Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Dalerba P, Sahoo D, Paik S, et al. CDX2 as a prognostic biomarker in stage II and stage III colon cancer. N Engl J Med 2016;374:211-22. DOI: 10.1056/NEJMoa1506597

Supplementary Appendix

CDX2 as a prognostic and predictive biomarker in Stage-II/III colon cancer.

Piero Dalerba, M.D., Debashis Sahoo, Ph.D., Soonmyung Paik, M.D., Xiangqian Guo, Ph.D., Greg Yothers, Ph.D., Nan Song, Ph.D., Nate Wilcox-Fogel, M.S., Erna Forgó, M.D., Pradeep S. Rajendran, B.S., Stephen P. Miranda, B.A., Shigeo Hisamori, M.D., Ph.D., Jacqueline Hutchison, Tomer Kalisky, Ph.D., Dalong Qian, M.D., Stephen R. Quake, Ph.D., Norman Wolmark, M.D., George A. Fisher, M.D., Ph.D., Matt van de Rijn, M.D., Ph.D., and Michael F. Clarke, M.D.

Table of Contents:

Supplementary Methods

Assembly and normalization of gene-expression array databases used for the bioinformatics search of novel biomarkers of colon epithelial differentiation. The bioinformatics search of novel biomarkers of colon epithelial differentiation started from a large collection (n = 47,240) of publicly available human gene-expression array experiments downloaded from the *National Center for Biotechnology Information* (NCBI) *Gene Expression Omnibus* (GEO) database (http://www.ncbi.nlm.nih.gov/geo). This collection, which we called the *"Human NCBI-GEO Global Database"*, was assembled by pooling experiments from a heterogeneous repertoire of NCBI-GEO data-series (GSEs), all of which contained experiments performed using Affymetrix platforms on different types of human samples. To avoid redundancies (i.e. identical samples deposited two or more times across multiple datasets) all data-series used to assemble the *"Human NCBI-GEO Global Database"* were cross-checked for duplications, and all duplicated data-files were removed. The *"Human NCBI-GEO Global Database"* latest update was performed on February 1st, 2015, at which time it was composed of 47,240 experiments, performed on four distinct platforms: a) the human Affymetrix HG-U133 Plus 2.0 platform (GPL570; n = 25,955); b) the human Affymetrix HG-U133A platform (GPL96; n = 17,001); c) the human Affymetrix HG-U133A2 platform (GPL571; n = 4,033); d) the human Affymetix HT-HG-U133A 2.0 platform (GPL3921; n = 251). A complete list of all the individual NCBI-GEO sample number identifiers (GSMIDs) of the experiments contained within the *"Human NCBI-GEO Global Database"* is provided in Table S2, which contains both an *"aggregate"* list of all experiments (n = 47,240; Table S2a) and four platform-specific lists, one for each of the four Affymetrix platforms included in the database: GPL570 (n = 25,955; Table S2b), GPL96 (n = 17,001; Table S2c), GPL571 (n = 4,033; Table S2d) and GPL3921 (n = 251; Table S2e). After having been downloaded and cross-checked for duplications, all gene-

expression arrays contained within the *"Human NCBI-GEO Global Database"* were pooled, normalized using the RMA (*Robust Multi-chip Average*) algorithm and transformed to log2 values. Normalization was performed either independently for each of the four different Affymetrix platforms or on the whole array collection, using a modified CDF (chip description file) reduced to contain only the probes that were shared across the four platforms.

From the *"Human NCBI-GEO Global Database"*, which contains experiments performed on all types of human samples, we then extracted a subset collection of 2,466 unique and non-redundant experiments performed on human colon epithelial tissues. We named this subset collection the "*Human Colon Global Database*", and we annotated all samples contained within it either as colorectal cancer (n = 2,239) or normal colon mucosa (n = 227). This subset collection contained experiments from 28 independent NCBI-GEO data-series (GSEs). A list of the 28 NCBI-GEO GSEs contained within the "*Human Colon Global Database*" is provided in both Table S1 and Figure S1 (Panel A). A complete list of all the GSMIDs of the experiments contained within the "*Human Colon Global Database*" (n = 2,466) is provided in Table S3a.

Calculation of gene-expression thresholds to separate negative and positive samples. To define the gene-array expression thresholds used to separate negative from positive samples for a specific mRNA, we used the *StepMiner* algorithm.[1] Briefly, for each gene, the normalized log2 expression values of all samples in the database 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. The *StepMiner* algorithm identifies the *"step"* as the point of largest jump from low to high values (but only if there are sufficient numbers of 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. An intermediate region is 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.[1,2] All the samples below the intermediate region (< *StepMiner* threshold - 0.5) are considered negative, and all those above the intermediate region (> *StepMiner* threshold + 0.5) are considered positive.

Computer-assisted data mining of gene-expression array databases using Boolean logic. The Boolean search for novel biomarkers of colon epithelial differentiation was conducted on our annotated "*Human Colon Global Database*" (Table S1, Figure S1). As previously mentioned, this database was assembled by pooling data from 28 independent NCBI-GEO data-series (GSEs), containing gene-expression information from 2,466 independent tissue samples, including both human colorectal carcinomas (n = 2,239) and human normal colon epithelia (n = 227). To minimize the risk that gene-associations 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 $EpCAM^{pos}/Albumin^{neg}$. EpCAM (*Epithelial Cell Adhesion Molecule*) is also designated as TACSTD1 (*Tumor-Associated Calcium Signal Transducer 1*), ESA (*Epithelial Specific Antigen*) or CD326, and 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. Threshold gene expression levels were calculated using the *StepMiner* algorithm, based on the expression distribution of the 2,466 arrays contained within the "*Human Colon Global Database*" itself ($EpCAM^{pos}$ defined as Affymetrix probe 201839_s_at >10.0; $Albumin^{neg}$ defined as Affymetrix probe 211298_s_at <7.8; Figure S1). This operation removed 137 arrays (6%) and left 2,329 arrays (94%) for subsequent analysis (colorectal carcinoma: n = 2,115; normal colon mucosa: n = 214; Figure S1). A complete list of all GSMIDs of the samples contained within *"Human Colon Global Database"* after *"purging"* based on the fulfillment of the $EpCAM^{pos}/Albumin^{neg}$ condition (n = 2,329) is provided in Table S3b.

We then used the *BooleanNet* software[2] to perform a systematic screen for genes that would fulfill the *"X^{neg} implies ALCAM^{pos}"* Boolean implication in the *"purged"* subset (n = 2,329) of the "*Human Colon Global Database*" (Figure S2). ALCAM *(Activated Leukocyte Cell Adhesion Molecule*), also known as CD166, was chosen as a marker of immature colon epithelial cells, because of its preferential expression at the bottom of colon crypts[3,4] and on human colon cancer cells with enriched tumorigenic capacity in mouse xenotransplantation models.[5] To perform the search, the threshold expression levels for *ALCAM* and all other genes represented in the arrays were calculated using the *StepMiner* algorithm, based on the expression distribution of the 47,240 arrays contained within the *"Human NCBI-GEO Global Database"* (Figure S2). Gene-expression patterns were considered to fulfill the *"X^{neg} implies ALCAM^{pos}"* Boolean implication when the false-discovery rate (FDR) of a sparsity test in the lower-left quadrant of a *"X vs. ALCAM"* two-axis plot was < 0.0001 (10^{-4}), as described in Figure S2. The search yielded 16 candidate genes (Figure S3), of which only one encoded for a protein whose expression levels could be studied by immunohistochemistry using a clinical-grade diagnostic test: CDX2 (*Caudal Type Homeobox Transcription Factor 2*).

Analysis of the relationship between *CDX2* mRNA expression and molecular features frequently observed in human colorectal cancer, such as microsatellite instability (MSI) and mutations of *Tumor Protein 53* (TP53). The relationship between *CDX2* mRNA expression levels and other molecular features frequently observed in human colorectal carcinomas, such as microsatellite instability (MSI) and TP53 mutations, was studied in two *ad-hoc* databases, obtained by pooling a limited collection of gene-expression array datasets annotated with the molecular variables of interest, and publicly available from the NCBI-GEO on-line repository (Table S1). The database used to test the relationship between *CDX2* mRNA expression and MSI was assembled by pooling seven gene-expression array datasets (GSE13067,

GSE13294, GSE24514, GSE26682, GSE35896, GSE39084, GSE41258) and contained a total of 862 independent primary tumors (Figure S4). In this database, all samples originally classified as MSI-low were re-classified as microsatellite stable (MSS). The database used to test the relationship between *CDX2* mRNA expression and TP53 mutations was assembled by pooling two gene-expression array datasets (GSE39084, GSE41258) and contained a total of 214 independent primary tumors (Figure S5). In both databases, tumor samples were stratified in *CDX2*^{neg} and *CDX2*^{pos} subgroups using the *StepMiner* algorithm, based on the expression distribution of the 47,240 arrays contained within the *"Human NCBI-GEO Global Database"* (*CDX2*^{neg} defined as Affymetrix probe 206387_at < 6.46; Figure S2, Panel D). A complete list of all GSMIDs of the experiments contained within the two databases is provided in Table S4a (MSI/MSS) and Table S4b (TP53).

Stratification of colon cancer patients in distinct gene-expression subgroups and comparative analysis of their survival outcomes. The association between the mRNA expression of selected genes (i.e. *CDX2, ALCAM*) and patient survival was tested in a subset series of the "*Human Colon Global Database*" where each tumor had been annotated with the disease-free survival (DFS) information of the corresponding patient. We used this subset series as the *"discovery dataset"* for our study (Figure 1). The discovery dataset included gene-expression data from four publicly available NCBI-GEO data-series (GSE14333, GSE17538, GSE31595, GSE37892; Figure S6)[6-9], and contained information on 466 unique primary colon carcinoma samples, collected from patients at various clinical stages (AJCC Stage I-IV/Duke's Stage A-D) by five independent institutions: 1) the *H. Lee Moffit Cancer Center* in Tampa, Florida, USA (n = 164); 2) the *Vanderbilt Medical Center* in Nashville, Tennessee, USA (n = 55); 3) the *Royal Melbourne Hospital* in Melbourne, Australia (n = 80); 4) the *Institut Paoli-Calmette* in Marseille, France (n = 130); 5) the *Roskilde Hospital* in Copenhagen, Denmark (n =

37). As mentioned previously, in order to avoid bias due to possible redundancies (i.e. identical samples replicated two or more times across multiple NCBI-GEO datasets) all 466 samples contained in this subset were cross-checked to exclude the presence of duplicates (Figure S6). A complete list of all GSMIDs of the experiments contained within the NCBI-GEO discovery dataset is provided in Table S5, which contains both an *"aggregate"* list of all experiments (n = 466; Table S5a) and three sub-lists: 1) samples annotated with both DFS and pathological grade information (n = 216; Table S5b); 2) Stage-II samples annotated with DFS and adjuvant chemotherapy information (n = 114; Table S5c); 3) Stage-III samples annotated with DFS and adjuvant chemotherapy information (n = 108; Table S5c).

To investigate the relationship between the mRNA expression levels of selected genes (i.e. *CDX2*, *ALCAM*) and the clinical outcomes of the 466 colon cancer patients represented within the NCBI-GEO discovery dataset, we applied the *Hegemon* software tool.[10] The *Hegemon* software is an upgrade of the *BooleanNet* software,[2] where individual gene-expression arrays, after having been plotted on a two-axis chart based on the expression levels of any two given genes, can be stratified using the *StepMiner* algorithm and automatically compared for survival outcomes using Kaplan-Meier curves and log-rank tests. Since all 466 samples contained in the discovery dataset had been analyzed using the Affymetrix HG-U133 Plus 2.0 platform (GPL570), the threshold gene-expression levels for *CDX2* and *ALCAM* were calculated using the *StepMiner* algorithm based on the expression distribution of the 25,955 experiments performed on the Affymetrix HG-U133 Plus 2.0 platform (Table S2b). $CDX2^{neg}$ tumors were defined as Affymetrix probe 206387_at < 6.5, while $ALCAM^{neg}$ tumors were defined as Affymetrix probe 201951_at < 6.8 (Figure S2, Panel D). Based on these definitions, we stratified the patient population of the NCBI-GEO discovery dataset in different gene-expression subgroups, based on either the mRNA expression levels of *CDX2* alone (i.e. $CDX2^{neg}$ vs. $CDX2^{pos}$; Figure 2), *ALCAM*

alone (i.e. *ALCAM*^{neg} vs. *ALCAM*^{pos}; Figure S7), or a combination of both *CDX2* and *ALCAM* (i.e. *CDX2*^{neg}/*ALCAM*^{pos} vs. *CDX2*^{pos}/*ALCAM*^{pos} vs. *CDX2*^{pos}/*ALCAM*^{neg}; Figure S8-S9). Once grouped based on their gene-expression levels, patient subsets were compared for survival outcomes using both Kaplan-Meier survival curves and multivariate analysis based on the Cox proportional hazards method. In experiments involving comparisons to the *Ephrin-B2* (*EphB2*) *"intestinal stem cell"* (ISC) signature (EphB2-ISC; Figure S10), colon cancer patients were also grouped in three categories (EphB2-ISC^{low}, EphB2-ISC^{medium}, EphB2-ISC^{high}) based on the method described by Merlos-Suarez *et al*.[11]

CDX2 immunohistochemistry. Immunohistochemical analysis of tumor tissues was performed on *"formalin-fixed, paraffin-embedded"* (FFPE) tissue sections. Tissue sections were stained with a mouse anti-human CDX2 monoclonal antibody previously validated for diagnostic applications (clone CDX2-88, mouse IgG1-kappa, dilution 1:12.5, 4 mg/ml; BioGenex, USA).[12,13] The staining protocol was based upon recommendations from the *Nordic Immunohistochemical Quality Control* (NordiQC) organization (www.nordiqc.org), which suggests heat-induced antigen retrieval (HIER) with Tris/EDTA at pH 9.0 (Epitope Retrieval Solution pH9, Leica, Germany)[14]. Tissue slides were stained on a Bond-Max automatic stainer (Leica) and antigen detection was visualized using the BOND Polymer Refine detection kit (Leica).

Analysis of CDX2 protein expression levels in tissue microarrays (TMAs). Colon cancer TMAs fully annotated with clinical and pathological information were obtained from three independent sources: a) the *National Cancer Institute's* (NCI) *Cancer Diagnosis Program* (CDP; http://cdp.nci.nih.gov/colon), which provided us with a TMA that was purposefully designed to maximize the statistical power to find associations between individual biomarkers and clinical outcomes, based on a cohort of 367 independent tumor samples (Stage I: 49; Stage II: 122; Stage

III: 144; Stage IV: 52) that contained a balanced distribution of cases with different survival outcomes and long-term follow-up (validation dataset; Figure S11); b) the *National Surgical Adjuvant Breast and Bowel Project* (NSABP), which provided us with a TMA representative of the whole patient cohort recruited into the NSABP-C07 clinical trial (n = 1,519) and inclusive of both Stage-II (n = 435) and Stage-III (n = 1,084) colon cancer patients, all treated with adjuvant chemotherapy consisting of either 5-fuorouracil + leucovorin (FULV; n = 753) or 5-fuorouracil + leucovorin + oxaliplatin (FLOX; n = 766; expansion dataset #1, Figure S12)[15]; c) the *Stanford Tissue Micro-Array Database* (Stanford TMAD; https://tma.im/cgi-bin/home.pl), which provided us with a newly assembled TMA designed in collaboration with the *Stanford Cancer Registry*, representing a large sample (n = 321) of the colon cancer patient population treated at *Stanford Hospital* between 1992 and 2011 (Stage I: 19; Stage II: 122; Stage III: 151; Stage IV: 25; expansion dataset #2, Figure S13). All tumors were scored blindly by one of the authors. In cases where the TMA contained two tissue cores for each patient (i.e. two samples from distinct areas of the same tumor) the two cores were scored independently and paired at the end. Tumors with discordant scores on the two sections were upgraded to the highest score. A detailed description of the scoring system, together with representative photographs and scoring results, is provided in Figure S14. Briefly, we scored as $CDX2^{pos}$ all tumors whose malignant epithelial component displayed widespread nuclear expression of CDX2, either in all or a majority of cancer cells. We scored as $CDX2^{neg}$ all tumors whose malignant epithelial component either completely lacked CDX2 expression or showed faint nuclear expression in a minority of malignant epithelial cells. To assess the robustness of our scoring system, and control for possible bias introduced by inter-observer variability, we analyzed the concordance between the scoring results obtained by two independent investigators using contingency tables and calculating *Cohen's Kappa Indices* (Figure S15).[16] Finally, the association between CDX2

expression and survival outcomes was tested by a third investigator, who did not participate in the scoring process.

