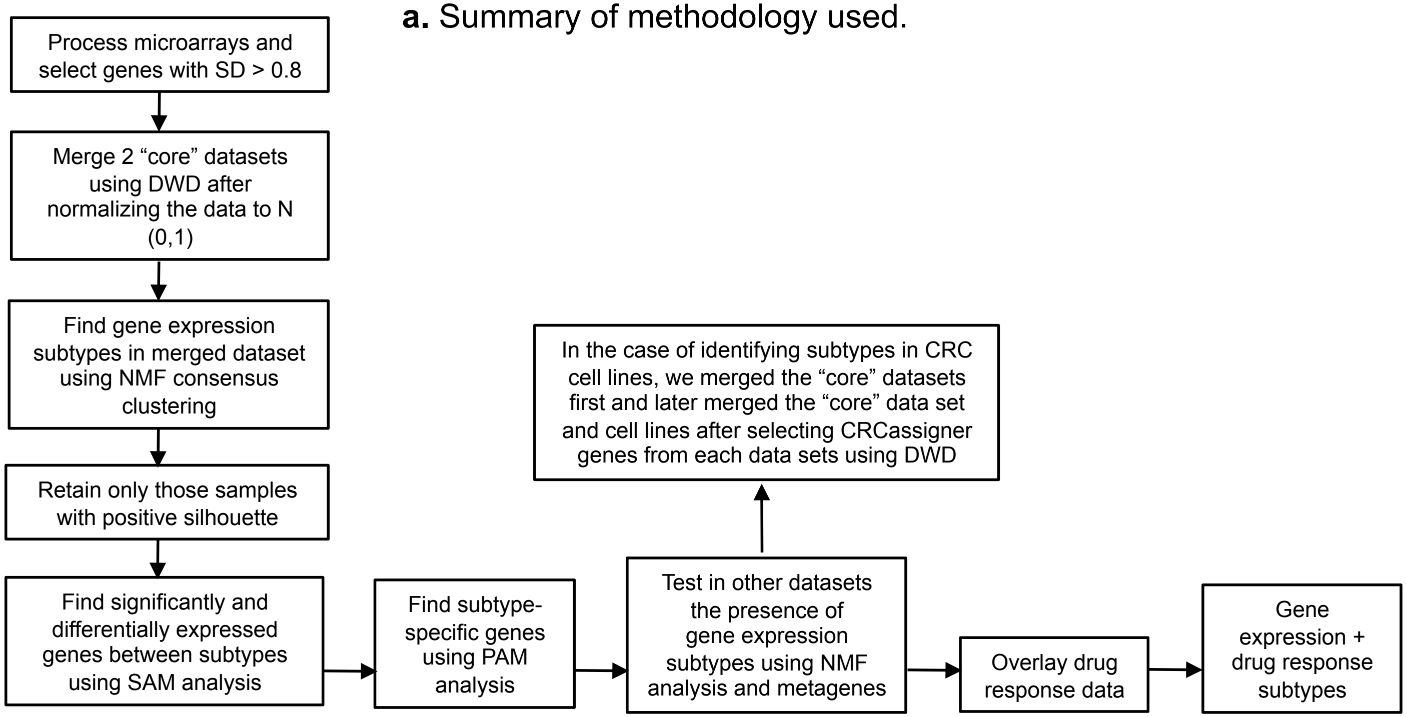
Supplementary Fig. 1

a. Summary of methodology used.



Process microarrays and select genes with SD > 0.8

Merge 2 “core” datasets using DWD after normalizing the data to N (0,1)

Find gene expression subtypes in merged dataset using NMF consensus clustering

Retain only those samples with positive silhouette

Find significantly and differentially expressed genes between subtypes using SAM analysis

Find subtype-specific genes using PAM analysis

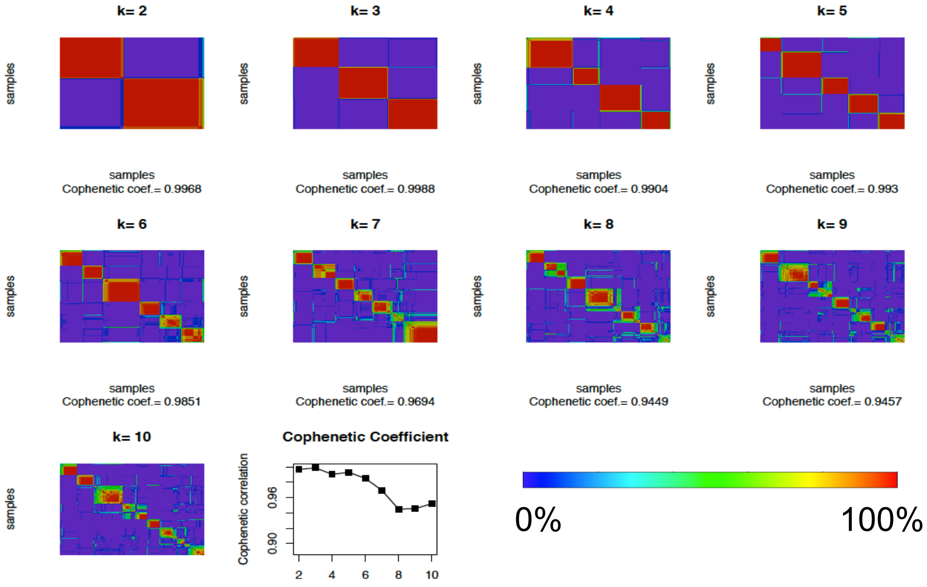
Test in other datasets the presence of gene expression subtypes using NMF analysis and metagenes

