.14–4.88	0.021*
Grade (G1–G4)	1.12	0.60–2.10	0.72
Stage (I–IV)	3.34	2.16–5.15	<0.001**
Age ^c	0.99	0.97–1.01	0.45
Sex (M/F) ^d	1.19	0.87–1.63	0.27

Multivariate analysis based on the Cox proportional hazards model, testing the *KRT20/top-crypt* two-gene scoring systems in parallel with pathological grading, clinical stage, age and sex, using the *KRT20/CA1* two-gene classifier, the *KRT20/MS4A12* two-gene classifier, the *KRT20/CD177* two-gene classifier or the *KRT20/SLC26A3* two-gene classifier. Contrary to pathological grade, *KRT20/top-crypt* gene expression groups were associated with statistically significant ($P < 0.05$) hazard ratios (HR), with the only exception of the *KRT20/CD177* two-gene classifier. The analysis was performed on a subset database of 181 microarrays annotated with grading information (database from ref. 42, $n = 181$, see **Supplementary Table 1**). *, $P < 0.05$; **, $P < 0.001$.

^aHR, hazard-ratio. ^bCI, confidence interval. ^cAge modeled as a continuous variable. ^dM/F, male versus female.

between positive and negative expression, based on the StepMiner algorithm (**Supplementary Fig. 20**)⁴⁰, and compared the clinical outcome of colon cancer patients in the three groups, using a pool of three independent data sets, containing 299 patients at different clinical stages (either AJCC stage I–IV or Dukes stage A–D) from the H. Lee Moffit Cancer Center, the Vanderbilt Medical Center and the Royal Melbourne Hospital^{41,42}, all of which were annotated with disease-free survival (DFS) data.

The three patient groups identified by these simple two-gene classifiers displayed substantially different clinical outcomes. An increasingly immature gene-expression profile corresponded to a progressively worse prognosis (**Fig. 4e–h**). This result was independent of the gene chosen as marker of *CA1*⁺/*SLC26A3*⁺ cells (that is, *CA1*, *MS4A12*, *CD177*, *SLC26A3*) and a multivariate analysis indicated that the prognostic value of the two-gene grouping system was not confounded by stage or other clinical variables (**Fig. 4i–l**).

Tumors with a more immature gene-expression profile (group 3, *KRT20*⁺/*top-crypt*^{-low}) were more likely to be of high pathological grade (G3–G4; **Supplementary Fig. 21**) and of microsatellite instability status (MSI; **Supplementary Fig. 22**). These enrichments, however, did not confound the prognostic value of the two-gene classifier system, as the high hazard-ratios associated with more immature gene-expression groups remained statistically significant ($P < 0.05$), when tested against pathological grade in multivariate analysis (**Table 1**;

with the exception of *KRT20/CD177*, $P = 0.06$), and because MSI⁺ tumors are known to be usually associated with a better prognosis⁴³. The prognostic effect of the two-gene classifier system was also independent of the recently described multigene EphB2 intestinal stem cell signature¹⁹, and was associated with comparable, if not superior, hazard ratios (**Supplementary Fig. 23**).

DISCUSSION

In this study, we implemented a method to investigate the cellular composition of solid tissues based on high-throughput parallel analysis of the gene-expression repertoire of single cells sorted by flow cytometry. We used this methodology to identify distinct cellular subsets of the human colon epithelium and to discover gene expression markers to define them. We then examined human colorectal tumors, both benign and malignant, and characterized them in terms of cell lineage composition and maturation. We showed that tumor tissues contain multiple cell types whose transcriptional identities mirror those of the cellular lineages of the normal epithelium. Moreover, we showed that tumor tissues generated from a single cell can recapitulate the lineage diversity of parent tumors, demonstrating that multilineage differentiation represents a key source of *in vivo* functional and phenotypic cancer cell heterogeneity.