Statistical tests. Once stratified based on gene or protein expression patterns, patient subsets were compared for survival outcomes, using both Kaplan-Meier survival curves and multivariate analyses based on the Cox proportional hazards method. Differences in Kaplan-Meier curves were tested for statistical significance using the log-rank test. *"Treatment-by-marker"* interactions between CDX2 status ($CDX2^{neg}$ vs. $CDX2^{pos}$) and adjuvant chemotherapy (no-chemotherapy vs. chemotherapy) were evaluated using the Cox proportional-hazards regression model in a 2 x 2 factorial design, by testing the statistical significance for the presence of an *"interaction factor"* (e.g. a multiplicative rather than merely additive effect) between the hazard-rates associated with each of the two variables individually.[17] The presence of an enrichment in $CDX2^{neg}$ carcinomas in tumors characterized by high pathological grade (G3/G4), microsatellite instability (MSI) or mutations in the TP53 gene was tested using Pearson's χ^2 test and by computing odds-ratios (OR) together with their 95% confidence intervals (CI). Differences in the expression levels of individual genes among different sample subgroups (e.g. $CDX2^{neg}$ vs. $CDX2^{pos}$, MSI vs. MSS, $TP53^{wild\text{-}type}$ vs. $TP53^{mutated}$ carcinomas) were evaluated using box-plots[18] and tested for statistical significance using a 2-sample t-test (2-tailed). The concordance between immunohistochemistry scoring results obtained by independent investigators was assessed using contingency tables and calculating *Cohen's Kappa Indices*.[16]

Figure S1. Assembly and purging of the "*Human Colon Global Database*"

A List of NCBI-GEO datasets used to assemble the *"Human Colon Global Database"*

GEO Dataset	Colon Cancer	Normal Colon	Total
GSE2109	427		427
GSE2361		1	1
GSE4045	37		37
GSE4107	12	10	22
GSE4183	15	8	23
GSE5851	80		80
GSE8671		32	32
GSE9254		19	19
GSE9348	70	12	82
GSE10714	7	3	10
GSE10961	18		18
GSE11831		17	17
GSE12945	62		62
GSE13067	74		74
GSE13294	155		155
GSE13471	4	4	8
GSE14333	226		226
GSE15960	6	6	12
GSE17538	65		65
GSE18088	53		53
GSE18105	94	17	111
GSE20916	91	44	135
GSE26682	331		331
GSE26906	58		58
GSE29623	1		1
GSE31595	37		37
GSE37892	130		130
GSE41258	186	54	240
Total	2239	227	2466

C Composition of the *"Human Colon Global Database"* after purging based on *EpCAM* and *Albumin.*

GEO Dataset	Colon Cancer	Normal Colon	Total
GSE2109	393		393
GSE2361		1	1
GSE4045	37		37
GSE4107	10	10	20
GSE4183	15	8	23
GSE5851	22		22
GSE8671		32	32
GSE9254		18	18
GSE9348	70	12	82
GSE10714	7	3	10
GSE11831		17	17
GSE10961	4		4
GSE12945	62		62
GSE13067	73		73
GSE13294	155		155
GSE13471	4	4	8
GSE14333	224		224
GSE15960	6	6	12
GSE17538	64		64
GSE18088	53		53
GSE18105	94	16	110
GSE20916	91	44	135
GSE26682	329		329
GSE26906	57		57
GSE29623	1		1
GSE31595	37		37
GSE37892	129		129
GSE41258	178	43	221
Total	2115	214	2329

B Purging of the *"Human Colon Global Database"* based on *EpCAM* and *Albumin* gene-expression levels

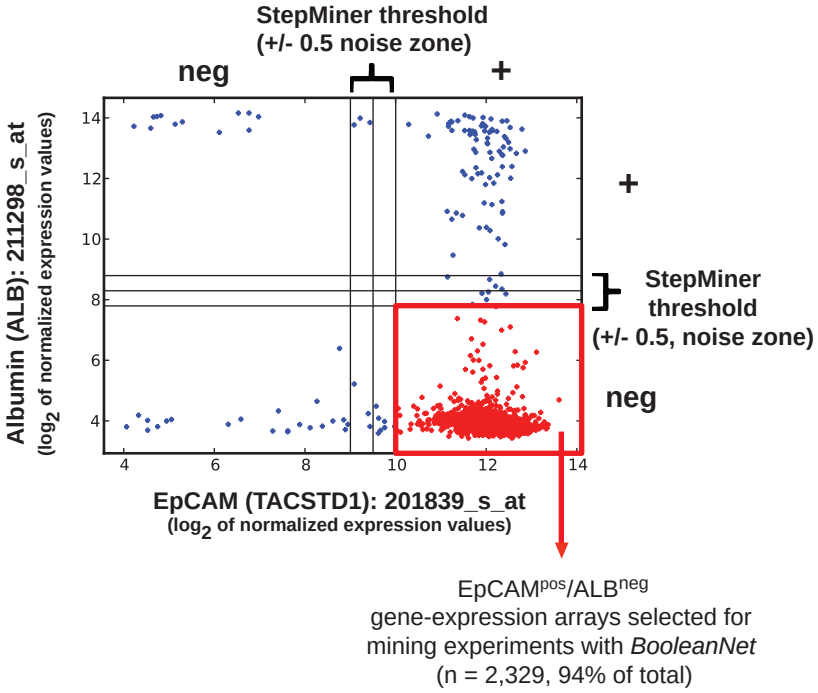
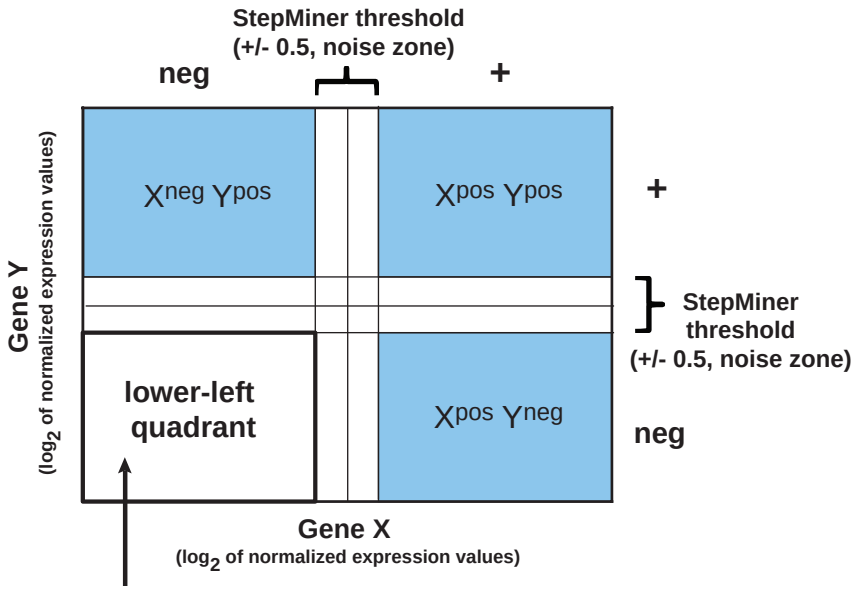


Figure S1. **Assembly and purging of the *"Human Colon Global Database".*** From the *"Human NCBI-GEO Global Database"*, we extracted 2,466 human gene-expression array experiments, belonging to 28 independent GEO data-series (GSEs), and performed on primary tissue samples of either colon cancer (n = 2,239) or normal colon epithelium (n = 227; Panel A). A detailed description of the 28 GEO data-series (GSEs) included in this study is provided in Table S1. To minimize the risk that mining 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, we focused our analysis on the subset of arrays whose gene-expression profile could be defined as *EpCAM^{pos}/Albumin^{neg}* (Panel B). 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 the 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 values, which is the minimum noise level in these type of datasets (Sahoo *et al., Genome Biology,* 9:R157, 2008).[1] All the data below the intermediate region (< *StepMiner* threshold - 0.5) were considered negative, and all above the intermediate region (> *StepMiner* threshold + 0.5) were considered positive. Based on these rules, *EpCAM^{pos}* samples were defined as Affymetrix probe 201839_s_at > 10.00 (i.e. 9.5+0.5), and *ALB^{neg}* samples were defined as Affymetrix probe 211298_s_at < 7.8 (i.e. 8.3-0.5). The *"purging"* operation removed 137 arrays (6%) and left 2,329 arrays (94%) for subsequent analysis (Panel C). A detailed description of all these operations is also provided in the Supplementary Methods. A complete list of all individual NCBI-GEO sample number identifiers (GSMIDs) of the 2,466 non-redundant experiments used to assemble the *"Human Colon Global Database"* is provided in Table S3.

Figure S2. High-throughput mining of gene-expression databases using Boolean logic

A Boolean Implication analysis: *"X^{neg} implies Y^{pos}"*

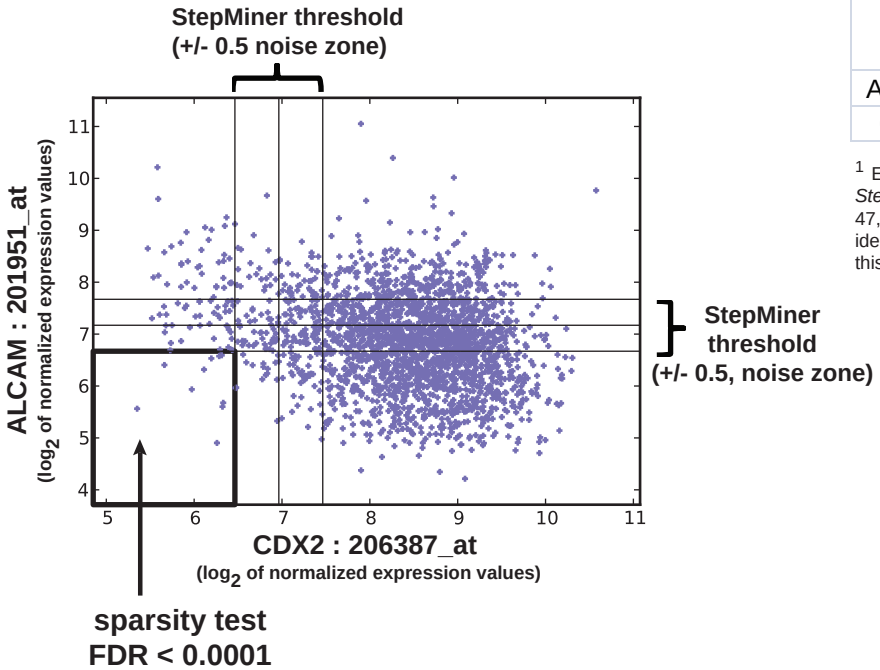


B *"Human NCBI-GEO Global Database"* [1] used for the calculation of gene-expression thresholds with the *StepMiner* algorithm

Affymetrix® Platform	No of Experiments
U133 Plus 2.0	25,955
U133A	17,001
U133A 2.0	4,033
HT-U133A	251
Total	47,240

[1] All experiments are publicly available and can be downloaded from the NCBI-GEO website (http://www.ncbi.nlm.nih.gov/geo) as of February 1st, 2015. A complete list of all individual NCBI-GEO sample number identifiers (GSMIDs) of the 47,240 non-redundant experiments used to assemble the *"Human NCBI-GEO Global Database"* is provided in Table S2.

C Boolean implication analysis: *"$CDX2^{neg}$ implies $ALCAM^{pos}$"*
"Human Colon Global Database" (n = 2,329)



D *ALCAM* and *CDX2* gene-expression thresholds[1]

Gene	Affymetrix® U133A probe set	*StepMiner* Threshold	*StepMiner* Threshold - 0.5
ALCAM	201951_at	7.17	6.67
CDX2	206387_at	6.96	6.46

[1] Expressed as $\log_2$ of gene-expression values and calculated using the *StepMiner* algorithm on the *"Human NCBI-GEO Global Database"* (n = 47,240). A complete list of all individual NCBI-GEO sample number identifiers (GSMIDs) of the 47,240 non-redundant experiments that form this database is provided in Table S2.

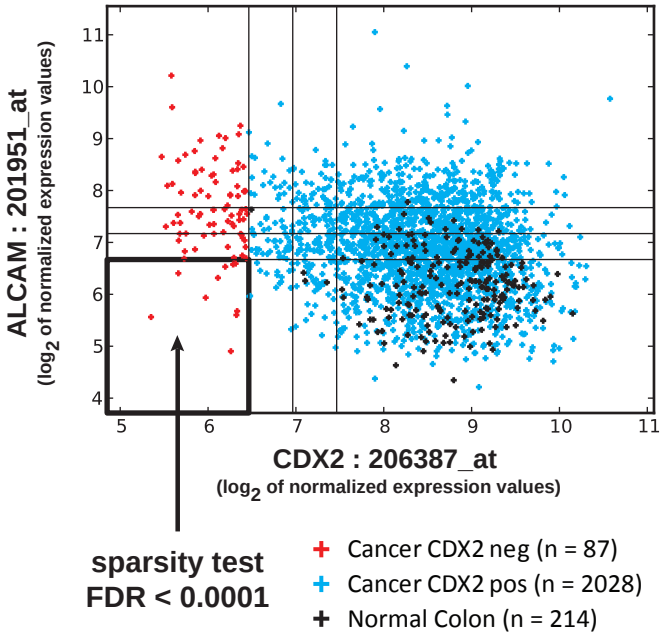
Figure S2. **High-throughput mining of gene-expression databases using Boolean logic.** To identify pairs of genes whose expression is regulated by Boolean implications, we exploited the previously described *BooleanNet* software algorithm (Sahoo *et al.*, *Genome Biology*, 9:R157, 2008).[2] In this study, we performed a search based on a Boolean implication of the "X^{neg} implies Y^{pos}" type (Panel A). Gene-expression patterns were considered to fulfill this type of implication when the false-discovery rate (FDR) of a sparsity test in the lower left quadrant was < 0.0001 (10^{-4}). Threshold gene expression levels were calculated using the *StepMiner* algorithm, based on the expression distribution of the 47,240 gene-expression arrays contained within the *"Human NCBI-GEO Global Database"* (Panel B), and an intermediate region (*"noise zone"*) 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. The fulfillment of the *"X^{neg} implies $ALCAM^{pos}$"* was tested on the *"Human Colon Global Database"* (n = 2.329 samples after *"purging"* based on the fulfillment of the *$EpCAM^{pos}/ALB^{neg}$* condition, as described in Figure S1). Among the genes that fulfilled the *"X^{neg} implies $ALCAM^{pos}$"* relationship was the gene encoding for the homeobox transcription factor CDX2 (Panel C). The threshold gene-expression levels for the lower left quadrant were: 6.67 (i.e. 7.17-0.5) for *ALCAM* (Affymetrix probe 201951_at) and 6.46 (i.e. 6.96-0.5) for *CDX2* (Affymetrix probe 206387_at; Panel D). Gene-expression levels were assigned for each gene in each array, using the log2 of the expression values. A step-by-step, detailed description of all these operations is also provided in the Supplementary Methods.

Figure S3. Identification of CDX2.

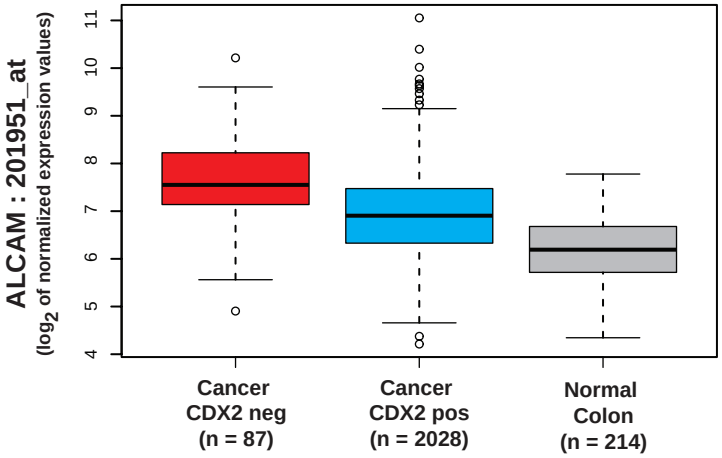
A List of genes fulfilling the Boolean implication "*X*^{neg} *implies ALCAM*^{pos}" (FDR < 10^{-4})

Affymetrix® U133A probe set	Gene Symbol	Dynamic Range
202831_at	GPX2	10.3
206387_at	CDX2	8.59
219404_at	EPS8L3	8.25
210264_at	GPR35	8.12
203287_at	LAD1	8.05
212611_at	DTX4	7.99
206430_at	CDX1	7.88
211184_s_at	USH1C	7.84
205506_at	VIL1	7.77
205137_x_at	USH1C	7.34
220082_at	PPP1R14D	6.64
214898_x_at	MUC3B	6.55
220073_s_at	PLEKHG6	6.5
215420_at	IHH	6
214763_at	ACOT11	5.7
219418_at	NHEJ1	5.38

B Boolean implication analysis "*CDX2*^{neg} *implies ALCAM*^{pos}"



C Box-plot analysis of *ALCAM* mRNA expression in normal and cancer tissues



D Statistical analysis of differences in *ALCAM*[1] mRNA expression levels

	ALCAM mRNA expression levels[2]		2-sample t-test (2-tailed)
population	**mean**	**95% CI[3]**	
Cancer CDX2 neg	7.62	7.42-7.81	p < 0.001
Cancer CDX2 pos	6.89	6.86-6.93	p < 0.001
Normal Colon	6.20	6.11-6.29	

[1] ALCAM is also known as CD166

[2] Affymetrix probe 201951_at, log2 of gene-expression values

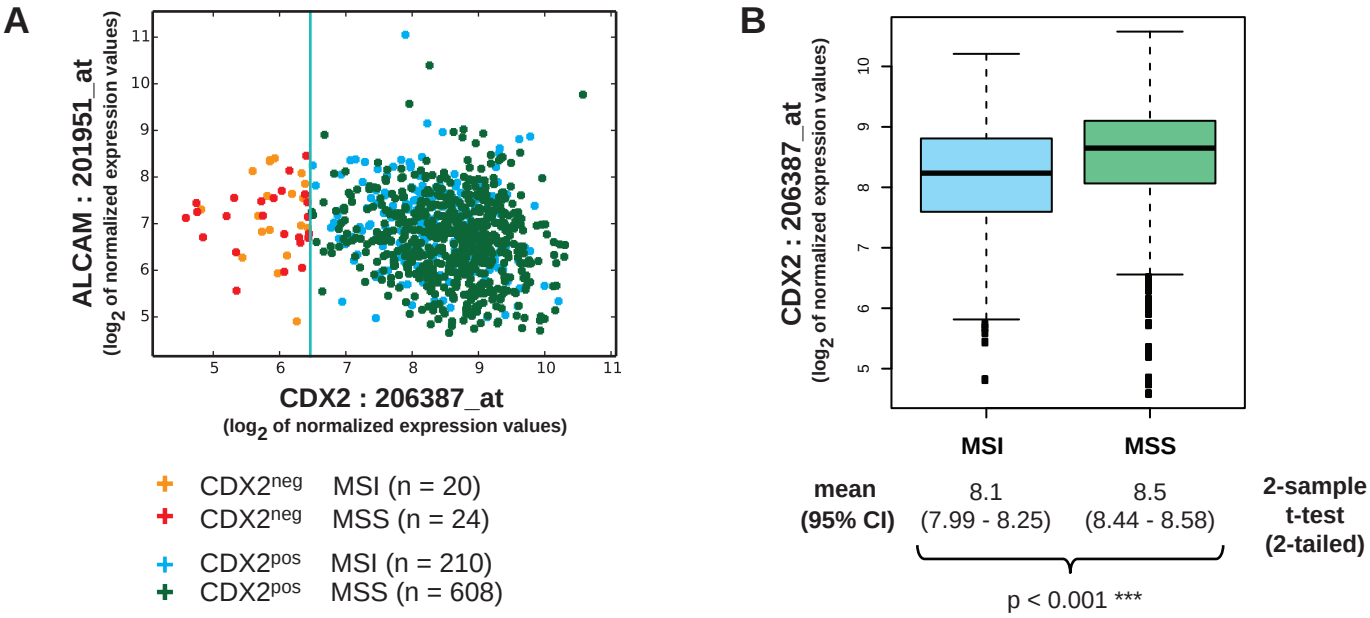
[3] CI: confidence interval

Figure S3. **Identification of CDX2.** A database containing 2,329 human gene expression arrays from both normal colon (n = 214), and colorectal cancer tissue samples (n = 2115), was mined to identify genes that fulfilled the *"X*^{neg} *implies ALCAM*^{pos}*"* Boolean implication. A sparsity test for the lower left quadrant was performed, after threshold definition using he *StepMiner* algorithm and using a false-discovery rate (FDR) < 0.0001 (10^{-4}). This screening yielded 16 candidate genes, that were ranked based on the dynamic range of their gene-expression values (Panel A). Among genes ranking at the top was the homeobox gene CDX2. A visual analysis of *CDX2* and *ALCAM* gene-expression relationships using two-axis scatter plots confirmed the *"CDX2*^{neg} *implies ALCAM*^{pos}*"* Boolean relationship (Panel B). A box-plot analysis (Panel C) indicated that mean ALCAM gene-expression levels were higher in *CDX2*^{neg} colorectal carcinomas (n = 87) as compared to *CDX2*^{pos} ones (n = 2028) and to normal colorectal epithelium (n = 214). A 2-sample t-test to compare mean *ALCAM* gene-expression levels in the three populations indicated that these differences were statistically significant (Panel D).