In the case of identifying subtypes in CRC cell lines, we merged the “core” datasets first and later merged the “core” data set and cell lines after selecting CRCassigner genes from each data sets using DWD

Overlay drug response data

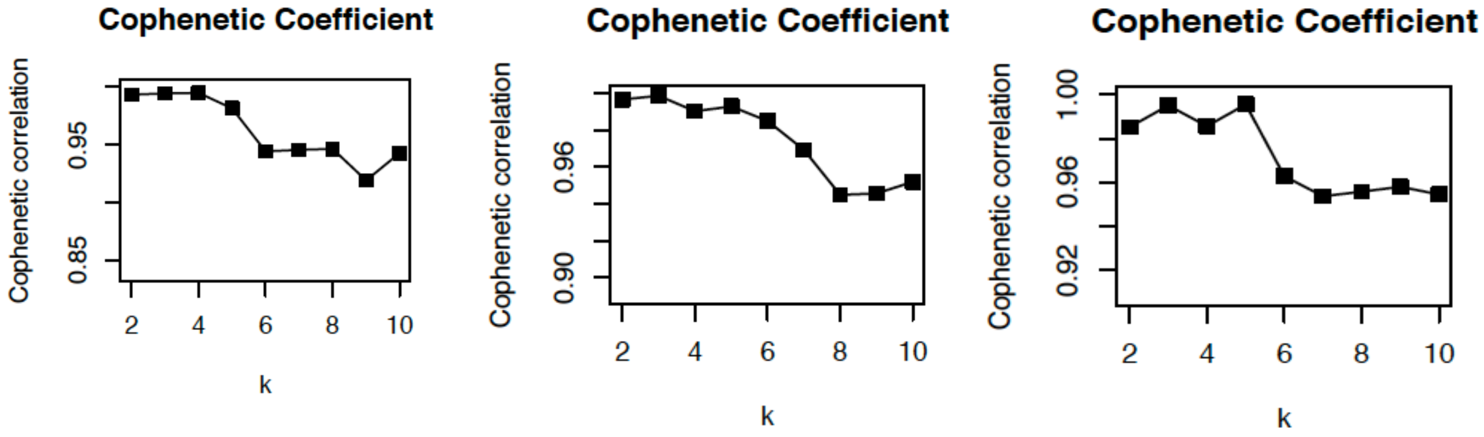
Gene expression + drug response subtypes

b. NMF consensus clusters for combined core datasets - GSE13294 & GSE14333



Supplementary Fig. 1, cont'd

c. Cophenetic coefficient plots from NMF clusters for genes selected using different SD cut-offs.

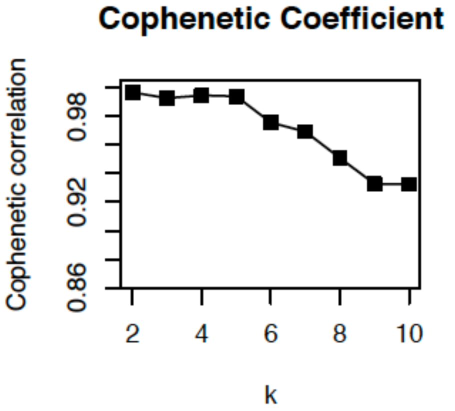


SD = 0.5

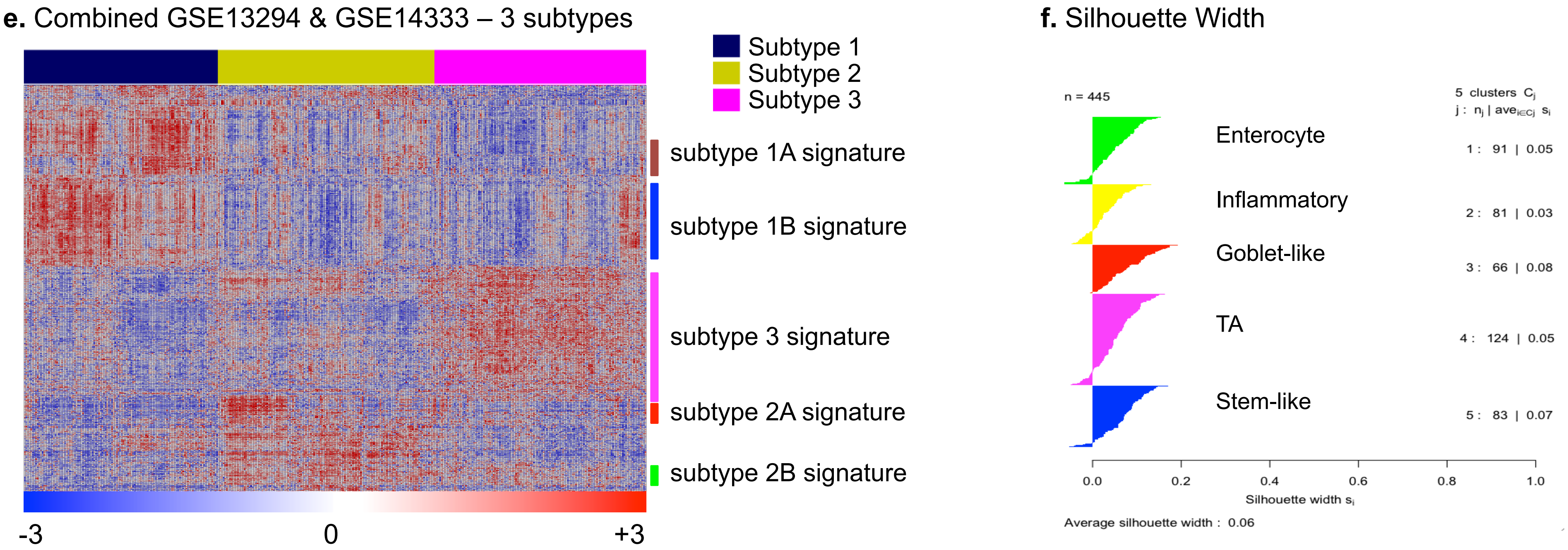
SD = 0.8

SD = 1

d. Cophenetic coefficient plots from NMF clusters for genes selected using fold change greater than 2 in at least 3 samples.