Using these concepts as a guide, we identified biological subsets of human colorectal cancer, based on the expression of genes characteristic of specific cell types. These biological subsets were associated with substantially different clinical outcomes and could be identified by a simple two-gene classifier system. This prognostic scoring system appeared independent of and superior to pathological grading, which is, to this date, one of the few parameters incorporated into the design of therapeutic algorithms for colon cancer patients⁴⁴. Owing to its simplicity and quantitative nature, this two-gene scoring system has the potential to move beyond the realm of purely experimental medicine and is a viable candidate for clinical applications.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturebiotechnology/>.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This study was supported by National Institutes of Health (NIH) grants U54-CA126524 and P01-CA139490 (to S.R.Q. and M.F.C.), the NIH Director's Pioneer Awards (to S.R.Q.) and a grant from the Ludwig foundation (to M.F.C.). P.D. was supported by a training grant from the California Institute for Regenerative Medicine (CIRM) and by a BD Biosciences Stem Cell Research Grant (Summer 2011). T.K. was supported by a fellowship from the Machiah Foundation. D.S. was supported by NIH grant K99-CA151673, by Department of Defense grant W81XWH-10-1-0500 and a grant from the Siebel Stem Cell Institute and the Thomas and Stacey Siebel Foundation. We wish to thank R. Tibshirani, D. Witten, L. Warren, R.A. White III, E. Gilbert, P. Lovelace, M. Palmor, C. Donkers and S.P. Miranda for helpful discussion and technical support in many moments during the completion of this study.

AUTHOR CONTRIBUTIONS

P.D., T.K., D.S., M.F.C. and S.R.Q. conceived the study and designed the experiments. P.S.R., M.E.R., A.A.L., M.Z., N.F.N., M.v.d.W. and H.C. provided intellectual guidance in the design of selected experiments. P.D., T.K., D.S., P.S.R., A.A.L., S.S., J.O., D.M.J., D.Q., J.W., and S.H. performed the experiments. P.D., T.K., D.S., N.F.N., Y.S., M.F.C. and S.R.Q. analyzed the data and/or provided intellectual guidance in their interpretation. J.B., A.A.S. and B.V. provided samples and reagents. P.D., T.K., D.S., M.F.C. and S.R.Q. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/nbt/index.html>.

Published online at <http://www.nature.com/nbt/index.html>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