Figure S4. Relationship between *CDX2* mRNA expression and MSI/MSS status.

Distribution of *CDX2* mRNA expression levels in tumors with MSI and MSS.

Pooled gene-expression datasets (n = 862):
GSE13067, GSE13294, GSE24514, GSE26682, GSE35896, GSE39084, GSE41258



Frequency of *CDX2^{neg}* tumors after stratification for MSI vs. MSS status.

C

MSI/MSS	CDX2 status		% CDX2^{neg}	OR[1] (95% CI[2])
	CDX2^{neg}	CDX2^{pos}		
MSS (n = 632)	24	608	**3.8 %** (24/632)	1
MSI (n = 230)	20	210	**8.7 %** (20/230)	2.4 (1.3-4.5)

Pearson's Chi-square Test
$\chi^2 = 8.4$
p = 0.004 **

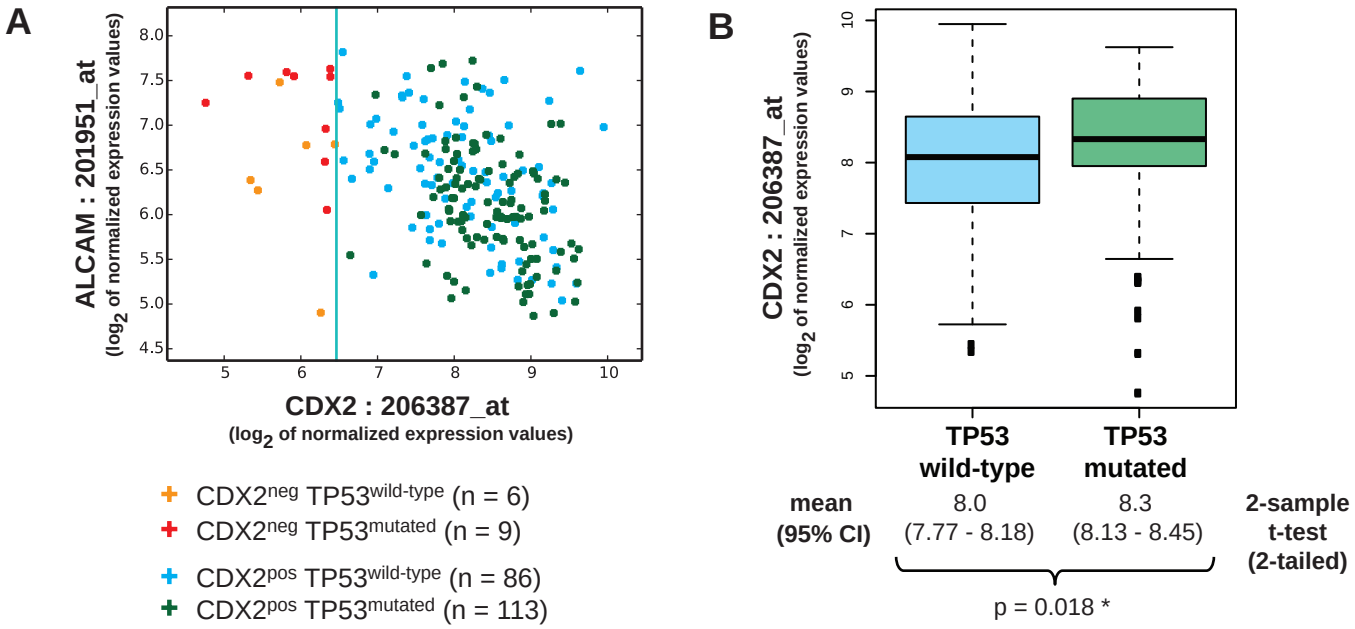
[1]OR: Odds ratio;
[2]CI: Confidence interval

Figure S4. **Relationship between *CDX2* expression and MSI/MSS status.** The relationship between the *CDX2^{neg}* phenotype and microsatellite instability (MSI) was investigated in a database of 862 independent primary colorectal carcinomas annotated for MSI/MSS status (MSI = 230, MSS = 632). The database was assembled by pooling seven independent gene-expression array datasets downloaded from the NCBI-GEO online repository (GSE13067, GSE13294, GSE24514, GSE26682, GSE35896, GSE39084, GSE41258; Tables S1 and S4a). In this database all tumors classified as MSI-low were re-classified as MSS. The database showed a classical tumor distribution with regard to *ALCAM* and *CDX2* mRNA expression levels (Panel A), which recapitulated the distribution observed in the *"Human Colon Global Database"* (Figure S3). The distribution of *CDX2* mRNA expression levels in tumors with and without MSI (MSI vs. MSS) was compared using box-plots, and showed that MSI tumors were characterized by a slightly lower mean *CDX2* mRNA expression value as compared to MSS ones (8.1 vs. 8.5 $\log_2$ of normalized expression values, p < 0.001, Panel B). The association between MSI and the *CDX2^{neg}* status was evaluated using 2x2 contingency tables (Panel C), after stratification of tumors in *CDX2^{neg}* and *CDX2^{pos}* groups based on the *StepMiner-0.5* threshold (Figure S2). Overall, *CDX2^{neg}* tumors showed only limited overlap with tumors characterized by MSI. However, tumors with MSI appeared to be characerized by an enrichment in *CDX2^{neg}* tumors as compared to MSS ones (MSS vs. MSI: 3.8% vs. 8.7%; OR = 2.4, 95%CI = 1.3-4.5; $\chi^2 = 8.4$, p = 0.004).

Figure S5. Relationship between CDX2 mRNA expression and TP53 mutations.

Distribution of *CDX2* mRNA expression levels in TP53$^{wild\text{-}type}$ and TP53^{mutated} tumors.

Pooled gene-expression datasets (n = 214):
GSE39084, GSE41258



Frequency of *CDX2^{neg}* tumors after stratification for the presence of TP53 mutations.

C

TP53 mutation status	CDX2 status		% CDX2^{neg}	OR[1] (95% CI[2])
	CDX2^{neg}	CDX2^{pos}		
wild-type (n = 92)	6	86	**6.5 %** (6/92)	1
mutated (n = 122)	9	113	**7.4 %** (9/122)	1.1 (0.4-3.3)

Pearson's Chi-squared Test
$\chi^2 = 0.06$
p = 0.81

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

Figure S5. **Relationship between *CDX2* mRNA expression and TP53 mutations.** The relationship between the *CDX2^{neg}* phenotype and genetic mutations in the *TP53* gene was investigated in a database of 214 independent primary colorectal carcinomas annotated for TP53 mutational status (TP53$^{wild\text{-}type}$ = 92, TP53^{mutated} = 122). The database was assembled by pooling two independent gene-expression array datasets downloaded from the NCBI-GEO online repository (GSE39084, GSE41258; Tables S1 and S4b), and showed a classical tumor distribution with regard to *ALCAM* and *CDX2* mRNA expression levels (Panel A), which recapitulated the distribution observed in our *"Human Colon Global Database"* (Figure S3). The distribution of *CDX2* mRNA expression levels in tumors with and without intact TP53 (TP53$^{wild\text{-}type}$ vs. TP53^{mutated}) was compared using box-plots, and showed that TP53$^{wild\text{-}type}$ tumors were characterized by a slightly lower mean *CDX2* mRNA expression value as compared to TP53^{mutated} ones (TP53$^{wild\text{-}type}$ vs. TP53^{mutated}: 8.0 vs. 8.3 $\log_2$ of normalized expression values, p = 0.018, Panel B). The association between the TP53^{mutated} and the *CDX2^{neg}* status was evaluated using 2x2 contingency tables, after tumor stratification in *CDX2^{neg}* and *CDX2^{pos}* groups based on the *StepMiner-0.5* threshold (Figure S2). Overall, CDX2^{neg} tumors showed only partial overlap with TP53^{mutated} ones, and their frequency among TP53$^{wild\text{-}type}$ tumors was not statistically different from that observed among TP53^{mutated} ones (TP53$^{wild\text{-}type}$ vs. TP53^{mutated}: 6.5% vs. 7.4%; OR = 1.1, 95%CI = 0.4-3.3; $\chi^2 = 0.06$, p = 0.81, Panel C).

Figure S6. Patient composition of the “Discovery Dataset” (NCBI-GEO).

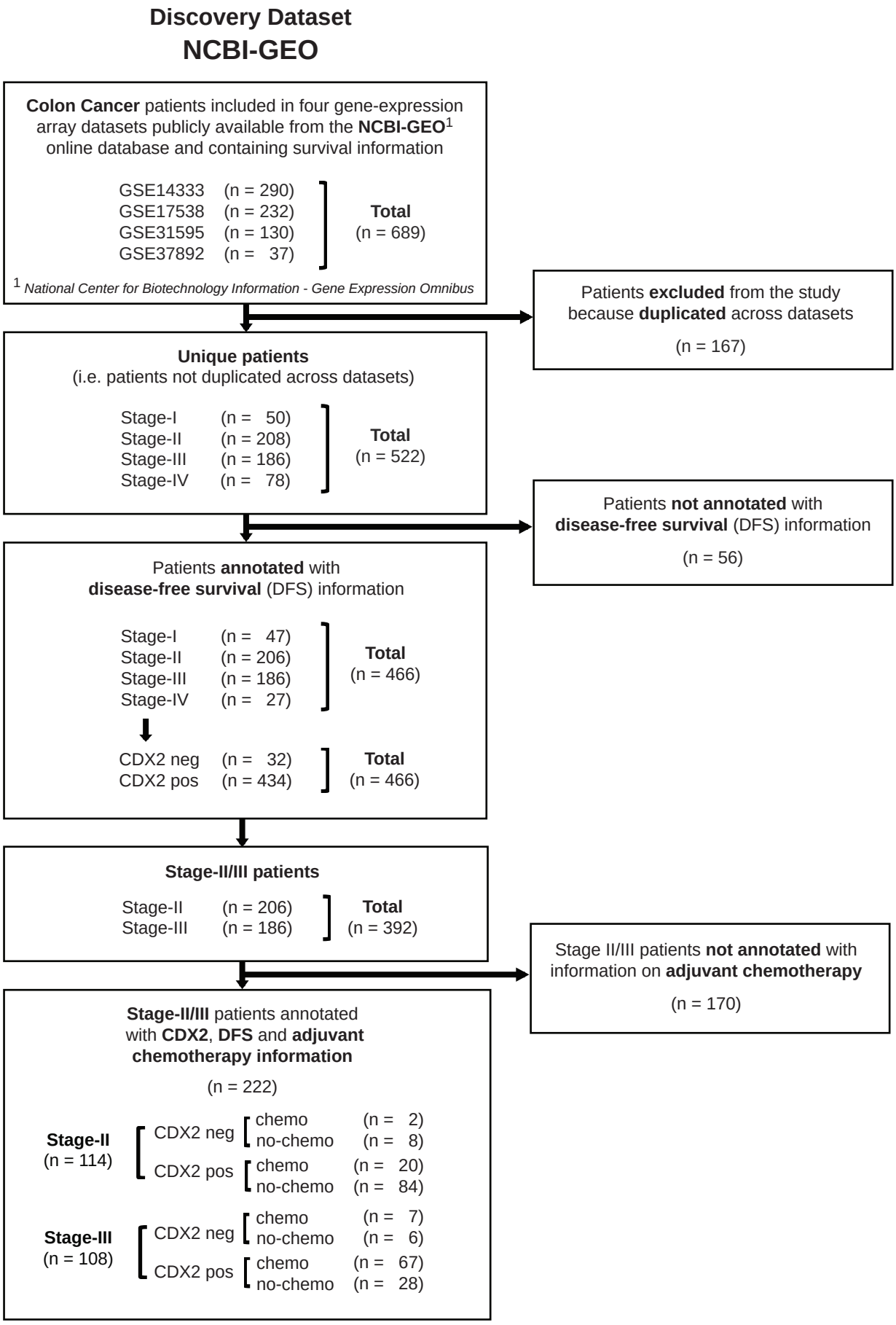
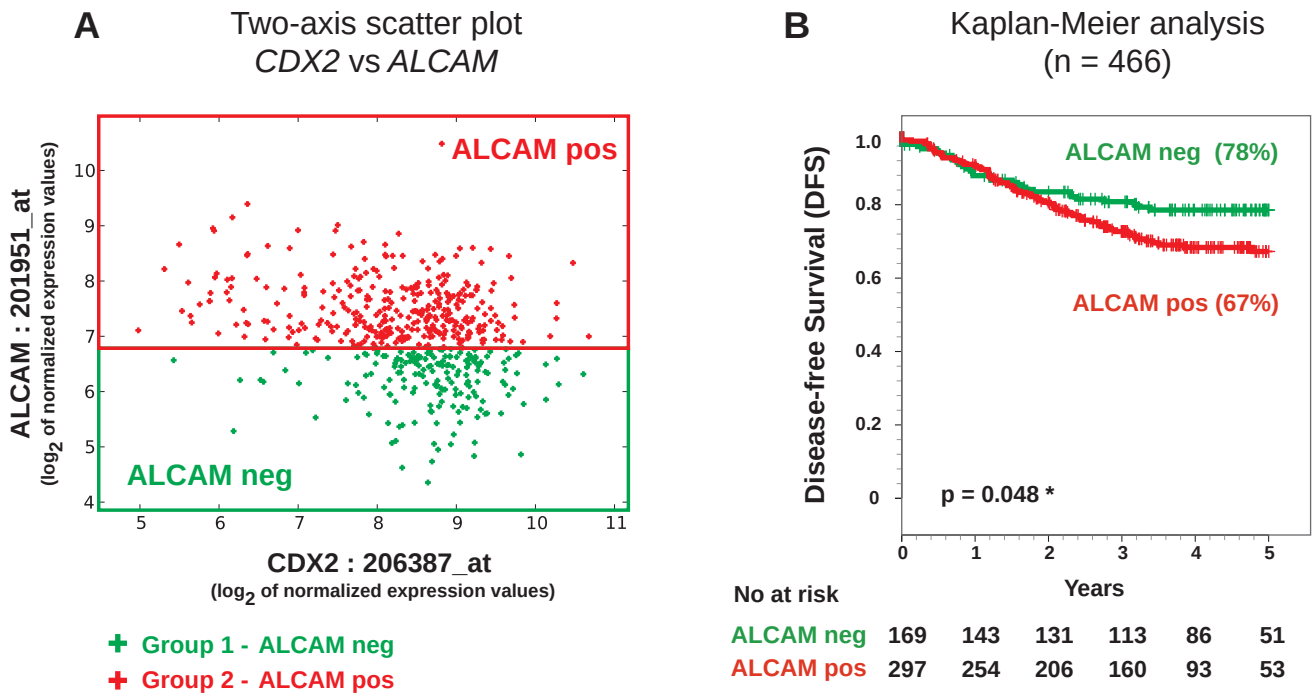


Figure S7. Relationship between *ALCAM* expression and disease-free survival (DFS) in the discovery dataset (NCBI-GEO).



C Multivariate analysis - Cox proportional hazards model

	Univariate				Multivariate			
All patients (n = 466)	**HR[1]**	**95% CI[2]**	**p-value**		**HR[1]**	**95% CI[2]**	**p-value**	
ALCAM	1.49	1.00 - 2.23	**0.050**	*	1.66	1.11 - 2.49	**0.014**	*
Stage (I-IV)	3.47	2.62 - 4.59	**< 0.001**	***	3.47	2.61 - 4.60	**< 0.001**	***
Age[3]	0.99	0.97 - 1.00	0.058		0.99	0.98 - 1.01	0.29	
Sex (M/F)[4]	1.07	0.89 - 1.28	0.49		1.07	0.89 - 1.29	0.48	

Patients annotated with grading information (n = 216)	**HR[1]**	**95% CI[2]**	**p-value**		**HR[1]**	**95% CI[2]**	**p-value**	
ALCAM	2.19	1.15 - 4.17	**0.017**	*	2.33	1.21 - 4.49	**0.011**	*
Stage (I-IV)	3.13	2.14 - 4.60	**< 0.001**	***	3.15	2.10 - 4.73	**< 0.001**	***
Grade (G1-G3)	1.63	0.94 - 2.82	0.08		1.17	0.66 - 2.06	0.59	
Age[3]	0.99	0.97 - 1.01	0.20		0.99	0.97 - 1.02	0.54	
Sex (M/F)[4]	1.15	0.88 - 1.51	0.32		1.17	0.87 - 1.57	0.29	

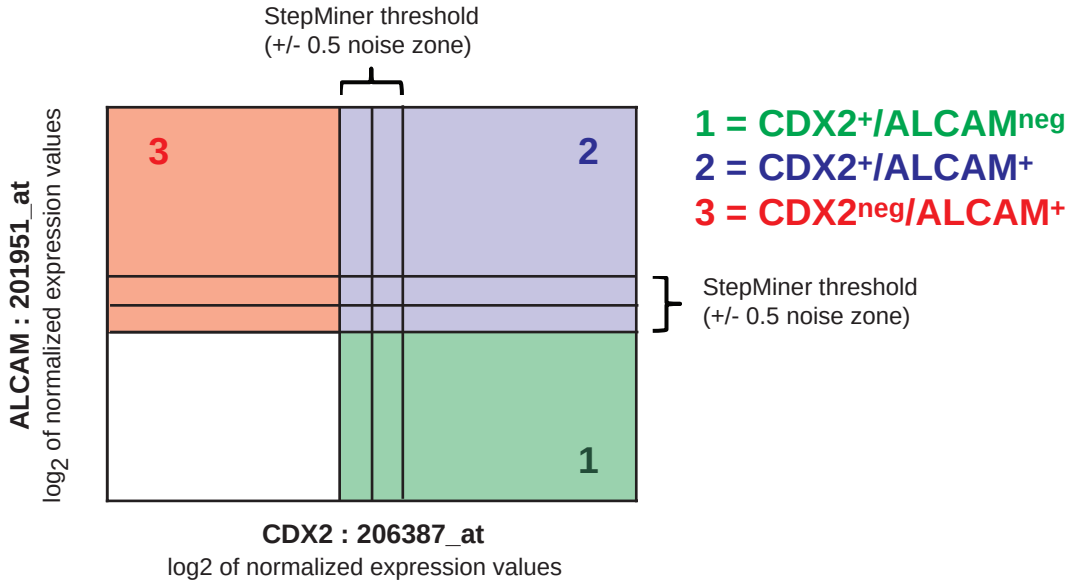
[1] HR: hazard ratio,
[2] CI: confidence interval
[3] Age modeled as a continuous variable
[4] M/F: male vs female

* p < 0.05
** p < 0.01
*** p < 0.001

Figure S7. **Relationship between *ALCAM* expression and DFS in the NCBI-GEO discovery dataset.** To study the relationship between *ALCAM* mRNA expression levels and DFS, we used the *StepMiner* algorithm[1] to stratify the NCBI-GEO discovery dataset (n = 466) in two groups: *ALCAM*^{neg} (< *StepMiner* threshold - 0.5) and *ALCAM*^{pos} (> *StepMiner* threshold - 0.5). The threshold for *ALCAM*^{neg} vs. *ALCAM*^{pos} mRNA expression levels was: 6.8 (Affymetrix probe 201951_at; 7.3 – 0.5 = 6.8; Panel A). *ALCAM*^{pos} tumors displayed a statistically significant association with reduced 5-year DFS, although the effect appeared of relatively small magnitude (*ALCAM*^{pos} vs. *ALCAM*^{neg}: 67% vs. 78%, p = 0.048, Panel B). This finding remained statistically significant when tested in mutivariate analysis using the Cox proportional hazards method (Panel C).