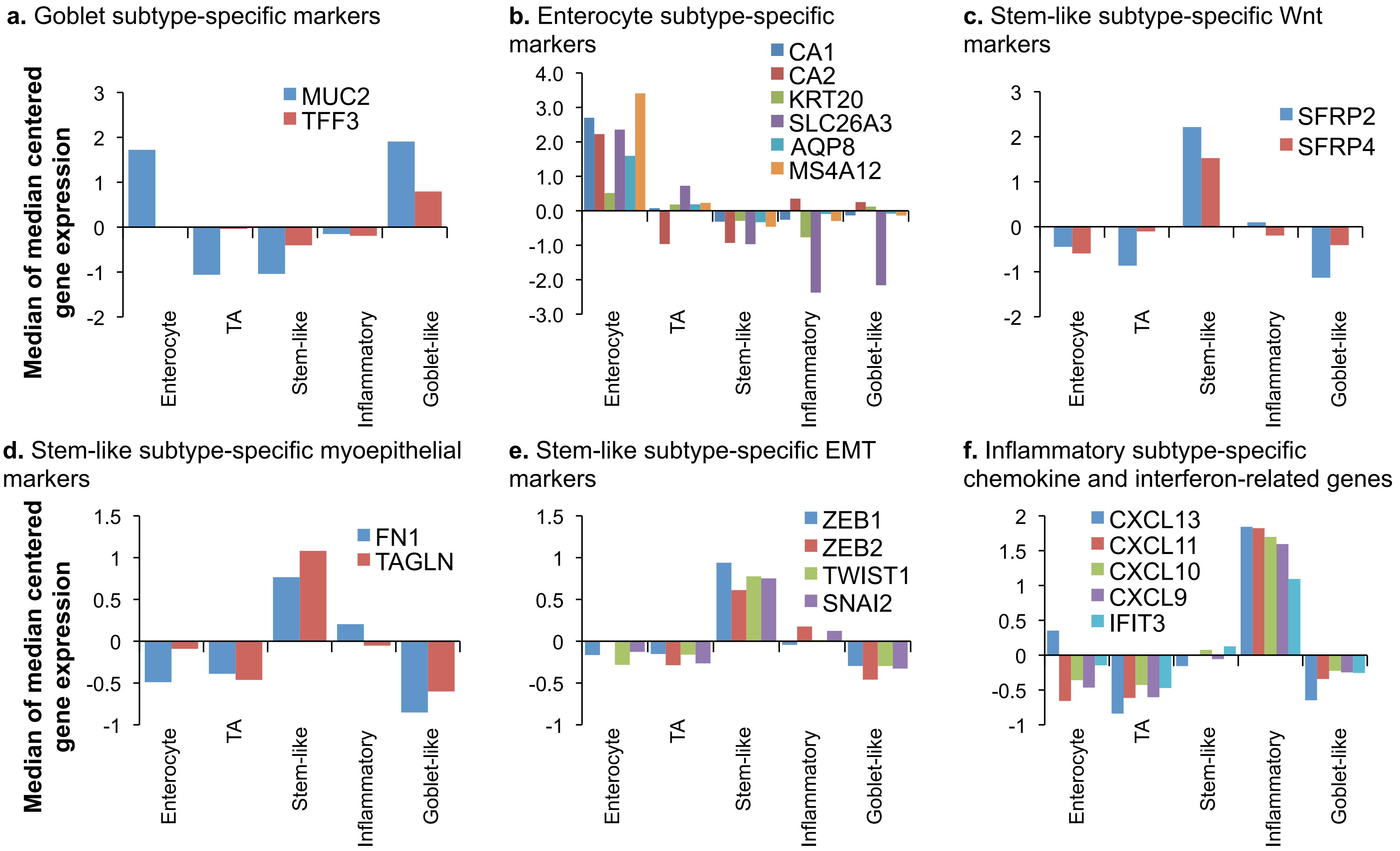
Supplementary Fig. 1, cont'd



e. Combined GSE13294 & GSE14333 – 3 subtypes

f. Silhouette Width

Supplementary Fig. 2



a. Goblet subtype-specific markers

b. Enterocyte subtype-specific markers

c. Stem-like subtype-specific Wnt markers

d. Stem-like subtype-specific myoepithelial markers

e. Stem-like subtype-specific EMT markers

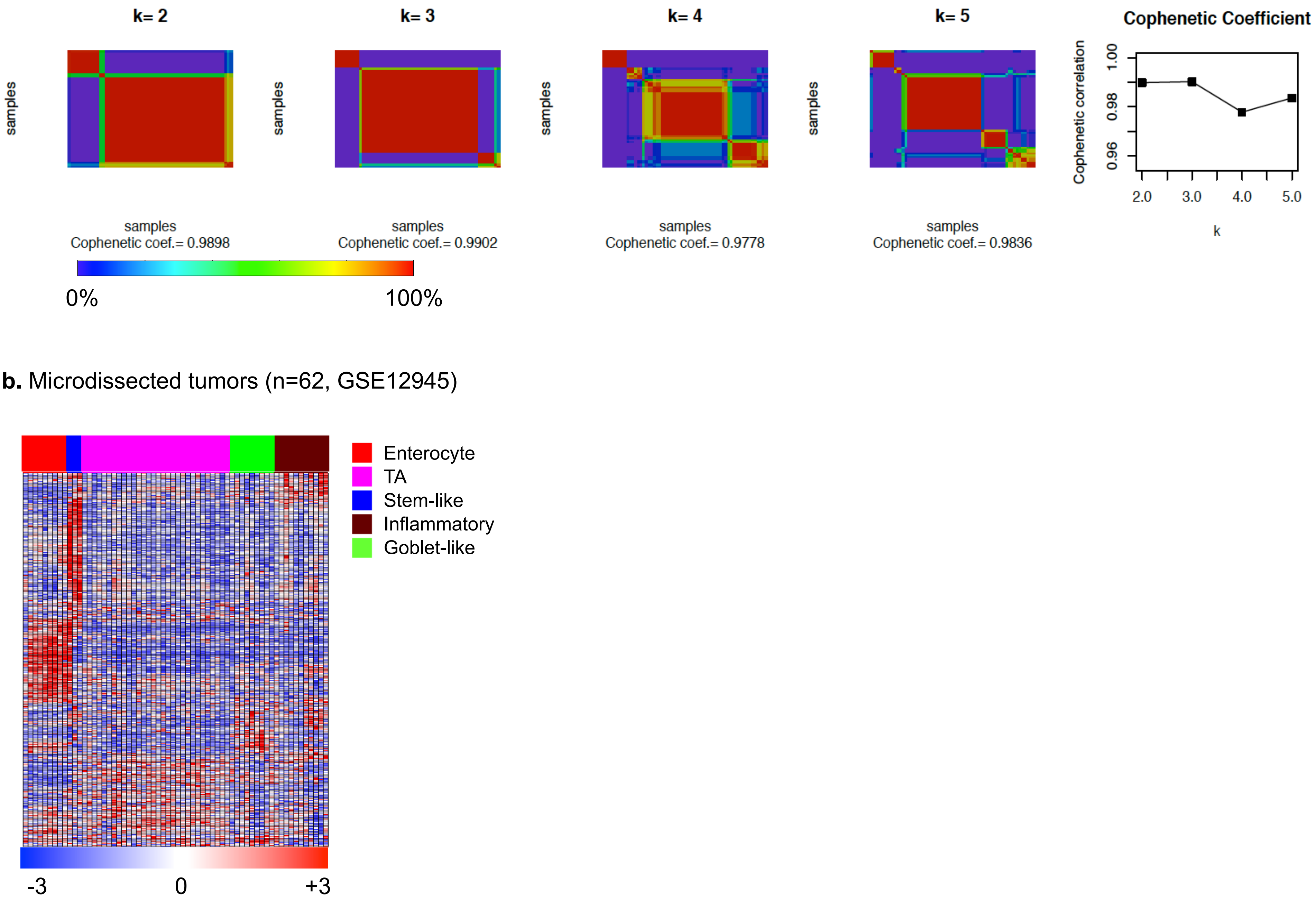
f. Inflammatory subtype-specific chemokine and interferon-related genes

Supplementary Fig. 2, con't

g. Quantitative RT-PCR assays for the identification of subtypes using patient CRC samples