1. Reya, T., Morrison, S.J., Clarke, M.F. & Weissman, I.L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
2. Jordan, C.T., Guzman, M.L. & Noble, M. Cancer stem cells. *N. Engl. J. Med.* **355**, 1253–1261 (2006).
3. Dalerba, P., Cho, R.W. & Clarke, M.F. Cancer stem cells: models and concepts. *Annu. Rev. Med.* **58**, 267–284 (2007).
4. Shackleton, M., Quintana, E., Fearon, E.R. & Morrison, S.J. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell* **138**, 822–829 (2009).
5. Campbell, L.L. & Polyak, K. Breast tumor heterogeneity: cancer stem cells or clonal evolution? *Cell Cycle* **6**, 2332–2338 (2007).
6. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805**, 105–117 (2010).
7. Kirkland, S.C. Clonal origin of columnar, mucous, and endocrine cell lineages in human colorectal epithelium. *Cancer* **61**, 1359–1363 (1988).
8. Odoux, C. *et al.* A stochastic model for cancer stem cell origin in metastatic colon cancer. *Cancer Res.* **68**, 6932–6941 (2008).
9. Vermeulen, L. *et al.* Single-cell cloning of colon cancer stem cells reveals a multi-lineage differentiation capacity. *Proc. Natl. Acad. Sci. USA* **105**, 13427–13432 (2008).
10. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469**, 415–418 (2011).
11. Warren, L., Bryder, D., Weissman, I.L. & Quake, S.R. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc. Natl. Acad. Sci. USA* **103**, 17807–17812 (2006).
12. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene-expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
13. White, A.K. *et al.* High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl. Acad. Sci. USA* **108**, 13999–14004 (2011).
14. Jiao, Y.F., Nakamura, S., Sugai, T., Yamada, N. & Habano, W. Serrated adenoma of the colorectum undergoes a proliferation versus differentiation process: new conceptual interpretation of morphogenesis. *Oncology* **74**, 127–134 (2008).
15. Wielenga, V.J. *et al.* Expression of CD44 in Apc and Tcf mutant mice implies regulation by the WNT pathway. *Am. J. Pathol.* **154**, 515–523 (1999).
16. Prall, F. *et al.* CD66a (BGP), an adhesion molecule of the carcinoembryonic antigen family, is expressed in epithelium, endothelium, and myeloid cells in a wide range of normal human tissues. *J. Histochem. Cytochem.* **44**, 35–41 (1996).
17. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).
18. Becker, L., Huang, Q. & Mashimo, H. Immunostaining of Lgr5, an intestinal stem cell marker, in normal and premalignant human gastrointestinal tissue. *Scientific World Journal* **8**, 1168–1176 (2008).
19. 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).
20. 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 Biol.* **9**, R157 (2008).
21. Hoglund, P. *et al.* Mutations of the down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea. *Nat. Genet.* **14**, 316–319 (1996).
22. Fischer, H., Stenling, R., Rubio, C. & Lindblom, A. Differential expression of aquaporin 8 in human colonic epithelial cells and colorectal tumors. *BMC Physiol.* **1**, 1 (2001).
23. Koslowski, M., Sahin, U., Dhaene, K., Huber, C. & Tureci, O. MS4A12 is a colon-selective store-operated calcium channel promoting malignant cell processes. *Cancer Res.* **68**, 3458–3466 (2008).
24. Noah, T.K., Kazanjian, A., Whitsett, J. & Shroyer, N.F. SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells. *Exp. Cell Res.* **316**, 452–465 (2010).
25. Gregorieff, A. *et al.* The ets-domain transcription factor Spdef promotes maturation of goblet and paneth cells in the intestinal epithelium. *Gastroenterology* **137**, 1333–1345 (2009).
26. van der Flier, L.G. *et al.* Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* **136**, 903–912 (2009).
27. Ezhkova, E. *et al.* Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell* **136**, 1122–1135 (2009).
28. Park, I.K. *et al.* Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. *Nature* **423**, 302–305 (2003).
29. Sangiorgi, E., Capecchi, M.R. & Giall, R.C. Bmi1 is expressed *in vivo* in intestinal stem cells. *Nat. Genet.* **40**, 915–920 (2008).
30. Zeng, Y.A. & Nusse, R. Wnt proteins are self-renewal factors for mammary stem cells and promote their long-term expansion in culture. *Cell Stem Cell* **6**, 568–577 (2010).
31. Beider, K., Abraham, M. & Peled, A. Chemokines and chemokine receptors in stem cell circulation. *Front. Biosci.* **13**, 6820–6833 (2008).
32. Jensen, K.B. *et al.* Lrig1 expression defines a distinct multipotent stem cell population in mammalian epidermis. *Cell Stem Cell* **4**, 427–439 (2009).
33. Dalla-Favera, R., Wong-Staal, F. & Gallo, R.C. Onc gene amplification in promyelocytic leukaemia cell line HL-60 and primary leukaemic cells of the same patient. *Nature* **299**, 61–63 (1982).
34. Hoey, T. *et al.* DLL4 blockade inhibits tumor growth and reduces tumor-initiating cell frequency. *Cell Stem Cell* **5**, 168–177 (2009).
35. Park, S.Y., Gonen, M., Kim, H.J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120**, 636–644 (2010).
36. Losi, L., Baisse, B., Bouzourene, H. & Benhattar, J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. *Carcinogenesis* **26**, 916–922 (2005).
37. Dalerba, P. *et al.* Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. USA* **104**, 10158–10163 (2007).
38. Oien, K.A. Pathologic evaluation of unknown primary cancer. *Semin. Oncol.* **36**, 8–37 (2009).
39. Lugli, A., Tzankov, A., Zlobec, I. & Terracciano, L.M. Differential diagnostic and functional role of the multi-marker phenotype CDX2/CK20/CK7 in colorectal cancer stratified by mismatch repair status. *Mod. Pathol.* **21**, 1403–1412 (2008).
40. 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).
41. 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).
42. 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).
43. Guastadisegni, C., Colafranceschi, M., Ottini, L. & Dogliotti, E. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. *Eur. J. Cancer* **46**, 2788–2798 (2010).
44. Bardia, A. *et al.* Adjuvant chemotherapy for resected stage II and III colon cancer: comparison of two widely used prognostic calculators. *Semin. Oncol.* **37**, 39–46 (2010).