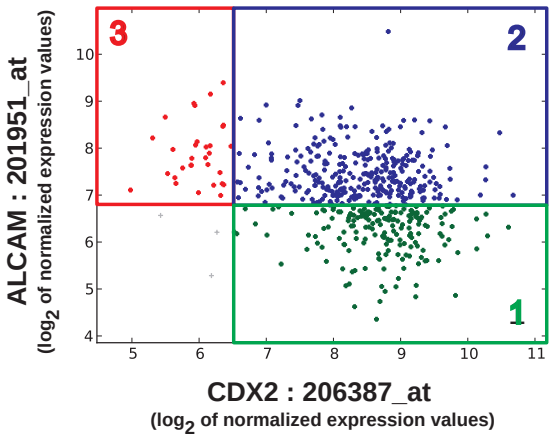
Figure S8. Definition of *CDX2/ALCAM* gene-expression groups for survival analysis

A General approach to define gene-expression groups using the *Hegemon* software



B Definition of *CDX2/ALCAM* gene-expression groups

NCBI-GEO discovery dataset (n = 466)



C *ALCAM* and *CDX2* gene-expression thresholds [1]

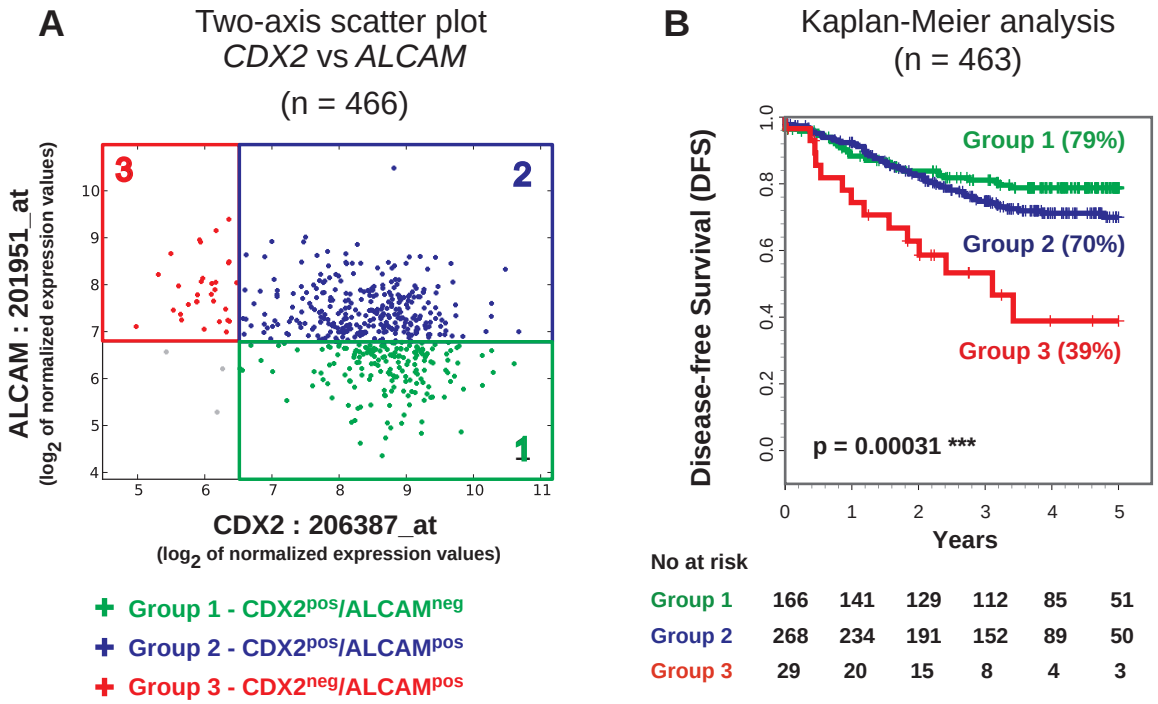
Gene	Affymetrix® U133 Plus 2.0 probe set	*StepMiner* Threshold	*StepMiner* Threshold - 0.5
ALCAM	201951_at	7.3	6.8
CDX2	206387_at	7.0	6.5

[1] Calculated using the *StepMiner* algorithm on the HG-U133 Plus 2.0 subset (n = 25,955) of the *"Human NCBI-GEO Global Database"*. A complete list of all individual NCBI-GEO sample number identifiers (GSMIDs) of the 25,955 non-redundant experiments that form this subset database is provided in Table S2b.

[2] $\log_2$ of normalized expression values

Figure S8. **Definition of *CDX2/ALCAM* gene-expression groups for survival analysis.** We used the *Hegemon* software tool to stratify the NCBI-GEO discovery dataset into three subgroups, from more to less differentiated (Panel A): *CDX2^{pos}/ALCAM^{neg}* (Group 1, green), *CDX2^{pos}/ALCAM^{pos}* (Group 2, blue), *CDX2^{neg}/ALCAM^{pos}* (Group 3, red). The thresholds used to define positive and negative samples were calculated using the *StepMiner* algorithm[1] 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. All patients whose gene-expression values were below the intermediate region (< *StepMiner* threshold - 0.5) were considered negative, and the rest (> *StepMiner* threshold - 0.5) were considered positive. The NCBI-GEO discovery dataset was obtained by pooling four NCBI-GEO data-series (GSE14333, GSE17538, GSE31595, GSE37892), which contained data from 466 independent colon cancer patients, all annotated with disease-free survival (DFS) information (Panel B; see also Table S1 and Table S5a). Since this database contained data generated exclusively on the Affymetrix HG-U133 Plus 2.0 platform (GPL570), *CDX2* and *ALCAM StepMiner* gene-expression thresholds were calculated on the Affymetrix HG-U133 Plus 2.0 subset of the *"Human Global NCBI-GEO Database"* (n = 25,955; Table S2b). The threshold gene-expression levels were: 6.8 for *ALCAM* (Affymetrix probe 201951_at; 7.3 – 0.5 = 6.8) and 6.5 for *CDX2* (Affymetrix probe 206387_at; 7.0 – 0.5 = 6.5; Panel C). A step-by-step, detailed description of all these operations is also provided in the Supplementary Methods.

Figure S9. Relationship between *CDX2/ALCAM* gene-expression groups and disease-free survival (DFS) in the discovery dataset (NCBI-GEO).



C **Multivariate analysis - Cox proportional hazards model**

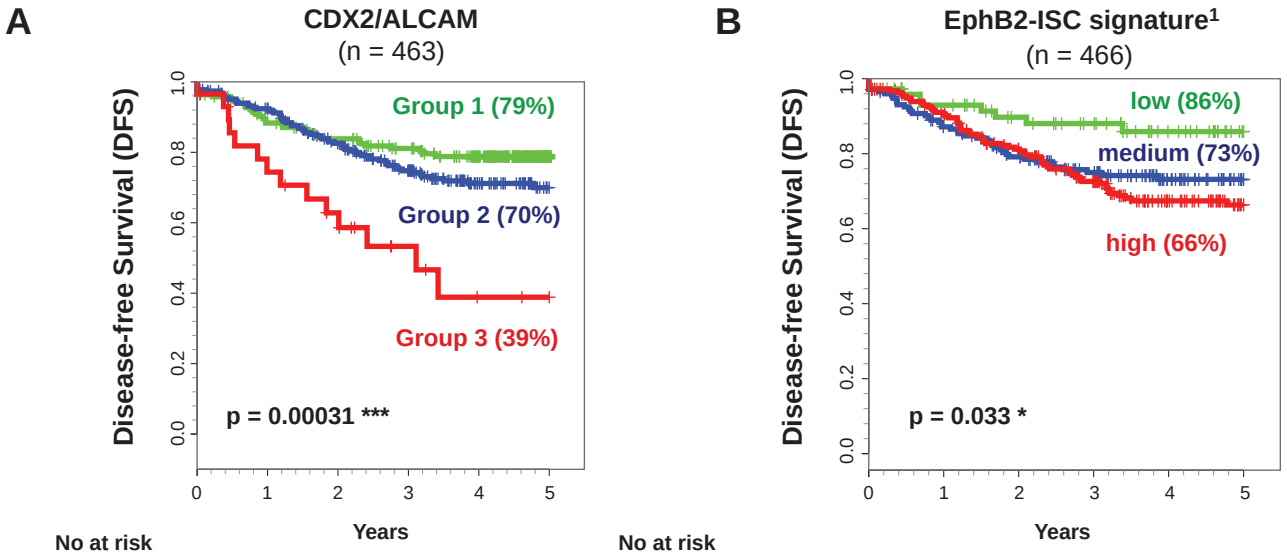
		Univariate				Multivariate		
	HR[1]	**95% CI[2]**	**p-value**		**HR[1]**	**95% CI[2]**	**p-value**	
All patients (n = 463)								
CDX2/ALCAM	1.69	1.22 - 2.34	**0.002**	**	1.84	1.33 - 2.55	**< 0.001**	***
Stage (I-IV)	3.47	2.62 - 4.59	**< 0.001**	***	3.53	2.65 - 4.70	**< 0.001**	***
Age[3]	0.99	0.97 - 1.00	**0.044**	*	0.99	0.98 - 1.00	0.19	
Sex (M/F)[4]	1.06	0.88 - 1.27	0.55		1.06	0.88 - 1.28	0.56	
Patients annotated with grading information (n = 214)								
CDX2/ALCAM	2.14	1.37 - 3.34	**< 0.001**	***	2.47	1.51 - 4.06	**< 0.001**	***
Stage (I-IV)	3.14	2.13 - 4.62	**< 0.001**	***	3.36	2.20 - 5.14	**< 0.001**	***
Grade (G1-G3)	1.71	0.98 - 2.98	0.059		1.02	0.58 - 1.82	0.93	
Age[3]	0.99	0.97 - 1.01	0.15		0.99	0.97 - 1.01	0.29	
Sex (M/F)[4]	1.12	0.85 - 1.48	0.40		1.19	0.88 - 1.61	0.26	

[1] HR: hazard ratio,
[2] CI: confidence interval
[3] Age modeled as a continuous variable
[4] M/F: male vs female

* p < 0.05
** p < 0.01
*** p < 0.001

Figure S9. **Relationship between *CDX2/ALCAM* gene-expression groups and DFS in the NCBI-GEO discovery dataset.** To study the relationship between *CDX2/ALCAM* mRNA expression levels and DFS, we used the *Hegemon* software tool[10] to visualize three distinct gene-expression groups, from more to less differentiated: Group 1 (green), defined as *CDX2^{pos}/ALCAM^{neg}*; Group 2 (blue), defined as *CDX2^{pos}/ALCAM^{pos}*; Group 3 (red), defined as *CDX2^{neg}/ALCAM^{pos}* (Panel A). The gene-expression groups correlated with progressively worse prognosis both in Kaplan-Meier ($p < 0.001$, Panel B) and in univariate Cox proportional hazards analysis ($p < 0.001$, Panel C). Multivariate analysis of survival data indicated that the association between *CDX2/ALCAM* gene-expression groups and reduced 5-year DFS was not confounded by stage, pathological grade, age or gender ($p < 0.001$, Panel C) and was superior to grade, age and gender in magnitude of hazard ratio (HR, Panel C).

Figure S10. The association between *CDX2/ALCAM* gene-expression subgroups and reduced DFS is not confounded by the EphB2-ISC gene-expression signature.



No at risk						
Group 1	166	141	129	112	85	51
Group 2	268	234	191	152	89	50
Group 3	29	20	15	8	4	3

- Group 1 - CDX2^{pos}/ALCAM^{neg}
- Group 2 - CDX2^{pos}/ALCAM^{pos}
- Group 3 - CDX2^{neg}/ALCAM^{pos}

No at risk						
EphB2-ISC low	74	64	53	46	32	17
EphB2-ISC medium	175	145	125	99	65	34
EphB2-ISC high	217	188	159	128	82	53

- low expression of the EphB2-ISC signature
- medium expression of the EphB2-ISC signature
- high expression of the EphB2-ISC signature

1 EphB2-ISC = gene expression signature of Ephrin-B2+ intestinal stem cells (ISC) according to Merlos-Suarez *et al.* (*Cell Stem Cell* , 8:511-524, 2011).

C Cox proportional hazards model

All patients (n = 463)

	Univariate				Multivariate			
	HR[1]	95% CI[2]	p-value		HR[1]	95% CI[2]	p-value	
CDX2/ALCAM	1.69	1.22 - 2.34	**0.002**	**	1.74	1.25 - 2.44	**0.001**	**
EphB2-ISC[3]	1.39	1.06 - 1.82	**0.018**	*	1.24	0.93 - 1.65	0.14	
Stage (I-IV)	3.47	2.62 - 4.59	**< 0.001**	***	3.54	2.65 - 4.73	**< 0.001**	***
Age[4]	0.87	0.75 - 1.00	**0.044**	*	0.91	0.79 - 1.06	0.22	
Sex (M/F)[5]	1.06	0.88 - 1.27	0.55		1.07	0.89 - 1.29	0.47	

Patients annotated with grading information (n = 214)

	HR[1]	95% CI[2]	p-value		HR[1]	95% CI[2]	p-value	
CDX2/ALCAM	2.14	1.37 - 3.34	**< 0.001**	***	2.37	1.43 - 3.92	**< 0.001**	***
EphB2-ISC[3]	1.47	0.97 - 2.23	0.069		1.22	0.78 - 1.90	0.38	
Stage (I-IV)	3.14	2.13 - 4.62	**< 0.001**	***	3.31	2.15 - 5.09	**< 0.001**	***
Grade (G1-G3)	1.71	0.98 - 2.98	0.059		1.11	0.60 - 2.03	0.74	
Age[4]	0.86	0.71 - 1.05	0.15		0.88	0.70 - 1.11	0.29	
Sex (M/F)[5]	1.12	0.85 - 1.48	0.40		1.22	0.90 - 1.65	0.21	

[1] HR: hazard ratio,
[2] CI: confidence interval
[3] EphB2-ISC: "low" vs "medium" vs "high"
[4] Age modeled as a continuous variable
[5] M/F: male vs female

* p < 0.05
** p < 0.01
*** p < 0.001

Figure S10. **The association between *CDX2/ALCAM* gene-expression subgroups and reduced DFS is not confounded by the EphB2-ISC gene-expression signature.** Both the *CDX2/ALCAM* grouping system and the gene-expression signature derived from Ephrin-B2^{+} mouse intestinal stem cells (EphB2-ISC; Merlos-Suarez *et al., Cell Stem Cell*, 8:511-524, 2011)[11] can be used to stratify colon cancer patients in subgroups characterized by different 5-year DFS rates (Panel A and Panel B). A multivariate analysis comparing the two scoring systems (*CDX2/ALCAM* gene-expression vs. EphB2-ISC signature) indicated that the association between *CDX2/ALCAM* subgroups and reduced 5-year DFS is not confounded by the EphB2-ISC signature (p = 0.0011) and is associated to a higher hazard-ratio (HR) as compared to the EphB2-ISC signature (Panel C).

Figure S11. Patient composition of the “Validation Dataset” (NCI-CDP).

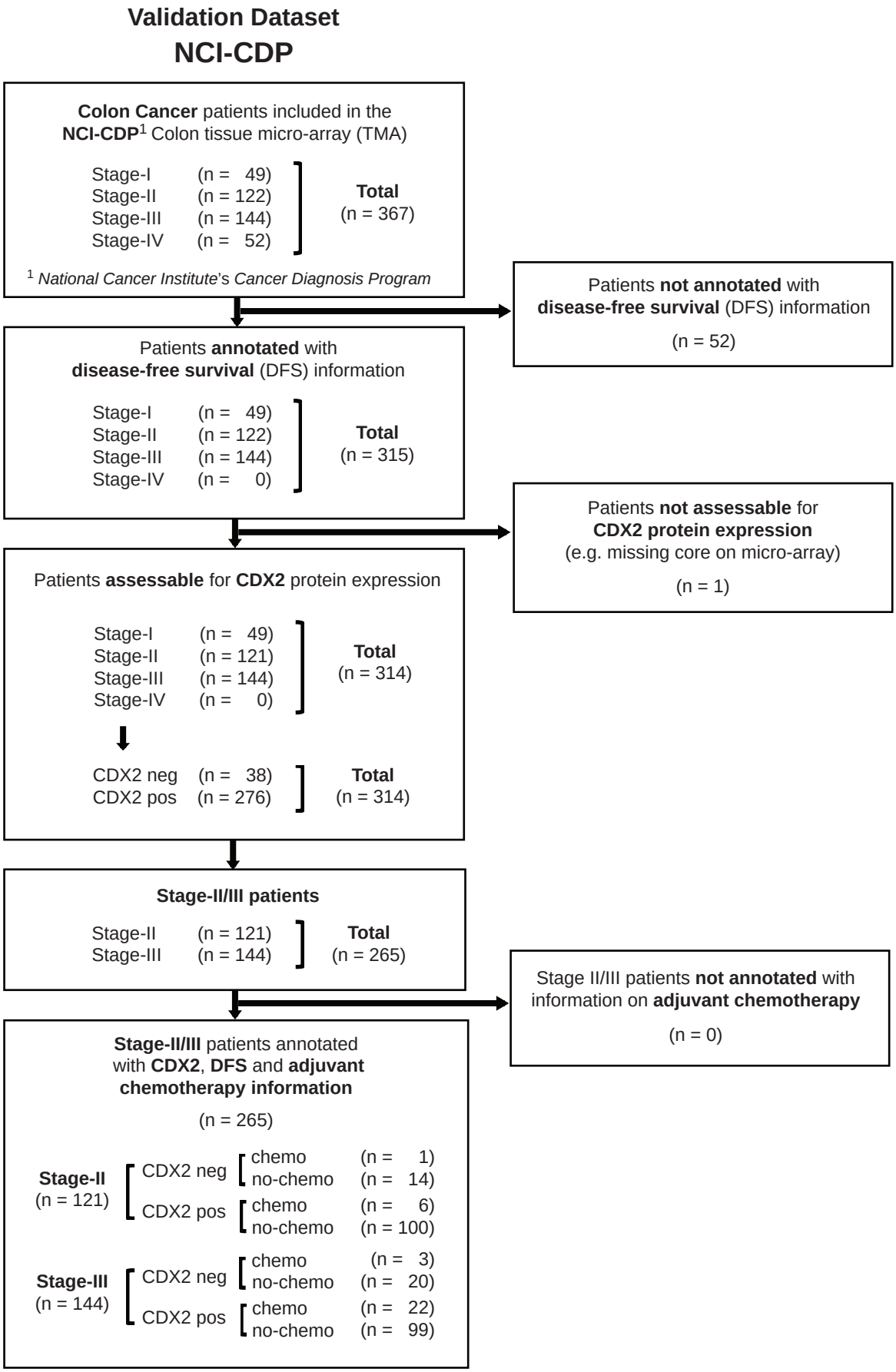


Figure S12. Patient composition of the NSABP-C07 dataset.