Samples/ Markers	MUC2	TFF3	SFRP2	RARRES3	CFTR	Subtypes
CR560671	Neg	Neg	Neg	Neg	Pos	TA
CR560974	Neg	Neg	Neg	Neg	Pos	TA
CR560973	Neg	Neg	Neg	Neg	Pos	TA
CR560603	Pos	Neg	Neg	Neg	Neg	Enterocyte
CR561060	Pos	Pos	Neg	Neg	Neg	Goblet-like
CR559521	Pos	Pos	Neg	Neg	Neg	Goblet-like
CR560367	Neg	Neg	Neg	Pos	Neg	Inflammatory
CR559251	Neg	Neg	Pos	Neg	Neg	Stem-like
CR560476	Neg	Neg	Pos	Neg	Neg	Stem-like
CR560080	Neg	Neg	Pos	Neg	Neg	Stem-like

Supplementary Fig. 3

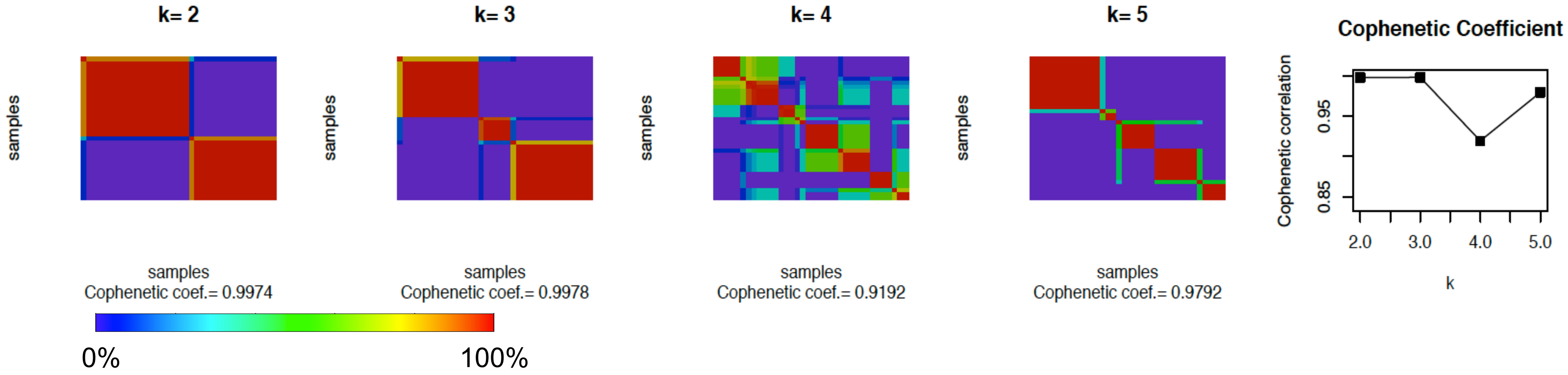


a. Microdissected tumors (n=62, GSE12945)