ONLINE METHODS

Human primary tissues and colon cancer xenografts. Human primary colon tissues, normal or malignant, were collected according to guidelines from Stanford University's institutional review board. Human colon cancer tissues used in this study, either from primary samples or xenograft lines, are listed in **Supplementary Table 4**, together with clinical information relative to corresponding patients. Human colon cancer xenograft lines were established and serially passaged in immunodeficient mice following previously published protocols³⁷. A detailed description of these protocols is provided in the **Supplementary Methods**.

Cell lines. Calibration experiments to measure accuracy and precision of single-cell sorting by flow cytometry, as well as to measure single-cell sensitivity of single-cell PCR, were performed on a clone of the HCT116 human colon cancer cell line infected with the pLL3.7 lentivirus (Addgene no. 11795). HCT116 cells are available from the American Tissue-type Culture Collection (ATCC; CCL-247) and were maintained in RPMI-1640 medium, supplemented with 10% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 120 µg/ml penicillin, 100 µg/ml streptomycin, 20 mM HEPES and 1 mM sodium pyruvate, as previously described⁴⁵.

Solid tissue disaggregation. Solid tissues, normal and neoplastic, collected from primary surgical specimens or mouse xenografts, were mechanically and enzymatically disaggregated into single-cell suspensions, following previously published protocols³⁷. Briefly, solid tissues were minced into small chunks (2 mm³), rinsed with Hank's balanced salt solution (HBSS), finely chopped with a razor blade into minute fragments (0.2–0.5 mm³), resuspended in serum-free RPMI-1640 medium (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, 1 mM sodium pyruvate) together with 100 units/ml DNase-I and 200 units/ml Collagenase-III (Worthington) and incubated for 2 h at 37 °C to obtain enzymatic disaggregation. Cell suspensions were serially filtered with sterile gauze, 70-µm and 40-µm nylon meshes. Red blood cells were removed by osmotic lysis with ACK hypotonic buffer (150 mM NH₄Cl, 1 mM KHCO₃; 5 min on ice).

Flow cytometry and single-cell sorting experiments. To minimize loss of cell viability, we performed experiments on fresh cell suspensions, prepared shortly before flow cytometry³⁷. 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, 1 mM sodium pyruvate and 5 mM EDTA. To minimize unspecific binding of antibodies, cells were first incubated with 0.6% human IgGs (Gammagard Liquid; Baxter) for 10 min on ice, at a concentration of 3–5 × 10⁵ cells/100 µl. Cells were subsequently washed and stained with antibodies at dilutions determined by appropriate titration experiments. Antibodies used in this study include anti-human EpCAM-FITC or PE (clone EBA-1; BD Biosciences), anti-human CD44-APC (clone G44-26; BD Biosciences) and anti-human CD66a-PE (clone 283340; R&D Systems). Cells positive for expression of nonepithelial lineage markers (Lin⁺) were excluded by staining with PE.Cy5-labeled antibodies using different strategies for primary tissues and mouse xenografts. In experiments on primary human tissues, stromal cells were excluded by 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 staining with anti-mouse CD45-PE/Cy5 (clone 30-F11; BD Biosciences) and anti-mouse H-2K^d-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 with 2% heat-inactivated calf serum, 20 mM HEPES, 5 mM EDTA, 1 mM sodium pyruvate and 1.1 µM DAPI dilactate (Molecular Probes). Flow-cytometry analysis was performed using a BD FACSAriaII cell-sorter (Becton Dickinson). Forward-scatter height versus forward-scatter width (FSC-H versus FSC-W) and side-scatter height versus side-scatter width (SSC-H versus 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 software package, with appropriate adjustments (device: 96-well plate; precision: single-cell; nozzle: 130 µm).