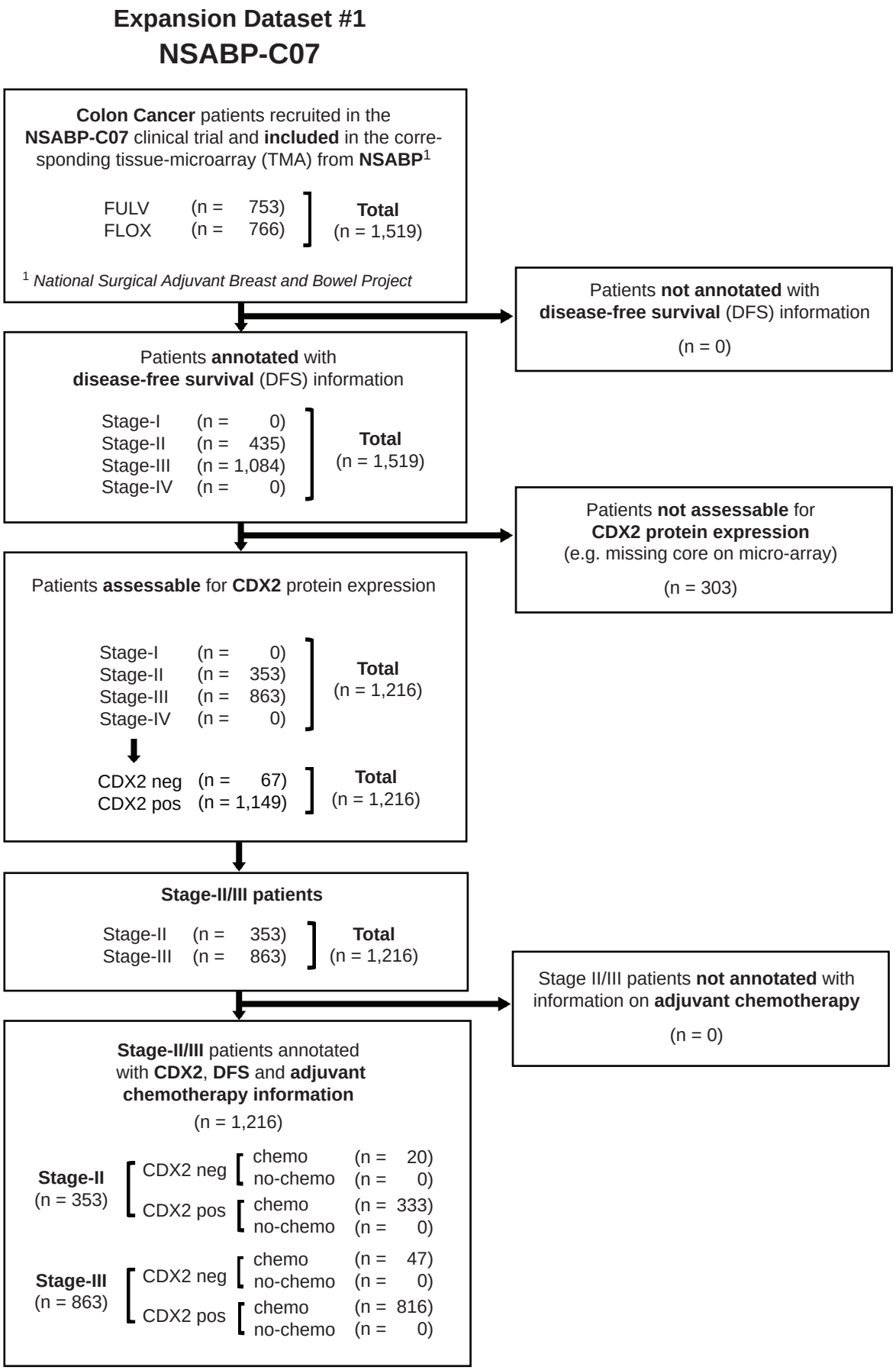


Figure S13. Patient composition of the Stanford TMAD dataset.

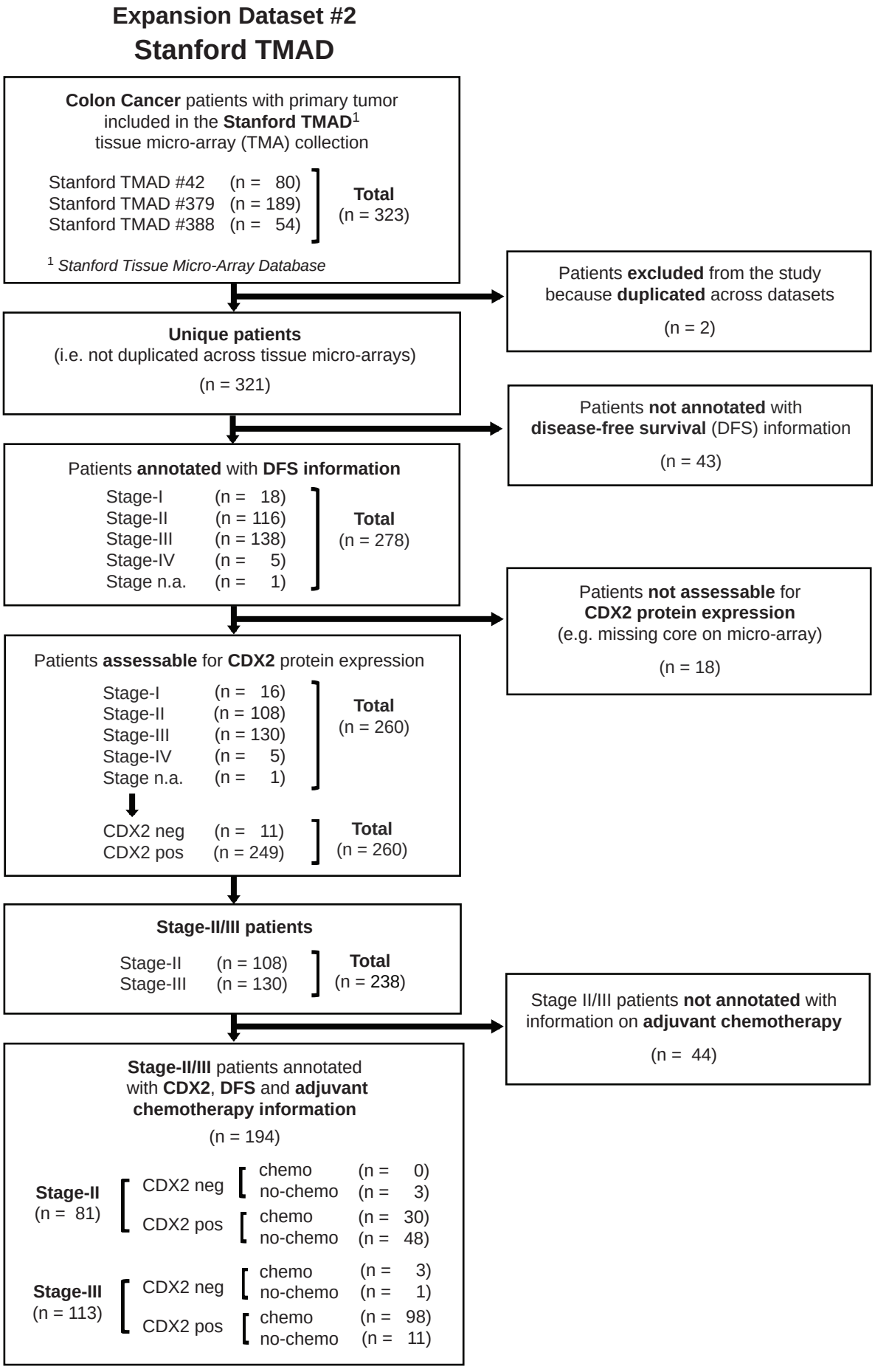
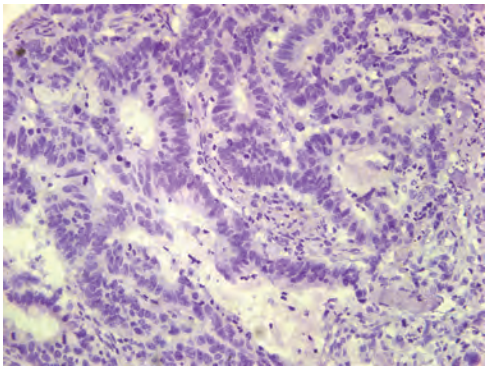


Figure S14. Scoring system for CDX2 protein expression in immunohistochemistry.



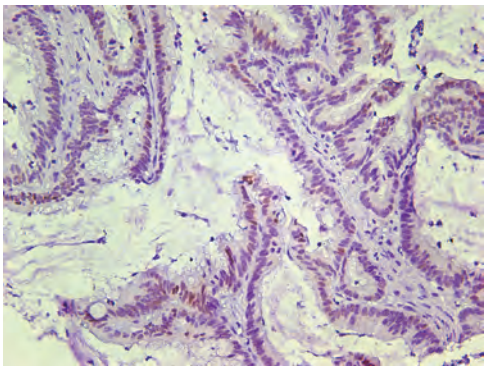
A

Score 0

No staining

Observed frequency (NCI-CDP):

whole population n = 34/366 (9%)
subset with DFS n = 27/314 (9%)



B

Score 0.5

Weak/scattered staining in a minority of cells

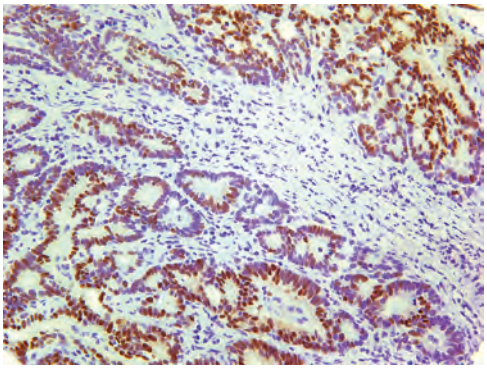
Observed frequency (NCI-CDP):

whole population n = 14/366 (4%)
subset with DFS n = 11/314 (3%)

CDX2 negative

Observed frequency (NCI-CDP):

whole population n = 48/366 (13%)
subset with DFS n = 38/314 (12%)



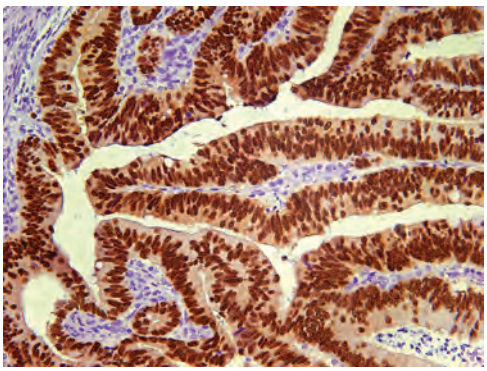
C

Score 2

Moderate/strong staining in a majority of cells

Observed frequency (NCI-CDP):

whole population n = 36/366 (10%)
subset with DFS n = 31/314 (10%)



D

Score 3

Strong staining in all cells

Observed frequency (NCI-CDP):

whole population n = 282/366 (77%)
subset with DFS n = 245/314 (78%)

CDX2 positive

Observed frequency (NCI-CDP):

whole population n = 318/366 (87%)
subset with DFS n = 276/314 (88%)

Figure S14. **Scoring system for CDX2 protein expression in immunohistochemistry.** Based on our gene-expression data, we decided to stratify colon carcinomas in two categories: $CDX2^{neg}$ or $CDX2^{pos}$. We scored as $CDX2^{neg}$ all tumors whose malignant epithelial component either completely lacked CDX2 expression or showed faint nuclear expression in a minority of malignant epithelial cells, a feature observed in 13% (n = 48/366) of colon carcinomas in the NCI-CDP validation dataset. Tumors scored as $CDX2^{neg}$ fell into two staining patterns: a) complete lack of CDX2 expression (Score 0), observed in 9% (n = 34/366) of NCI-CDP cases (Panel A); b) scattered and faint nuclear expression in a minority fraction of cancer cells (Score 0.5), observed in 4% (n = 14/366) of NCI-CDP cases (Panel B). We scored as $CDX2^{pos}$ all tumors whose malignant epithelial component displayed widespread nuclear expression of CDX2, a feature observed in 87% (n = 318/366) of colon carcinomas in the NCI-CDP validation dataset. Tumors scored as $CDX2^{pos}$ also fell into two staining patterns: a) strong staining in a majority fraction of cancer cells (Score 2), observed in 10% (n = 36/366) of NCI-CDP0 cases (Panel C); b) strong staining in all cancer cells (Score 3), observed in 77% (n = 282/366) of NCI-CDP cases (Panel D). For each tumor, two independent tissue cores from distinct areas of the same lesion were analyzed. Tumors with discordant scores on the two cores were upgraded to the highest score.

Figure S15. Inter-observer agreement in the evaluation of CDX2 protein expression.

Inter-observer agreement in CDX2 scoring results on individual tissue cores (n = 368) [1].

A

CDX2 score of individual tissue cores (0, 0.5, 2, 3)		Observer #1 Score 0	Score 0.5	Score 2	Score 3
Observer #2	Score 0	33	0	0	0
	Score 0.5	5	6	0	0
	Score 2	1	3	43	23
	Score 3	0	0	4	224

Cohen's Kappa Index (with linear weighting)

$\kappa = 0.87$

(95% CI [2] = 0.83 - 0.91)

[2] CI: confidence interval

[1] The contingency table reports scoring data on 342 independent tissue cores (93%, 342/368) and excludes 26 cores (7%, 26/368) that either one observer (n = 18) or both observers (n = 8) deemed not-assessable (e.g. no recognizable tumor tissue, damaged tissue core). The original cohort of 368 cores originated from 184 independent primary tumors, each sampled in two distinct locations.

Agreement in the final evaluation of the CDX2 score in individual patients (n = 184) [2].

B

Final evaluation of patient CDX2 score (0, 0.5, 2, 3)		Observer #1 Score 0	Score 0.5	Score 2	Score 3
Observer #2	Score 0	15	0	0	0
	Score 0.5	3	2	0	0
	Score 2	1	0	16	12
	Score 3	0	0	3	127

Cohen's Kappa Index (with linear weighting)

$\kappa = 0.85$

(95% CI [2] = 0.78 - 0.92)

[2] CI: confidence interval

[2] The contingency table reports scoring data on 179 independent patients (97%, 179/184) and excludes 5 patients (3%, 5/184) that at least one observer deemed not-assessable (e.g. absence of recognizable tumor tissue in both cores from the same patient). For each patient, the final CDX2 score corresponds to the higher of the two CDX2 scores obtained from the two paired cores.

Agreement in the final evaluation of the CDX2 status in individual patients (n = 184) [3].

C

Final evaluation of patient status $CDX2^{neg}$ vs. $CDX2^{pos}$		Observer #1 $CDX2^{neg}$	$CDX2^{pos}$
Observer #2	$CDX2^{neg}$	20	0
	$CDX2^{pos}$	1	158

Cohen's Kappa Index (no weighting)

$\kappa = 0.97$

(95% CI [2] = 0.92 - 1.00)

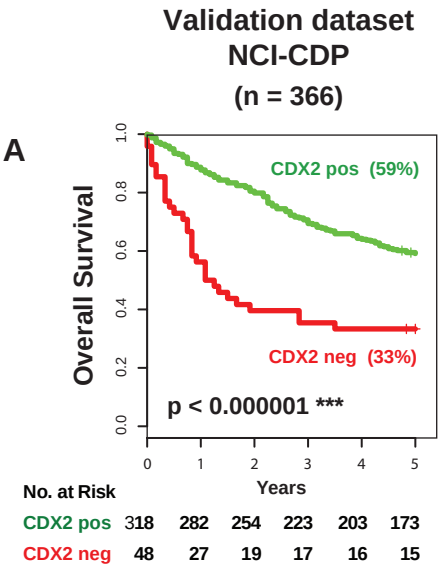
[2] CI: confidence interval

[3] The contingency table reports data on 179 independent patients (97%, 179/184) and excludes 5 patients (3%, 5/184) that at least one obesrver deemed not-assessable (e.g. absence of recognizable tumor tissue in both cores from the same patient). For each patient, the final CDX2 status is defined $CDX2^{neg}$ for tumors with scores of 0 and 0.5, and $CDX2^{pos}$ for tumors with scores of 2 and 3.

Figure S15. **Inter-observer agreement in the evaluation of CDX2 protein expression.** Two independent investigators used the same criteria (Figure S14) to independently score CDX2 protein expression levels in 184 primary colon carcinomas from the NCI-CDP tissue-microarray (TMA), where each tumor was sampled twice, for a total number of 368 independent tissue cores. The concordance between the two observers was analyzed using contingency tables to calculate the Cohen's Kappa Index. The results showed an excellent agreement (k > 0.8), both in terms of CDX2 scoring of the individual cores (Panel A) and in the final CDX2 score of individual patients (Panel B). Most importantly, the results showed a near-perfect agreement (k > 0.97) with regard to the final assessment of the patients' CDX2 status (Panel C).

Figure S16. Relationship between CDX2 expression, overall survival (OS) and disease-specific survival (DSS) in the NCI-CDP validation dataset.