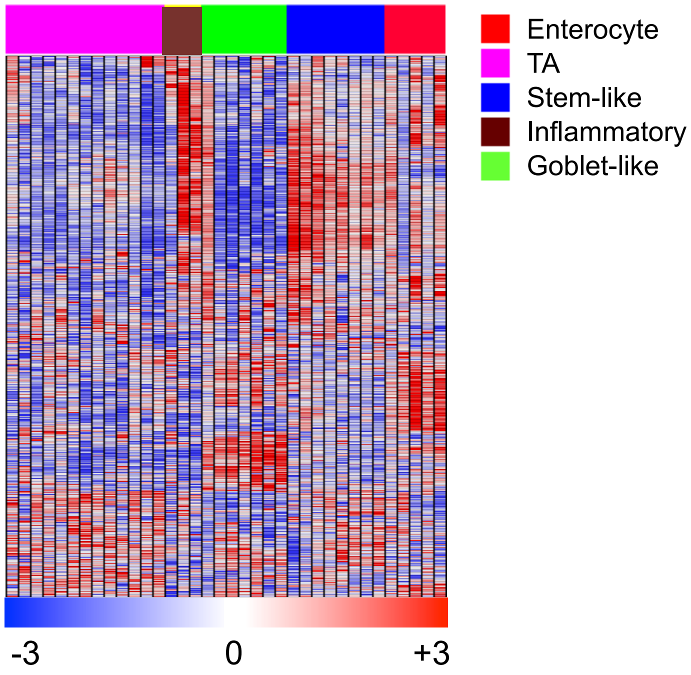
b. Microdissected tumors (n=62, GSE12945)

Supplementary Fig. 3, cont'd

c. Exon array data (n=36, GSE16125)

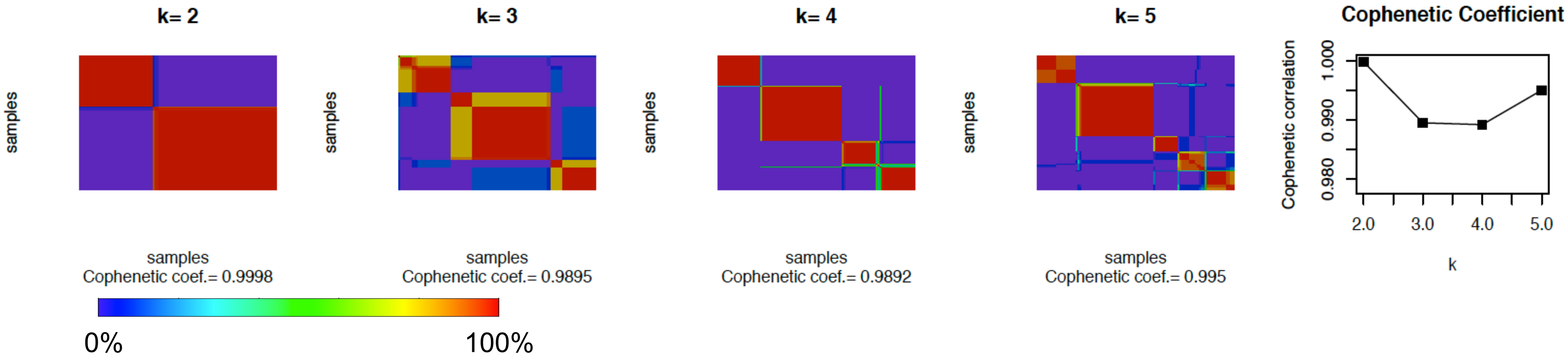


d. Exon array data (n=36, GSE16125)

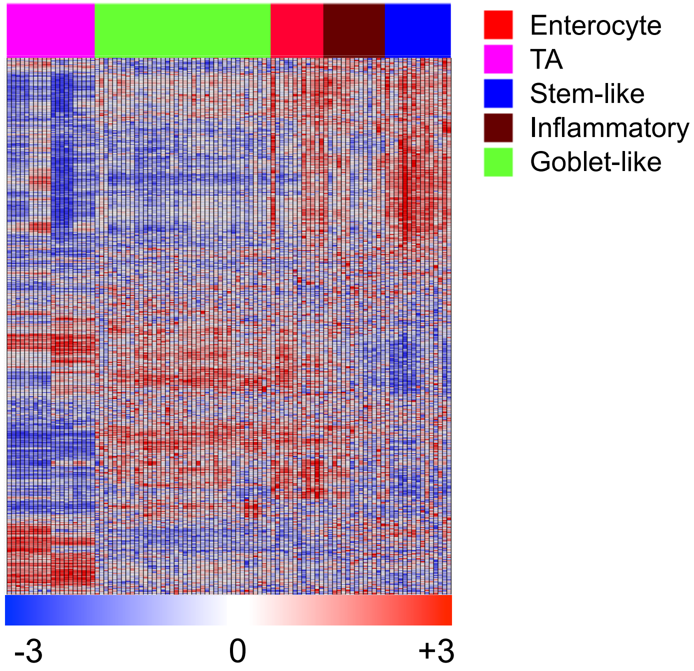


Supplementary Fig. 3, cont'd

e. Whole-tumor (n=101, GSE20916)