Single-cell PCR. Single-cell gene-expression experiments were performed using Fluidigm's M96 quantitative PCR (qPCR) DynamicArray microfluidic chips (Fluidigm). Single cells were sorted by FACS into individual wells of 96-well PCR plates as described above. Each 96-well plate was preloaded with 5 µl/well of CellsDirect PCR mix (Invitrogen) and 0.1 µl/well (2 U) of SupraseIn RNase-inhibitor. Following single-cell sorting, each well was supplemented with 1 µl (Applied Biosystems) 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) containing each assay at 1:100 dilution. Single-cell mRNA was directly reverse transcribed into cDNA (50 °C for 15 min, 95 °C for 2 min), pre-amplified for 20 cycles (each cycle: 95 °C for 15 s, 60 °C for 4 min) and diluted 1:3 with TE buffer. A 2.25 µl aliquot of amplified cDNA was then mixed with 2.5 µl of TaqMan Universal PCR Master Mix (Applied Biosystems) and 0.25 µl of Fluidigm's "sample loading agent," then inserted into one of the chip "sample" inlets. Individual TaqMan assays were diluted at 1:1 ratios with TE. A 2.5 µl aliquot of each diluted TaqMan assay was mixed with 2.5 µl of Fluidigm's "assay loading agent" and individually inserted into one of the chip "assay" inlets. Samples and probes were loaded into M96 chips using an IFC Controller HX (Fluidigm), then transferred to a BioMark real-time PCR reader (Fluidigm) following manufacturer's instructions. A list of the 57 TaqMan assays used in this study is provided in **Supplementary Table 2**.

Analysis and graphic display of single-cell PCR data. Single-cell PCR data were analyzed and displayed using MATLAB (MathWorks) as summarized in **Supplementary Figure 2**. A minimum of 336 cells were analyzed for each phenotypic population, corresponding to four PCR plates, each containing 84 single cells (84 × 4 = 336), eight positive and four negative controls. As positive controls, we used replicates of a 1:1:1 mixture of total RNA standards from human normal colon (AM7986), human normal testes (AM7972) and HeLa cells (AM7852), all from Applied Biosystems. Results from cells not expressing *ACTB* (β-actin) and *GAPDH* (glyceraldehyde 3-phosphate dehydrogenase), or expressing them at extremely low values (Ct >35), were removed from the analysis. Gene-expression results were normalized by mean centering and dividing by 3 times the standard deviation (3 s.d.) of expressing cells (**Supplementary Fig. 2**), and visualized using both hierarchical clustering and PCA^{12,46}. Hierarchical clustering was performed both on cells and genes, based on Euclidean or correlation distance metric and complete linkage. Positive or negative associations between two genes were tested by Spearman correlation, and *P*-values calculated based on 10,000 permutations. Both hierarchical clustering and PCA were based on the results for 47 differentially expressed genes (51 assays), and excluded results from housekeeping (*ACTB*, *GAPDH*, *EpCAM*) and proliferation-related genes (*MKI67*, *TOP2A*, *BIRC*) to avoid noise based on proliferation status. A detailed description of all these procedures is provided in the **Supplementary Methods**.