A



B

Overall Survival (OS)

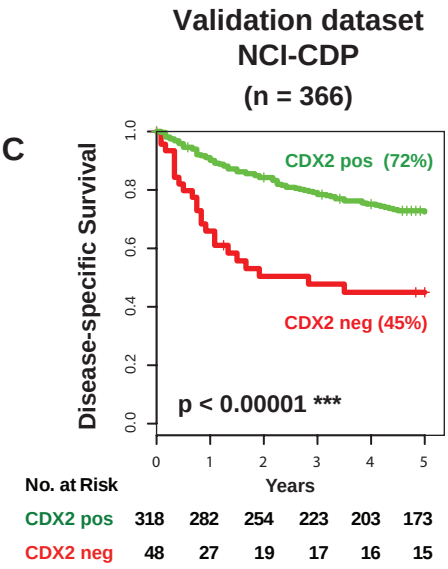
Multivariate analysis - Cox proportional hazards

	HR[1]	95% CI [2]	p-value	
CDX2	1.79	1.18 - 2.71	**0.006**	**
Stage (I-IV)	2.33	1.93 - 2.81	**< 0.001**	***
Grade (G1-G3)	0.90	0.75 - 1.07	0.22	
Age[3]	1.03	1.02 - 1.05	**< 0.001**	***
Sex (M/F)[4]	0.92	0.70 - 1.20	0.53	

[1] HR: hazard ratio,
[2] CI: confidence interval
[3] Age modeled as a continuous variable
[4] M/F: male vs female

* p < 0.05
** p < 0.01
*** p < 0.001

C



D

Disease-specific Survival (DSS)

Multivariate analysis - Cox proportional hazards

	HR[1]	95% CI [2]	p-value	
CDX2	2.09	1.22 - 3.58	**0.007**	**
Stage (I-IV)	5.33	3.94 - 7.21	**< 0.001**	***
Grade (G1-G3)	0.98	0.78 - 1.22	0.83	
Age[3]	1.01	1.00 - 1.03	0.11	
Sex (M/F)[4]	0.96	0.66 - 1.39	0.83	

[1] HR: hazard ratio
[2] CI: confidence interval
[3] Age modeled as a continuous variable
[4] M/F: male vs female

* p < 0.05
** p < 0.01
*** p < 0.001

Figure S16. **Relationship between CDX2 expression, overall survival (OS) and disease-specific survival (DSS) in the NCI-CDP validation dataset.** CDX2^{neg} carcinomas were associated with reduced 5-year overall survival (OS, CDX2^{neg} vs. CDX-2^{pos}: 33% vs. 59%, p<0.001, Panel A) and reduced 5-year disease-specific survival (DSS, CDX2^{neg} vs. CDX2^{pos}: 45% vs. 72%, p<0.001, Panel C) in the NCI-CDP validation dataset. The association between CDX2^{neg} tumors and reduced survival remained statistically significant in a multivariate analysis based on the Cox proportional hazards method, thus ruling out possible confounding effects of stage, grade, age or gender for both OS (HR = 1.79, 95%CI = 1.18 - 2.71, p = 0.006, Panel B) and DSS (HR = 2.09, 95%CI = 1.22 - 3.58, p = 0.007, Panel D). In terms of magnitude of effect, the hazard ratio (HR) associated with lack of CDX2 expression was second only to that associtaed with stage, and superior to that associated with pathological grade.

Figure S17. Relationship between CDX2 protein expression and pathological grade (G) in the NCI-CDP validation dataset.

Tumors with lack of CDX2 protein expression are enriched in tumors with high pathological grade

A

Pathological Grade	CDX2 status		% CDX2^{neg}	OR[1] (95% CI[2])
	CDX2^{neg}	CDX2^{pos}		
G1/G2 (n = 316)	23	293	**7.3%** (23/316)	1
G3/G4 (n = 50)	25	25	**50%** (25/50)	12.7 (6.3-25.6)

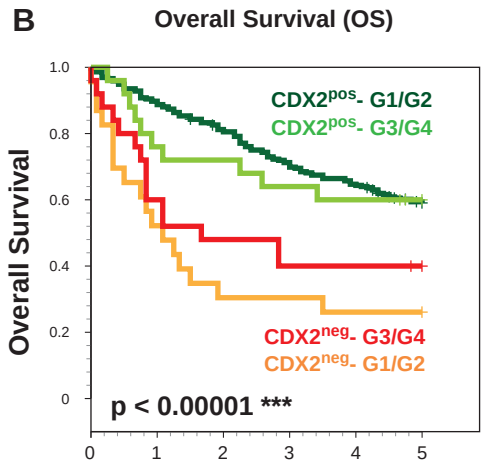
Pearson's Chi-squared Test

$\chi^2 = 69.15$

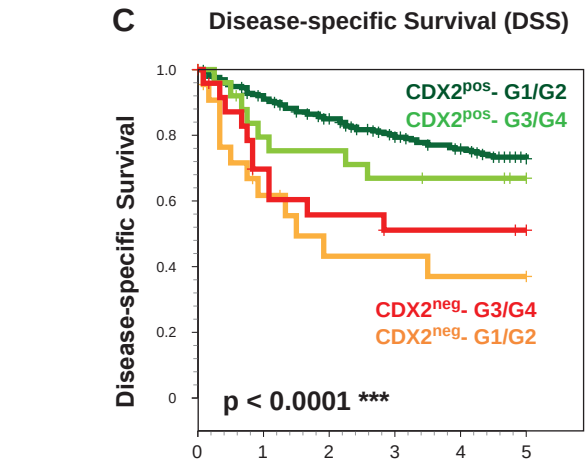
$p < 0.001$ ***

[1]OR: odds ratio; [2] CI: confidence interval

The prognostic effect of CDX2 protein expression is independent of pathological grade



No at risk						
CDX2^{pos}- G1/G2	293	263	236	207	188	160
CDX2^{pos}- G3/G4	25	19	18	16	15	13
CDX2^{neg}- G1/G2	23	12	7	7	6	6
CDX2^{neg}- G3/G4	25	15	12	10	10	9



No at risk						
CDX2^{pos}- G1/G2	293	263	236	207	188	160
CDX2^{pos}- G3/G4	25	19	18	16	15	13
CDX2^{neg}- G1/G2	23	12	7	7	6	6
CDX2^{neg}- G3/G4	25	15	12	10	10	9

CDX2^{pos}- G1/G2: CDX2 positive, low/intermediate grade (G1/G2)
CDX2^{pos}- G3/G4: CDX2 positive, high grade (G3/G4)
CDX2^{neg}- G1/G2 : CDX2 negative, low/intermediate grade (G1/G2)
CDX2^{neg}- G3/G4 : CDX2 negative, high grade (G3/G4)

Figure S17. **Relationship between CDX2 protein expression and pathological grade (G) in the NCI-CDP validation dataset.** An analysis of the distribution of low/intermediate grade (G1/G2) vs. high grade (G3/G4) tumors with respect to CDX2 protein expression in the NCI-CDP validation dataset (n = 366) showed that high grade tumors (G3/G4) were enriched in CDX2^{neg} tumors (Panel A). However, the association between CDX2^{neg} tumors and reduced survival appeared independent of pathological grade. CDX2^{neg} tumors with low/intermediate pathological grade (G1/G2) were characterized by poor clinical outcomes, similar to those observed in CDX2^{neg} tumors with high pathological grade (G3/G4), and substantially worse than those observed in CDX2^{pos} tumors, independently of their low/intermediate or (G1/G2) high (G3/G4) pathological grade. This effect was observed with respect to both 5-year overall survival (OS, Panel B) and 5-year disease-specific survival (DSS, Panel C).

Figure S18. Relationship between CDX2 protein expression, overall survival (OS) and disease-specific survival (DSS) in the NCI-CDP validation dataset, after patient stratification based on stage.

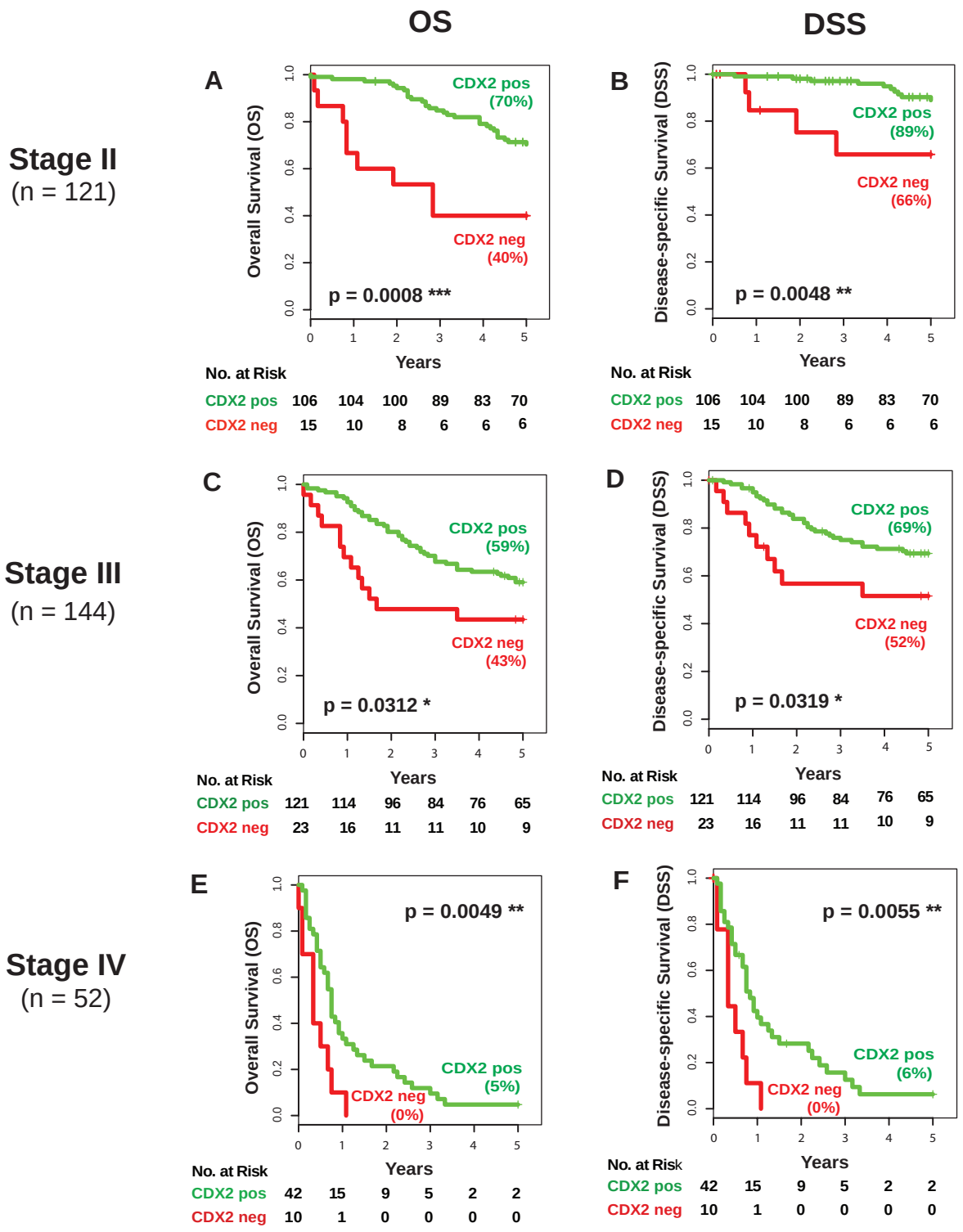


Figure S18. **Relationship between CDX2 protein expression, overall survival (OS) and disease-specific survival (DSS) in the NCI-CDP validation dataset, after patient stratification based on stage.** Lack of CDX2 protein expression was associated with a statistically significant reduction in 5-year OS and DSS across Stage-II, Stage-III and Stage-IV patients from the *National Cancer Institute*'s *Cancer Diagnosis Program* (NCI-CDP) tissue micro-array (TMA) validation dataset. Both the magnitude and the statistical significance of the reductions appeared highest in Stage-II patients (OS, $CDX2^{neg}$ vs. $CDX2^{pos}$: 40% vs. 70%, p = 0.0008, Panel A; DSS, $CDX2^{neg}$ vs. $CDX2^{pos}$: 66% vs. 89%, p = 0.0048, Panel B) as compared to Stage-III (OS: $CDX2^{neg}$ vs. $CDX2^{pos}$, 43% vs. 59%, p = 0.0312, Panel C; DSS: $CDX2^{neg}$ vs. $CDX2^{pos}$, 52% vs. 69%, p = 0.0319, Panel D) and Stage-IV patients (OS, $CDX2^{neg}$ vs. $CDX2^{pos}$: 0% vs. 5%, p = 0.0049, Panel E; DSS, $CDX2^{neg}$ vs. $CDX2^{pos}$: 0% vs. 6%, p = 0.0055, Panel F).

Figure S19. Relationship between CDX2 expression and 5-year DFS in **Stage-II** patients from the NCI-CDP validation dataset, after **stratification** for either size and depth of invasion of the primary tumor **(T3 vs. T4)** or number of regional lymph-nodes resected at surgery **(≥12 vs. <12)**.

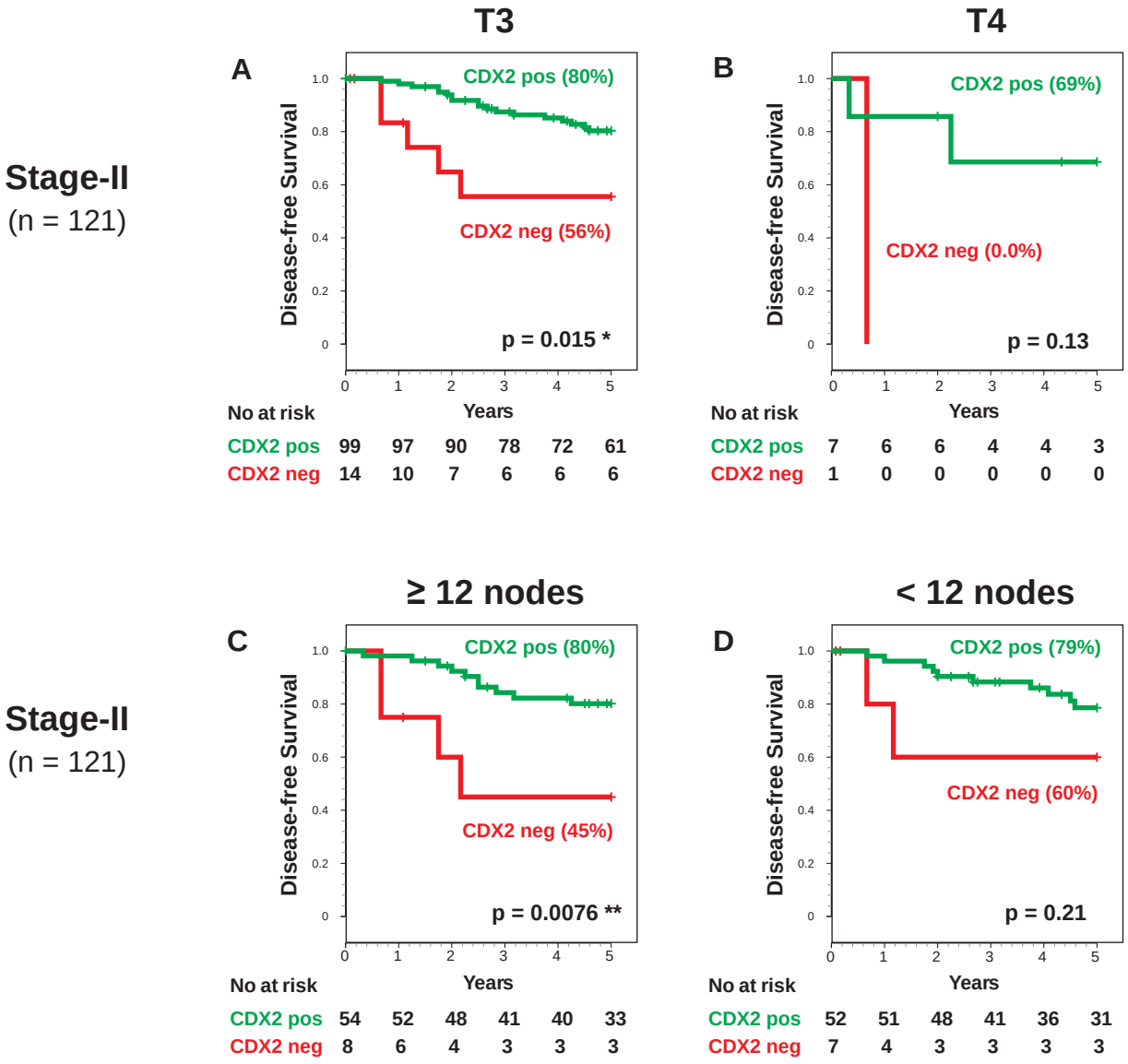


Figure S19. **Relationship between CDX2 protein expression and 5-year DFS in Stage-II patients from the NCI-CDP validation dataset, after stratification for either T-stage of the primary tumor (T3 vs. T4) or number of lymph-nodes resected at surgery (≥12 vs. <12).** In the Stage-II patient cohort from the *National Cancer Institute*'s *Cancer Diagnosis Program* (NCI-CDP) tissue micro-array (TMA) validation dataset (n = 121), lack of CDX2 protein expression was associated with reduced 5-year DFS independently of the T-stage of the primary tumor (T3, Panel A; T4, Panel B) and the number of regional lymph-nodes resected at surgery (≥12 nodes, Panel C; <12 nodes, Panel D). The difference in 5-year DFS observed between $CDX2^{neg}$ and $CDX2^{pos}$ tumors remained statistically significant in the Stage-II/T3 ($CDX2^{neg}$ vs. $CDX2^{pos}$: 56% vs. 80%, p = 0.015, Panel A) and Stage-II/≥12 nodes ($CDX2^{neg}$ vs. $CDX2^{pos}$: 45% vs. 80%, p = 0.0076, Panel C) subgroups.

Figure S20. Relationship between CDX2 expression and benefit from adjuvant chemotherapy in Stage-III patients from discovery (NCBI-GEO) and validation (NCI-CDP) datasets.