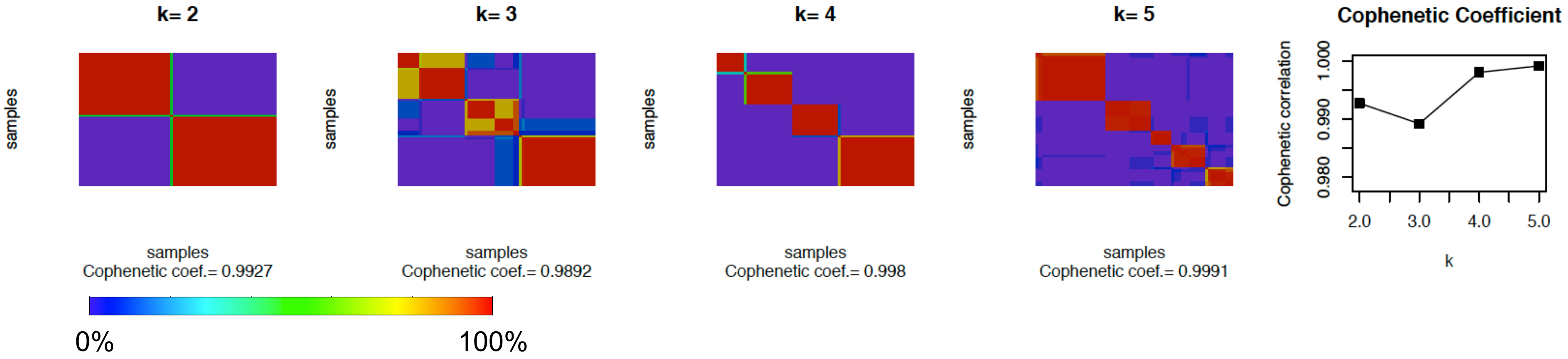


f. Whole-tumor (n=101, GSE20916)

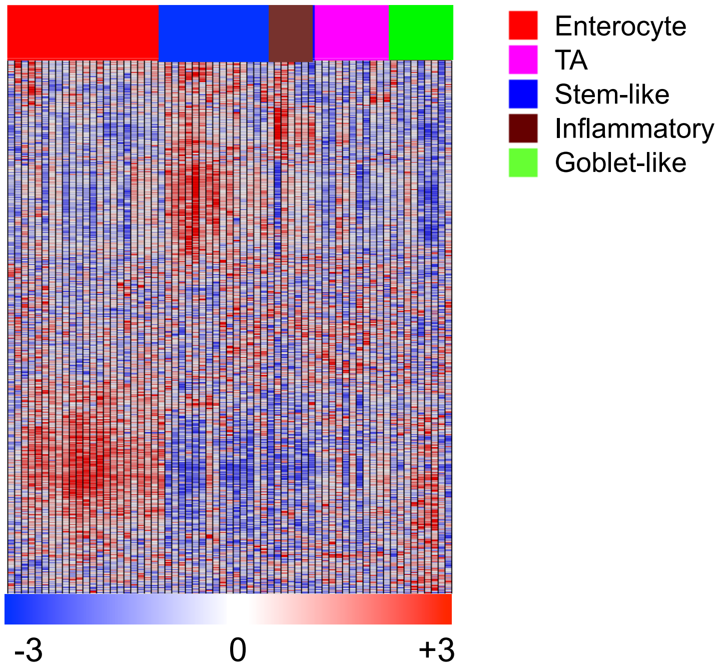


Supplementary Fig. 3, cont'd

g. Whole-tumor (n=65, GSE20842; Agilent-014850 Whole Human Genome Microarray 4×44K)

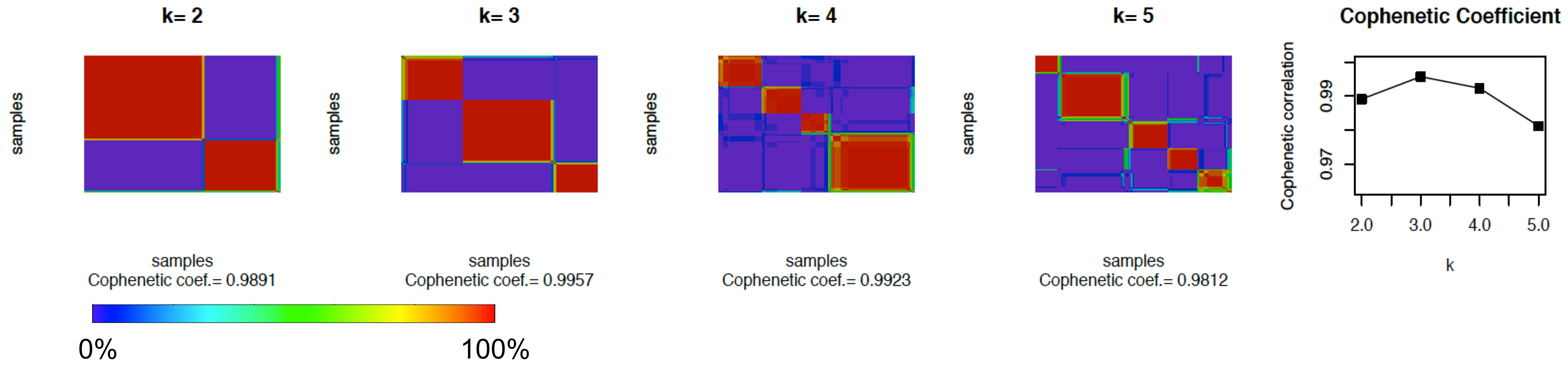


h. Whole-tumor (n=65, GSE20842; Agilent-014850 Whole Human Genome Microarray 4×44K)