Immunohistochemistry and immunofluorescence. Paraffin-embedded tissue sections were stained with anti-human CK20 (clone Ks20.8, DakoCytomation), MUC2 (clone Ccp58, Fitzgerald Industries), MKI67 (clone MIB-1, DakoCytomation), CEACAM1/CD66a (clone 283340; R&D Systems) and SLC26A3 (lot no. R32905, Sigma Life Science) antibodies, according to manufacturers' instructions. Frozen tissue sections were stained with an anti-human CD177 antibody (clone MEM-166, BD Biosciences) followed by secondary staining with goat anti-mouse IgG-Alexa488 (Invitrogen). A description of immunohistochemistry and immunofluorescence protocols is provided in the **Supplementary Methods**.

Generation and characterization of monoclonal tumors. EpCAM^{high}/CD44⁺ human colon cancer cells were infected with the pLL3.7 lentivirus (Addgene #11795)⁴⁷. Cells were infected by spin-inoculation for 4 h and injected in bulk



into the subcutaneous tissue of a NOD/SCID/IL2R $\gamma^{-/-}$ mice. The resulting tumors were analyzed to evaluate infection efficiency, and EGFP $^{+}$ /EpCAM $^{\text{high}}$ /CD44 $^{+}$ were re-sorted and injected as single cells, again into NOD/SCID/IL2R $\gamma^{-/-}$ mice. Monoclonal origin of tumors originated from single ($n = 1$) lentivirus-infected EpCAM $^{\text{high}}$ /CD44 $^{+}$ cancer cells was confirmed by ligation-mediated PCR (LM-PCR)⁴⁸, followed by DNA sequencing of LM-PCR amplification products. In the case of UM-COLON#4 clone 8, DNA sequencing of LM-PCR amplification products pinpointed the provirus integration-site on the long arm of human chromosome 19 (19q13.3), in proximity of the *AP3D1* gene (adaptor-related protein complex 3, delta 1 subunit). For a visual guide on how to interpret LM-PCR results refer to **Supplementary Figure 24**.

Tumorigenicity experiments. Tumorigenicity experiments were performed in NOD/SCID/IL2R $\gamma^{-/-}$ immunodeficient mice following previously published protocols^{37,49,50} and Stanford University's institutional animal welfare guidelines. Tumorigenic cell frequencies were calculated by limiting dilution using the L-Calcul software (StemCell Technologies). A detailed description of the protocols used for tumorigenicity experiments is provided in the **Supplementary Methods**.

Bioinformatic data collection and assemblage of the “human colon global database.” A collection of 46,047 publicly available human gene-expression arrays (25,721 arrays on Affymetrix U133 Plus 2.0, 16,357 arrays on Affymetrix U133A, 3,969 arrays on Affymetrix U133A 2.0) was downloaded from NCBI's GEO database and normalized using the RMA (Robust Multi-chip Average) algorithm. Normalization was done either independently for each platform or on the whole array collection, using a modified CDF (chip description file) reduced to contain only shared probes. From this general collection, which contained arrays from all types of human samples, we extracted a subset of 1,684 unique arrays from human colon tissues, either normal or cancerous. We named this subset the “human colon global database,” and we annotated all its samples as normal colon ($n = 173$), benign colonic adenoma ($n = 68$) or colorectal cancer ($n = 1443$). To avoid redundancies (that is, identical samples deposited two or more times in independent GEO data sets) we cross-checked all samples and removed duplicates. When available, we collected all available clinical, pathological and molecular information related to the corresponding patients. As not all arrays were annotated for all variables, individual hypotheses were tested on specific subsets of the “human colon global database.” A list of all GEO data sets used in this study, and of their contribution to different experiments, is provided in **Supplementary Table 1**.