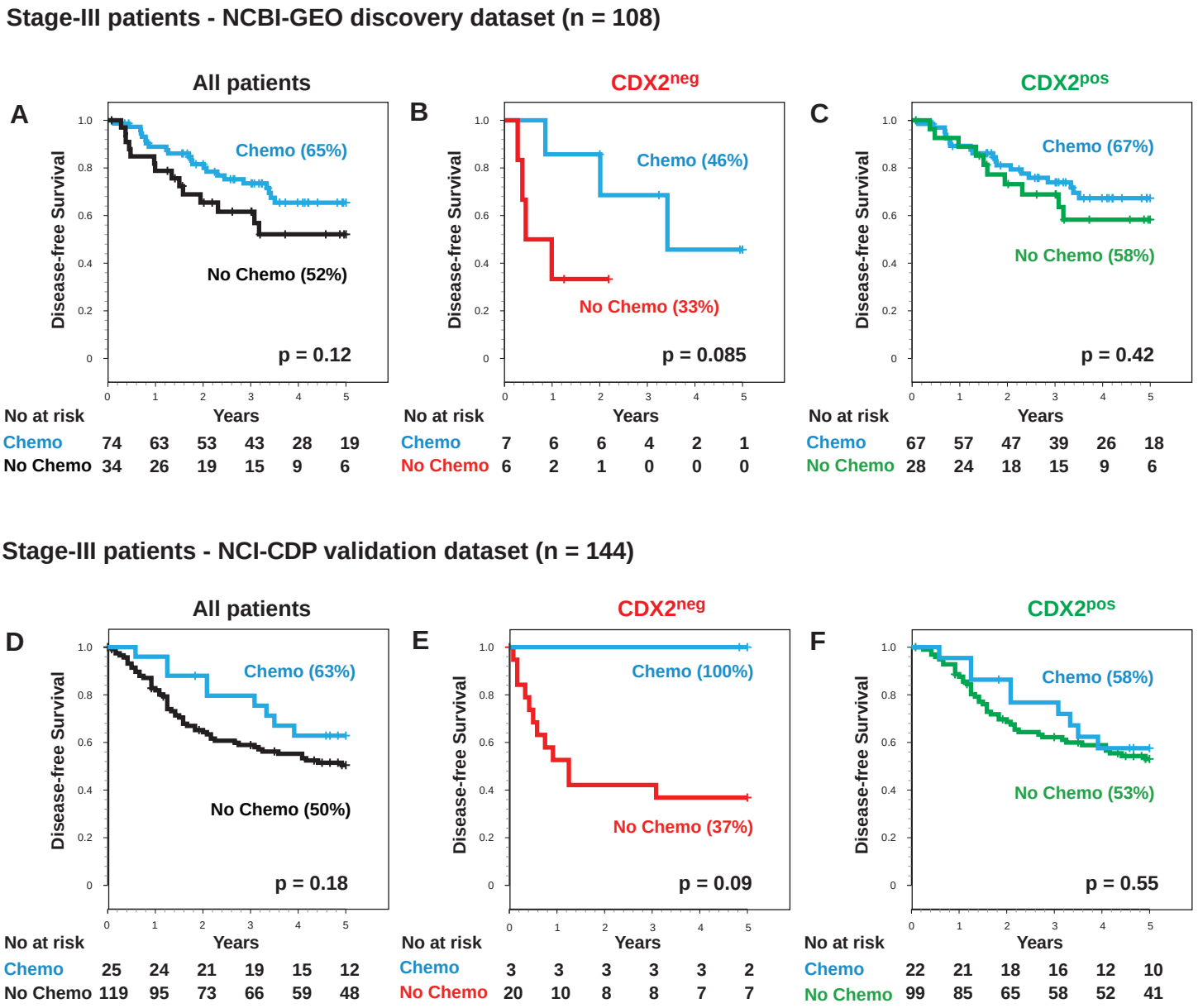


Figure S20. **Relationship between CDX2 expression and benefit from adjuvant chemotherapy in Stage-III patients from discovery (NCBI-GEO) and validation (NCI-CDP) datasets.** To evaluate whether patients with CDX2^{neg} tumors had benefited from adjuvant chemotherapy, we investigated the relationship between CDX2 status, 5-year DFS and treatment with adjuvant chemotherapy in Stage-III patients from both the NCBI-GEO discovery dataset (n = 108) and the NCI-CDP validation dataset (n=144). We stratified Stage-III patients according to CDX2 status (CDX2^{neg} vs. CDX2^{pos}) and compared the DFS of those treated with adjuvant chemotherapy with the DFS of those not treated with adjuvant chemotherapy (Chemo vs. No Chemo). Overall, treatment with adjuvant chemotherapy was associated with a trend towards improved 5-year DFS across all tested cohorts, although the difference did not reach statistical significance. The improvement in DFS associated with adjuvant chemotherapy appeared of greater magnitude in CDX2^{neg} patients as compared to CDX2^{pos} ones.

Figure S21. Relationship between CDX2 expression and benefit from adjuvant chemotherapy in **Stage-II** patients, after **stratification** for the size and depth of invasion of the primary tumor **(T3 vs. T4)**.

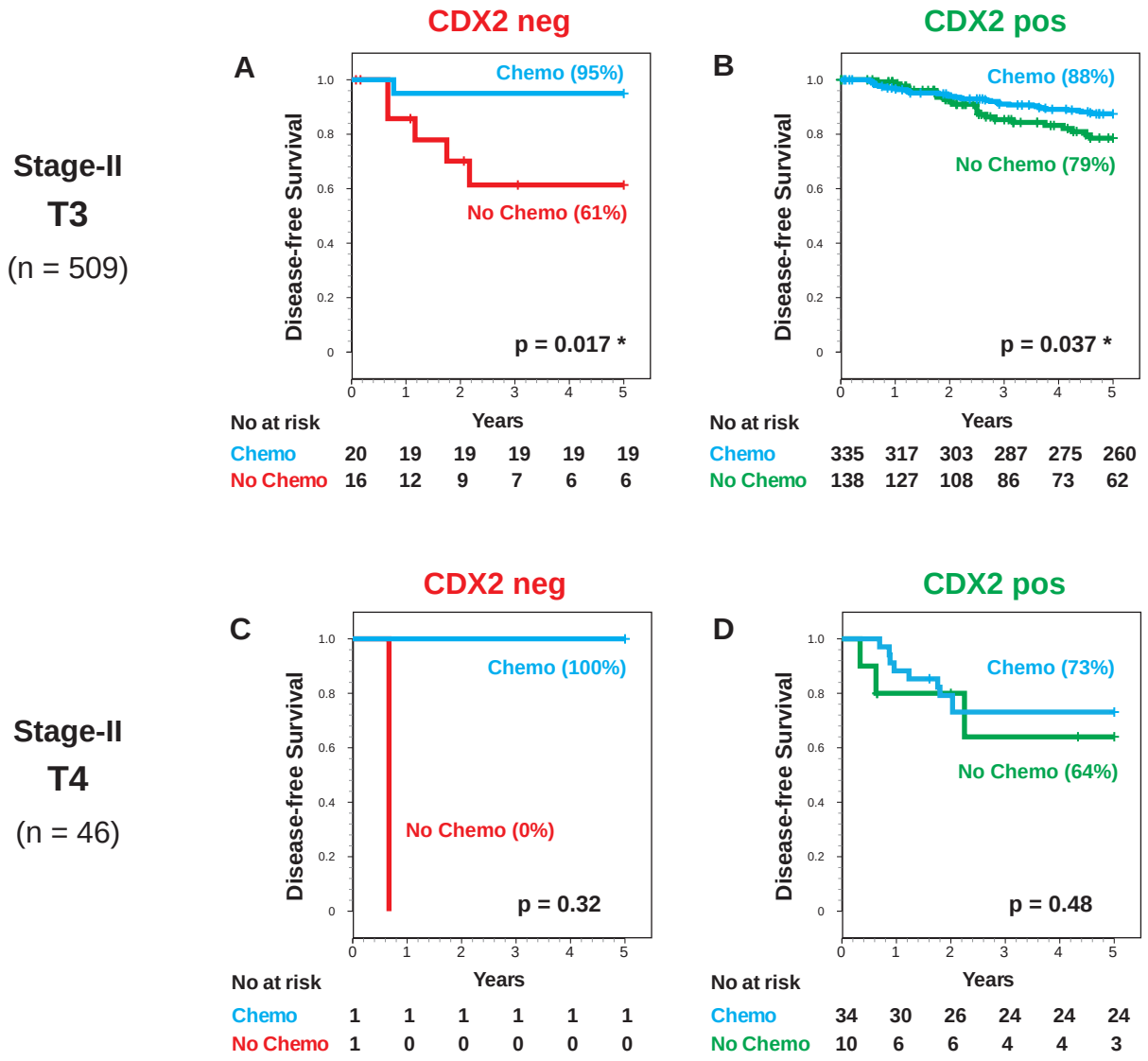


Figure S21. **Relationship between CDX2 protein expression and benefit from adjuvant chemotherapy in Stage-II patients, after stratification for the size and depth of invasion of the primary tumor (T3 vs. T4).** To evaluate whether the association between treatment with adjuvant chemotherapy and improved DFS in patients with Stage-II/CDX2^{neg} tumors could have been influenced by the size and depth of invasion (T-stage) of the primary tumor, we compared the 5-year DFS of treated and untreated patients from a pool of the three datasets annotated with T-stage information (NCI-CDP, NSABP-C07, Stanford TMAD), after stratification for CDX2 status (CDX2^{neg} vs. CDX2^{pos}) and T-stage itself (T3 vs. T4). Overall, treatment with adjuvant chemotherapy was associated with a trend towards improved 5-year DFS in both CDX2^{neg} and CDX2^{pos} cohorts, independently of the T-stage of the primary tumor being classified as T3 (CDX2^{neg}, Panel A; CDX2^{pos}, Panel B) or T4 (CDX2^{neg}, Panel C; CDX2^{pos}, Panel D). The difference in 5-year DFS observed between treated and untreated patients remained statistically significant in the Stage-II/T3 cohorts, in both CDX2^{neg} (No Chemo vs. Chemo: 61% vs. 95%, p = 0.017, Panel A) and CDX2^{pos} (No Chemo vs. Chemo: 79% vs. 88%, p = 0.037, Panel B) subgroups.

Figure S22. Relationship between CDX2 expression and benefit from adjuvant chemotherapy in **Stage-II** patients, after **stratification** for the number of regional lymph-nodes resected at surgery **(≥12 vs. <12)**.

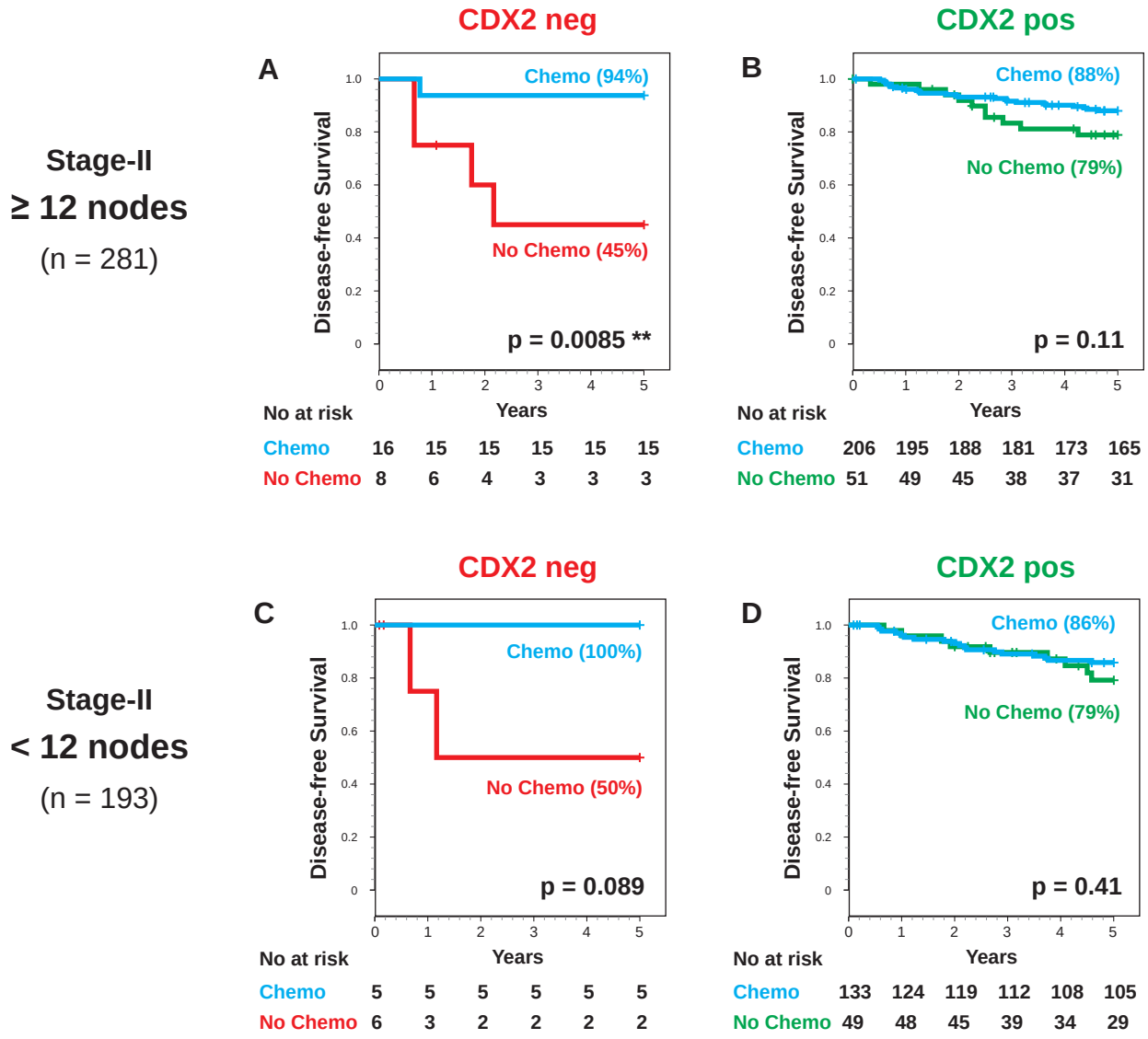


Figure S22. **Relationship between CDX2 protein expression and benefit from adjuvant chemotherapy in Stage-II patients, after stratification for the number of regional lymph-nodes resected at surgery (≥12 vs. <12).** To evaluate whether the association between treatment with adjuvant chemotherapy and improved DFS in patients with Stage-II/CDX2^{neg} tumors could have been influenced by parameters related to the quality of the regional lymph-node sampling during the surgical resection of the primary tumor, we compared the 5-year DFS of treated and untreated patients from a pool of the two datasets that were annotated with such information (NCI-CDP, NSABP-C07), after stratification for CDX2 status (CDX2^{neg} vs. CDX2^{pos}) and the number of regional lymph-nodes resected at surgery (≥12 vs. <12). Overall, treatment with adjuvant chemotherapy was associated with a trend towards improved 5-year DFS in both CDX2^{neg} and CDX2^{pos} cohorts, independently of the number of resected regional lymph-nodes being ≥12 (CDX2^{neg}, Panel A; CDX2^{pos}, Panel B) or <12 (CDX2^{neg}, Panel C; CDX2^{pos}, Panel D). The difference in 5-year DFS observed between treated and untreated patients remained statistically significant in the cohort of Stage-II/CDX2^{neg} patients who benefited from extensive sampling of regional lymph-nodes (≥12) during the surgical resection of the primary tumor (No Chemo vs. Chemo: 45% vs. 94%, p = 0.0085, Panel A).

Figure S23. Relationship between CDX2 expression and benefit from adjuvant chemotherapy in **Stage-III** patients, after **stratification** for the size and depth of invasion of the primary tumor **(T3 vs. T4)**.

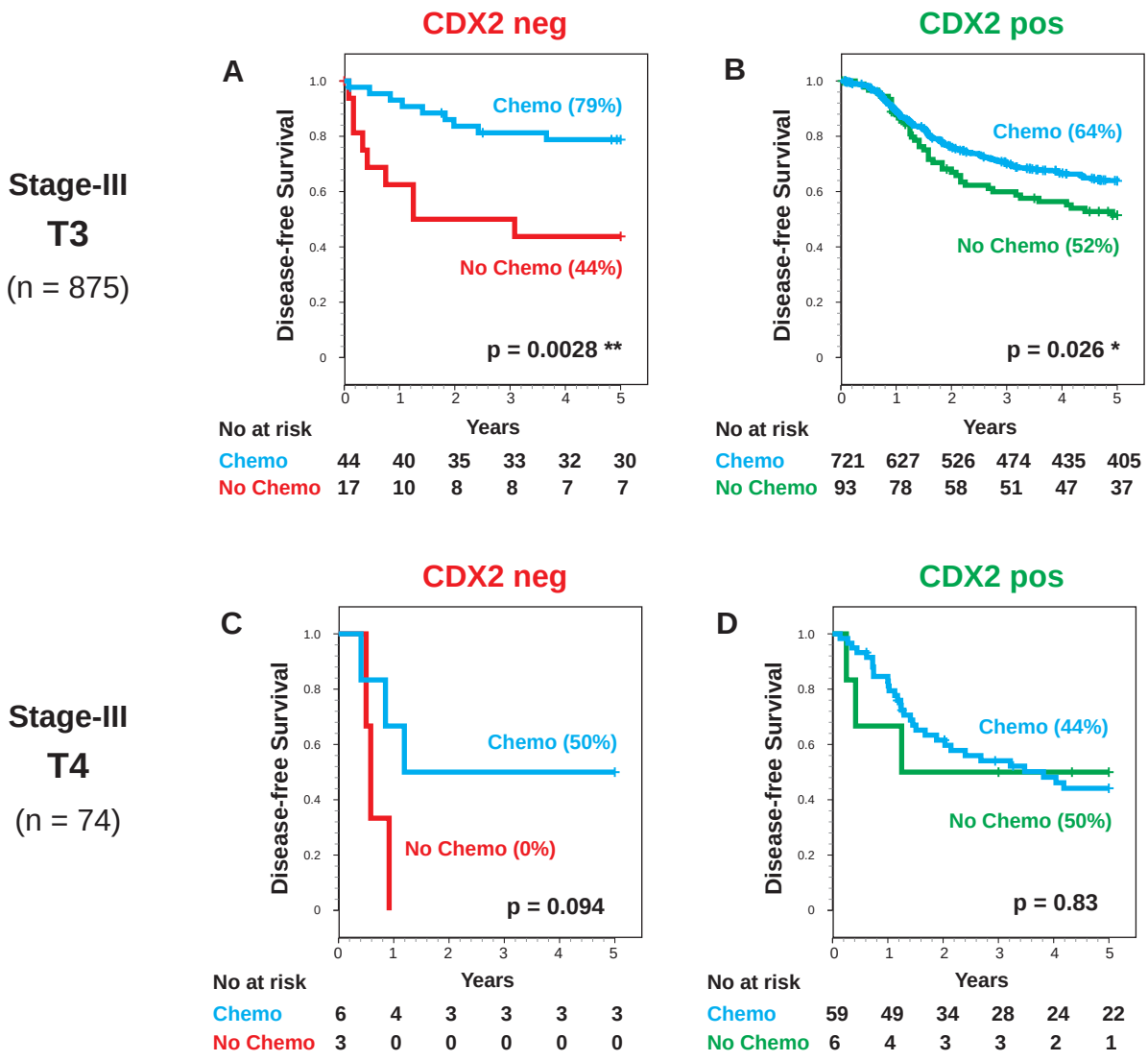


Figure S23. **Relationship between CDX2 protein expression and benefit from adjuvant chemotherapy in Stage-III patients, after stratification for the size and depth of invasion of the primary tumor (T3 vs. T4).** To evaluate whether the association between adjuvant chemotherapy and improved DFS in patients with Stage-III/CDX2^{neg} tumors could have been influenced by the size and depth of invasion (T-stage) of the primary tumor, we compared the 5-year DFS of treated and untreated patients from a pool of the three datasets annotated with T-stage information (NCI-CDP, NSABP-C07, Stanford TMAD), after stratification for CDX2 status (CDX2^{neg} vs. CDX2^{pos}) and T-stage itself (T3 vs. T4). Overall, treatment with adjuvant chemotherapy was associated with a trend towards improved 5-year DFS in both CDX2^{neg} and CDX2^{pos} cohorts, independently of the T-stage of the primary tumor being classified as T3 (CDX2^{neg}, Panel A; CDX2^{pos}, Panel B) or T4 (CDX2^{neg}, Panel C; CDX2^{pos}, Panel D). The difference in 5-year DFS observed between treated and untreated patients remained statistically significant in the Stage-III/T3 cohorts, in both CDX2^{neg} (No Chemo vs. Chemo: 44% vs. 79%, p = 0.0028, Panel A) and CDX2^{pos} (No Chemo vs. Chemo: 52% vs. 64%, p = 0.026, Panel B) subgroups.