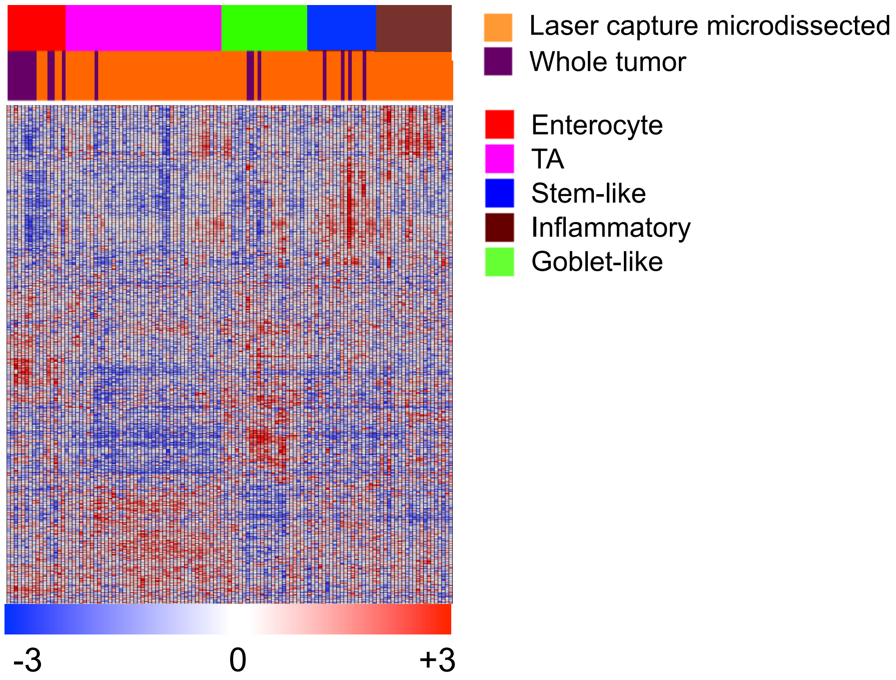


Supplementary Fig. 3, cont'd

i. LCM and whole tumor (n=123, GSE21510)

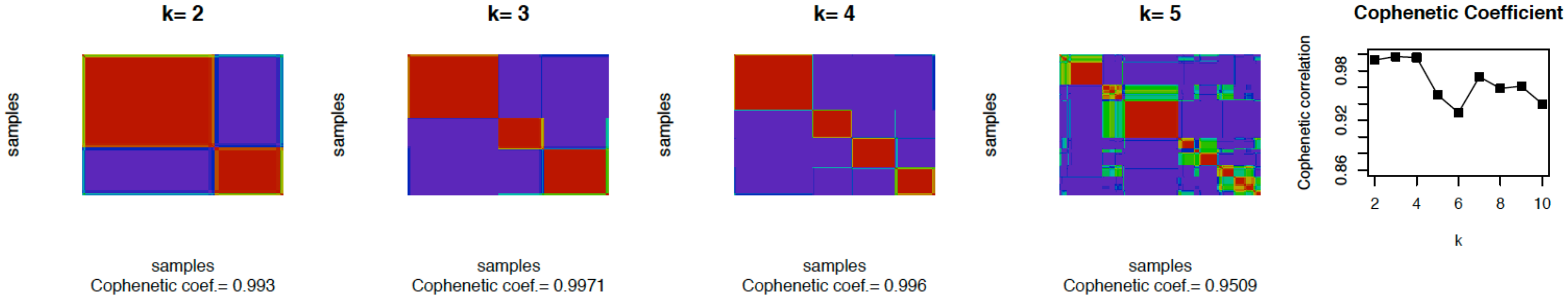


j. LCM and whole tumor (n=123, GSE21510)