Mining of gene-expression arrays using Boolean implications. Gene-expression thresholds between positive and negative samples were defined using the StepMiner algorithm⁴⁰, and Boolean implication relationships between pairs of genes using the BooleanNet software²⁰. Briefly, for each gene, individual samples were ordered from low-to-high based on their gene-expression values, and a rising step function was fit to the data, trying to minimize differences between fitted and measured values. This method identifies a “step” at the point of largest jump from low to high values (but only if a sufficient number of gene-expression values is present on each side of the jump to exclude a random oscillation due to noise) and sets the gene-expression threshold at the value corresponding to the step⁴⁰. An intermediate region is defined around the threshold, with a width of 1 (threshold ± 0.5), corresponding to a twofold change in expression levels, which represents the minimum noise in these data sets^{20,40}. All samples below the intermediate region ($< 1^{\text{st}}$ StepMiner threshold $- 0.5$) are considered negative, and all samples above the intermediate region ($> 1^{\text{st}}$ 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 (1^{st} StepMiner threshold) and a second threshold to split “positive” samples into two subgroups with “low” and “high” gene-expression (2^{nd} StepMiner threshold; **Supplementary Fig. 20**).

We started our search for developmentally regulated genes on the “human colon global database” (**Supplementary Table 1**). To minimize the risk of results being affected by samples containing substantial contaminations from tissues other than colorectal epithelium (e.g., normal liver tissue in hepatic metastases), we restricted our investigation to the subset of arrays

with an *EpCAM $^{+}$ /albumin $^{-}$* gene-expression profile (**Supplementary Fig. 6**). Threshold gene-expression levels were calculated using the StepMiner algorithm, based on the 1,684 arrays of the “human colon global database” (*EpCAM $^{+}$* : Affymetrix probe 201839_s_at > 10.05 ; *albumin $^{-}$* : Affymetrix probe 211298_s_at < 7.97). This operation removed 116 arrays (6.9%) and left 1,568 arrays (93.1%) for analysis (normal colon: $n = 170$; colorectal adenoma: $n = 68$; colorectal carcinoma: $n = 1,330$).

Boolean implication relationships between pairs of genes were systematically computed using the BooleanNet software²⁰. Mature enterocyte genes were predicted as genes highly expressed in *KRT20 $^{+}$* arrays and filtered based on the fulfillment of the “ X^{+} implies *KRT20 $^{+}$* ” Boolean implication (**Supplementary Fig. 7**). Goblet genes were predicted as genes highly expressed in *MUC2 $^{+}$* arrays and filtered based on the fulfillment of at least one of three independent Boolean implications: “*MUC2* is equivalent to X^{+} ”, “ X^{+} implies *MUC2 $^{+}$* ”, “*MUC2 $^{+}$* implies X^{+} ” (**Supplementary Fig. 8**). Immature genes were predicted as genes highly expressed in *KRT20 $^{-}$* arrays, and filtered based on the fulfillment of the “*KRT20 $^{-}$* implies X^{+} ” Boolean implication (**Supplementary Fig. 9**). Threshold gene-expression levels were calculated using the StepMiner algorithm, based on the global collection of 46,047 human arrays. Gene-expression patterns were considered to fulfill a Boolean implication when the false-discovery rate (FDR) of a sparsity test in the relevant quadrant was < 0.05 (ref. 20).

Differences in gene-expression levels among different sample groups (e.g., normal versus adenoma) were evaluated using box plots and tested for statistical significance using a 2-sample *t*-test (2-tailed). Correlations between two genes' expression levels were measured using Pearson correlation coefficients.

Stratification of human colon cancer patients in distinct gene-expression groups. Associations between gene-expression profiles and patient survival were investigated using a new bioinformatics tool, named Hegemon. Hegemon is an upgrade of the BooleanNet software, where individual gene-expression arrays, after being plotted on a two-axis chart based on the expression of two given genes²⁰, can be grouped and compared for survival outcomes, using both Kaplan-Meier curves and multivariate analysis based on the Cox proportional hazards method.

Survival analysis was done on a gene-expression database annotated with disease-free survival (DFS) information on 299 patients from three institutions: H. Lee Moffitt 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 and partially redundant GEO data sets (GSE14333, GSE17538; **Supplementary Table 1**)^{41,42}, both collected on Affymetrix U133 Plus 2.0. To avoid bias due to redundancies (that is, identical samples deposited in both GEO data sets), we cross-checked all samples and removed duplicates.