Figure S24. Relationship between CDX2 expression and benefit from adjuvant chemotherapy in **Stage-III** patients, after **stratification** for the extent of metastatic spread to regional lymph-nodes **(N1 vs. N2).**

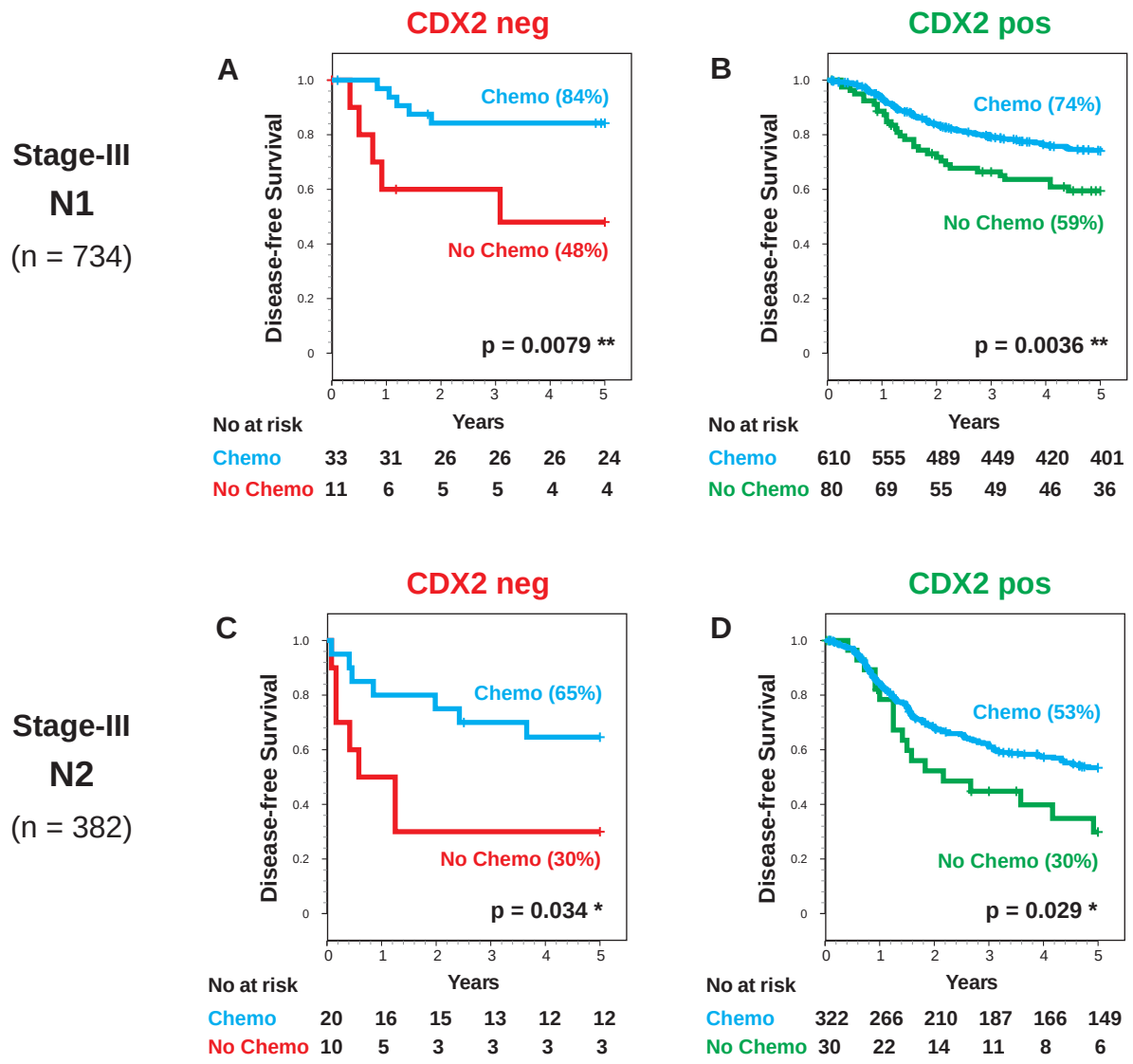


Figure S24. **Relationship between CDX2 protein expression and benefit from adjuvant chemotherapy in Stage-III patients, after stratification for the extent of metastatic spread to regional lymph-nodes (N1 vs. N2).** To evaluate whether the association between adjuvant chemotherapy and improved DFS in patients with Stage-III/CDX2^{neg} tumors could have been influenced by the extent of the tumor's metastatic spread to the regional lymph-nodes (N-stage), we compared the 5-year DFS of treated and untreated patients from a pool of the three datasets annotated with N-stage information (NCI-CDP, NSABP-C07, Stanford TMAD), after stratification for CDX2 status (CDX2^{neg} vs. CDX2^{pos}) and N-stage itself (N1 vs. N2). Overall, treatment with adjuvant chemotherapy was associated with improved 5-year DFS in both CDX2^{neg} and CDX2^{pos} cohorts, independently of the N-stage of the primary tumor being classified as N1 (CDX2^{neg}, Panel A; CDX2^{pos}, Panel B) or N2 (CDX2^{neg}, Panel C; CDX2^{pos}, Panel D). The difference in 5-year DFS observed between treated and untreated patients remained statistically significant across all Stage-III cohorts, including N1/CDX2^{neg} (No Chemo vs. Chemo: 48% vs. 84%, p = 0.0079, Panel A), N1/CDX2^{pos} (No Chemo vs. Chemo: 59% vs. 74%, p = 0.036, Panel B), N2/CDX2^{neg} (No Chemo vs. Chemo: 30% vs. 65%, p = 0.034, Panel C) and N2/CDX2^{pos} (No Chemo vs. Chemo: 30% vs. 53%, p = 0.029, Panel D) subgroups.

Table S1. List of publicly available NCBI-GEO[1] datasets used in the study.

NCBI - GEO dataset [1]	number of samples			Affymetrix® Platform	PubMed ID	Reference
	normal	cancer	total			
Human Colon Global Database						
GSE2109 (only colorectal cancer patients)		427	427	HG U133 Plus 2.0	n.a.	Expression Project for Oncology (expO) [2]
GSE2361 (only one normal colon sample)	1		1	HG U133A	PMID 15950434	Ge *et al.*, *Genomics*, 86:127-14 (2005)
GSE4045		37	37	HG U133A	PMID 16819509	Laiho *et al.*, *Oncogene*, 26:312-320 (2007)
GSE4107	10	12	22	HG U133 Plus 2.0	PMID 17317818	Hong *et al.*, *Clin. Cancer Res.*, 13:1107-1114 (2007)
GSE4183 (only normal colon and colorectal cancer)	8	15	23	HG U133 Plus 2.0	PMID 19461970	Gyorffy *et al.*, *PLoS ONE*, 4:e5645 (2009)
GSE5851		80	80	HG U133A 2.0	PMID 17664471	Khambata-Ford *et al.*, *J. Clin. Oncol.*, 25:3230-3237 (2007)
GSE8671 (only normal colon and colorectal cancer)	32		32	HG U133 Plus 2.0	PMID 18171984	Sabates-Bellver *et al.*, *Mol. Cancer Res.*, 5:1263-1275 (2007)
GSE9254	19		19	HG U133 Plus 2.0	PMID 18056783	La Pointe *et al.*, *Physiol. Genomics*, 33:50-64 (2008)
GSE9348	12	70	82	HG U133 Plus 2.0	PMID 20143136	Hong *et al.*, *Clin. Exp. Metastasis*, 27:83-90 (2010)
GSE10714 (only normal colon and colorectal cancer)	3	7	10	HG U133 Plus 2.0	PMID 20087348	Galamb *et al.*, *Br. J. Cancer*, 102:765-773 (2010)
GSE10961		18	18	HG U133 Plus 2.0	PMID 18827815	Pantaleo *et al.*, *Br. J. Cancer*, 99:1729-1734 (2008)
GSE11831	17		17	HG U133 Plus 2.0	PMID 19603079	Nielsen *et al.*, *PLoS ONE*, 4:e6210 (2009)
GSE12945		62	62	HG U133A	PMID 19399471	Staub *et al.*, *J. Mol. Med.*, 87:633-644 (2009)
GSE13067		74	74	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin. Cancer Res.*, 14:8061-8069 (2008)
GSE13294		155	155	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin. Cancer Res.*, 14:8061-8069 (2008)
GSE13471 (only colon samples)	4	4	8	HG U133A	PMID 19151715	Irizarry *et al.*, *Nat. Genet.*, 41:178-186 (2009)
GSE14333 (samples non-redundant with GSE13067)		226	226	HG U133 Plus 2.0	PMID 19996206	Jorissen *et al.*, *Clin. Cancer Res.*, 15:7642-7651 (2009)
GSE15960 (only normal colon and colorectal cancer)	6	6	12	HG U133 Plus 2.0	PMID 20087348	Galamb *et al.*, *Br. J. Cancer*, 102:765-773 (2010)
GSE17538 (samples non-redundant with GSE14333)		65 [3]	65 [3]	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968 (2010)
GSE18105	17	94	111	HG U133 Plus 2.0	PMID 20162577	Matsuyama *et al.*, *Int. J. Cancer*, 127:2292-2299 (2010)
GSE20916 (only normal colon and colorectal cancer)	44	91	135	HG U133 Plus 2.0	PMID 20957034	Skrzypczak *et al., PLoS ONE*, 5:e1309 (2010)
GSE26682 (samples analyzed using the HG U133A platform)		155	155	HG U133A	PMID 21300766	Vilar *et al.*, *Cancer Res.*, 71:2632-2642 (2011)
GSE26682 (samples analyzed using the HG U133 Plus 2.0 platform)		176	176	HG U133 Plus 2.0	PMID 21300766	Vilar *et al.*, *Cancer Res.*, 71:2632-2642 (2011)
GSE26906 (samples non-redundant with GSE37892)		58	58	HG U133 Plus 2.0	PMID 22496922	Birnbaum *et al.*, *Transl. Oncol.*, 5:72-76 (2012)
GSE29623 (samples non-redundant with GSE14333)		1	1	HG U133 Plus 2.0	PMID 22362069	Chen *et al.*, *J. Gastrointest. Surg.*, 16:905-912 (2012)
GSE31595		37	37	HG U133 Plus 2.0	PMID 22710688	Thorsteinsson *et al.*, *Int. J. Colorectal. Dis.*, 27:1579-1586 (2012)
GSE37892		130	130	HG U133 Plus 2.0	PMID 22917480	Laibe *et al.*, *OMICS*, 16:560-565 (2012)
GSE41258 (only normal colon and primary tumors)	54	186	240	HG U133A	PMID 19359472	Sheffer *et al.*, *P.N.A.S.*, 106:7131-7136 (2009)
Total number of samples	227	2239	2466			
Total number of samples after "purging" [4]	214	2115	2329			
Colon Cancer - disease-free survival (DFS) database						
GSE17538 (DFS data, VMC + MCC) [5]		200	200	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968 (2010)
GSE14333 (DFS data, Melbourne + MCC) [6]		99	99	HG U133 Plus 2.0	PMID 19996206	Jorissen *et al.*, *Clin. Cancer Res.*, 15:7642-7651 (2009)
GSE31595		37	37	HG U133 Plus 2.0	PMID 22710688	Thorsteinsson *et al.*, *Int. J. Colorectal. Dis.*, 27:1579-1586 (2012)
GSE37892		130	130	HG U133 Plus 2.0	PMID 22917480	Laibe *et al.*, *OMICS*, 16:560-565 (2012)
Total number of samples[7]			466			
Colon Cancer - DFS + pathological grading database						
GSE17538 (only patients with both DFS and grading data)		181	181	HG U133 Plus 2.0	PMID 19914252	Smith *et al.*, *Gastroenterology*, 138:958-968 (2010)
GSE31595 (only patients with both DFS and grading data)		35	35	HG U133 Plus 2.0	PMID 22710688	Thorsteinsson *et al.*, *Int. J. Colorectal. Dis.*, 27:1579-1586 (2012)
Total number of samples[7]			216			
Colon Cancer - Stage II + adjuvant chemotherapy database						
GSE14333 (only Stage II patients with DFS and adj. chemo. data)		94	94	HG U133 Plus 2.0	PMID 19996206	Jorissen *et al.*, *Clin. Cancer Res.*, 15:7642-7651 (2009)
GSE31595 (only Stage II patients with DFS and adj. chemo. data)		20	20	HG U133 Plus 2.0	PMID 22710688	Thorsteinsson *et al.*, *Int. J. Colorectal. Dis.*, 27:1579-1586 (2012)
Total number of samples[7]			114			
Colon Cancer - Stage III + adjuvant chemotherapy database						
GSE14333 (only Stage III patients with DFS and adj. chemo. data)		91	91	HG U133 Plus 2.0	PMID 19996206	Jorissen *et al.*, *Clin. Cancer Res.*, 15:7642-7651 (2009)
GSE31595 (only Stage III patients with DFS and adj. chemo. data)		17	17	HG U133 Plus 2.0	PMID 22710688	Thorsteinsson *et al.*, *Int. J. Colorectal. Dis.*, 27:1579-1586 (2012)
Total number of samples[7]			108			
Colon Cancer - MSI/MSS database						
GSE13067		74	74	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin. Cancer Res.*, 14:8061-8069 (2008)
GSE13294		155	155	HG U133 Plus 2.0	PMID 19088021	Jorissen *et al.*, *Clin. Cancer Res.*, 14:8061-8069 (2008)
GSE24514 (only tumor samples, all MSI)		34	34	HG U133A	PMID 21544814	Alhopuro *et al.*, *Int J. Cancer*, 130:1558-1566 (2012)
GSE26682 (only tumors annotated for MSI/MSS status)		140	140	HG U133A	PMID 21300766	Vilar *et al.*, *Cancer Res.*, 71:2632-2642 (2011)
GSE26682 (only tumors annotated for MSI/MSS status)		160	160	HG U133 Plus 2.0	PMID 21300766	Vilar *et al.*, *Cancer Res.*, 71:2632-2642 (2011)
GSE35896 (only tumors annotated for MSI/MSS status)		61	61	HG U133 Plus 2.0	PMID 23272949	Schlicker *et al.*, *BMC Med. Genomics*, 5:66 (2012)
GSE39084		70	70	HG U133 Plus 2.0	PMID 25083765	Kirzin *et al.*, *PLoS ONE*, 9:e103159 (2014)
GSE41258 (only primary tumors annotated for MSI/MSS status)		168	168	HG U133A	PMID 19359472	Sheffer *et al.*, *P.N.A.S.*, 106:7131-7136 (2009)
Total number of samples[7]			862			
Colon Cancer - TP53 mutation database						
GSE39084		70	70	HG U133 Plus 2.0	PMID 25083765	Kirzin *et al.*, *PLoS ONE*, 9:e103159 (2014)
GSE41258 (only primary tumors annotated for TP53 mutations)		144	144	HG U133A	PMID 19359472	Sheffer *et al.*, *P.N.A.S.*, 106:7131-7136 (2009)
Total number of samples[7]			214			

[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 (please refer to Figure S1)

[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] Not purged

SUPPLEMENTARY REFERENCES

1. Sahoo D, Dill DL, Tibshirani R, Plevritis SK. Extracting binary signals from microarray time-course data. **Nucleic Acids Research,** 35:3705-3712 (2007).

2. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK. Boolean implication networks derived from large scale, whole genome microarray datasets. **Genome Biology,** 9:R157 (2008).

3. Levin TG, Powell AE, Davies PS, et al. Characterization of the intestinal cancer stem cell marker CD166 in the human and mouse gastrointestinal tract. **Gastroenterology,** 139:2072-2082 (2010).

4. Weichert W, Knosel T, Bellach J, Dietel M, Kristiansen G. ALCAM/CD166 is overexpressed in colorectal carcinoma and correlates with shortened patient survival. **Journal of Clinical Pathology,** 57:1160-1164 (2004).

5. Dalerba P, Dylla SJ, Park IK, et al. Phenotypic characterization of human colorectal cancer stem cells. **Proceedings of the National Academy of Sciences USA (PNAS),** 104:10158-10163 (2007).

6. Jorissen RN, Gibbs P, Christie M, et al. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. **Clinical Cancer Research,** 15:7642-7651 (2009).

7. Smith JJ, Deane NG, Wu F, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. **Gastroenterology,** 138:958-968 (2010).

8. Thorsteinsson M, Kirkeby LT, Hansen R, et al. Gene expression profiles in stages II and III colon cancers: application of a 128-gene signature. **International Journal of Colorectal Disease,** 27:1579-1586 (2012).

9. Laibe S, Lagarde A, Ferrari A, et al. A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. **OMICS: a Journal of Integrative Biology,** 16:560-565 (2012).

10. Dalerba P, Kalisky T, Sahoo D, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. **Nature Biotechnology,** 29:1120-1127 (2011).

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

12. Li MK, Folpe AL. CDX-2, a new marker for adenocarcinoma of gastrointestinal origin. **Advances in Anatomic Pathology,** 11:101-105 (2004).

13. Werling RW, Yaziji H, Bacchi CE, Gown AM. CDX2, a highly sensitive and specific marker of adenocarcinomas of intestinal origin: an immunohistochemical survey of 476 primary and metastatic carcinomas. **The American Journal of Surgical Pathology,** 27:303-310 (2003).

14. Borrisholt M, Nielsen S, Vyberg M. Demonstration of CDX2 is highly antibody dependant. **Applied Immunohistochemistry and Molecular Morphology,** 21:64-72 (2013).

15. Kuebler JP, Wieand HS, O'Connell MJ, et al. Oxaliplatin combined with weekly bolus fluorouracil and leucovorin as surgical adjuvant chemotherapy for stage II and III colon cancer: results from NSABP C-07. **Journal of Clinical Oncology,** 25:2198-2204 (2007).

16. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. **Physical Therapy,** 85:257-268 (2005).

17. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 2 x k factorial design when time-to-failure is the outcome. **Controlled Clinical Trials,** 14:511-522 (1993).

18. Williamson DF, Parker RA, Kendrick JS. The box plot: a simple visual method to interpret data. **Annals of Internal Medicine,** 110:916-921 (1989).