Guided by single-cell PCR results, we chose to stratify patients using four genes characteristic of top-of-the-crypt *CA1 $^{+}$ /SLC26A3 $^{+}$* cells (*CA1*, *MS4A12*, *CD177*, *SLC26A3*) as markers of terminal differentiation, and using *KRT20*, whose expression is observed in both top-of-the-crypt *CA1 $^{+}$ /SLC26A3 $^{+}$* cells and a subset of *MUC2 $^{+}$ /TFF3 $^{\text{high}}$* goblet-type cells, as a more promiscuous marker of both intermediate and terminal differentiation. The hypothesis behind this approach was that, on average, a tumor's overall gene-expression profile would most closely resemble that of the most abundant cell population. Thus, tumors highly enriched in mature, terminally differentiated cell types would be characterized by a lower proliferation rate and/or a lower content of long-term self-renewing cells, and be associated with a better prognosis as compared to tumors predominantly composed by immature, progenitor-like cells.

Threshold gene-expression levels were calculated using the StepMiner algorithm, based on the 25,576 arrays on Affymetrix U133 Plus 2.0. *KRT20* expression (Affymetrix probe 213953_at) was tested as a marker to separate poorly differentiated tumors (*KRT20 $^{-}$*) from differentiated ones (*KRT20 $^{+}$*). Based on our previous experience⁴⁰, we defined as *KRT20 $^{-}$* all tumors whose *KRT20* expression values were $< 1^{\text{st}}$ 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-crypt $^{\text{high}}$*) from moderately differentiated ones (*top-crypt $^{\text{low}}$*). 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**). To maintain consistency in grouping samples with the highest expression levels, we adopted a scaled approach designed to match the different sensitivity of individual gene-expression probes (**Supplementary Fig. 20**). In the case of *CD177* and *SLC26A3*, we chose to separate negative samples from positive ones (*CD177*⁻ versus *CD177*⁺, *SLC26A3*⁻ versus *SLC26A3*⁺), whereas in the case of *CA1* and *MS4A12* we chose to separate low-to-negative expression samples from high expression ones (*CA1*^{-/low} versus *CA1*^{high}, *MS4A12*^{-/low} versus *MS4A12*^{high}). 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 *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 colon tumors into three “gene-expression groups”: Group 1 (*KRT20*⁺/*top-crypt*^{high}), Group 2 (*KRT20*⁺/*top-crypt*^{-/low}), Group 3 (*KRT20*⁻/*top-crypt*^{-/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*⁻/*top-crypt*^{high}), with the only exception of one isolated sample in the *KRT20/SLC26A3* experiment. In experiments involving comparisons to the EphB2⁺ “intestinal stem cell” (EphB2-ISC) signature (**Supplementary Fig. 23**),

tumors were grouped in three categories (EphB2-ISC^{low}, EphB2-ISC^{medium}, EphB2-ISC^{high}), as described in Merlos-Suarez *et al.*¹⁹.

Survival analysis and other statistical tests. Once grouped based on gene-expression profiles, patient subsets were compared for survival outcomes using Kaplan-Meier 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. 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 (Group-3, *KRT20*⁻/*top-crypt*^{-/low}) was measured using odds-ratios and tested for significance using Pearson’s χ^2 test.

45. 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).
46. Ringner, M. What is principal component analysis? *Nat. Biotechnol.* **26**, 303–304 (2008).
47. O’Doherty, U., Swiggard, W.J. & Malim, M.H. Human immunodeficiency virus type 1 spinoculation enhances infection through virus binding. *J. Virol.* **74**, 10074–10080 (2000).
48. Wang, G.P. *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.* **36**, e49 (2008).
49. Ishizawa, K. *et al.* Tumor-initiating cells are rare in many human tumors. *Cell Stem Cell* **7**, 279–282 (2010).
50. Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature* **456**, 593–598 (2008).