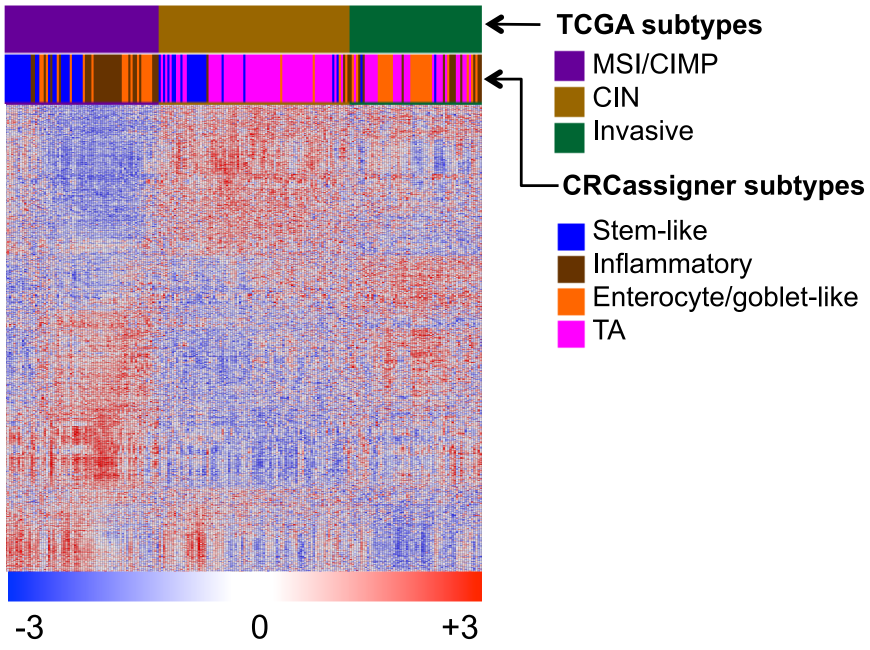


Supplementary Fig. 3, cont’d

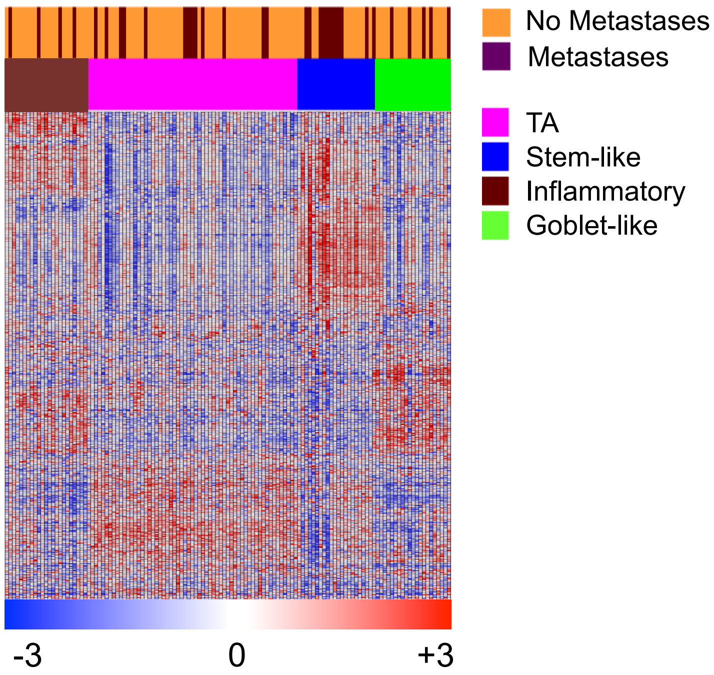
k. TCGA dataset (n=220)



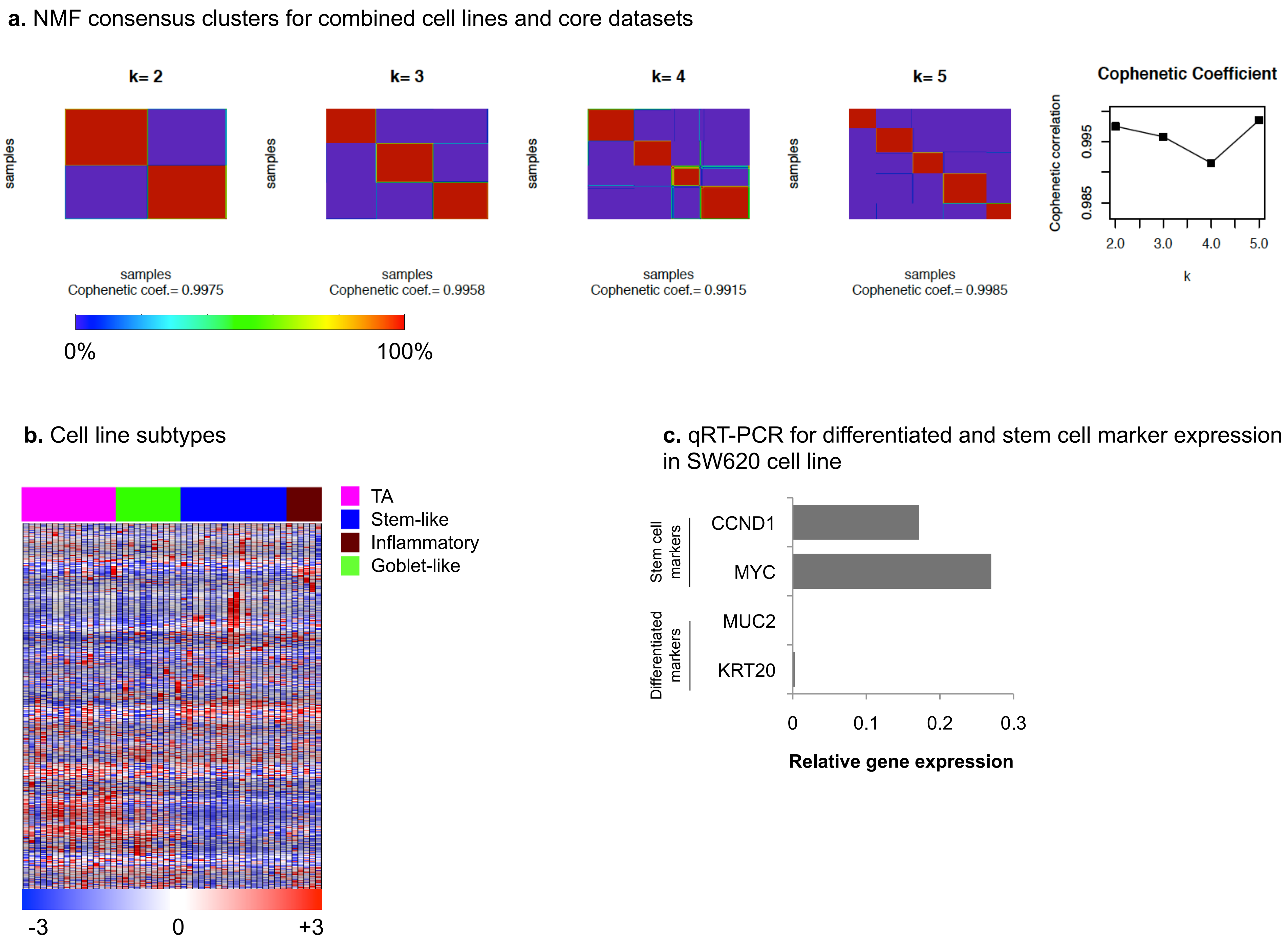
l. Comparison of TCGA subtypes and CRCassigner subtype using TCGA samples



m. Primary tumors with metastasis information (n=125, GSE28722)



Supplementary Fig. 4

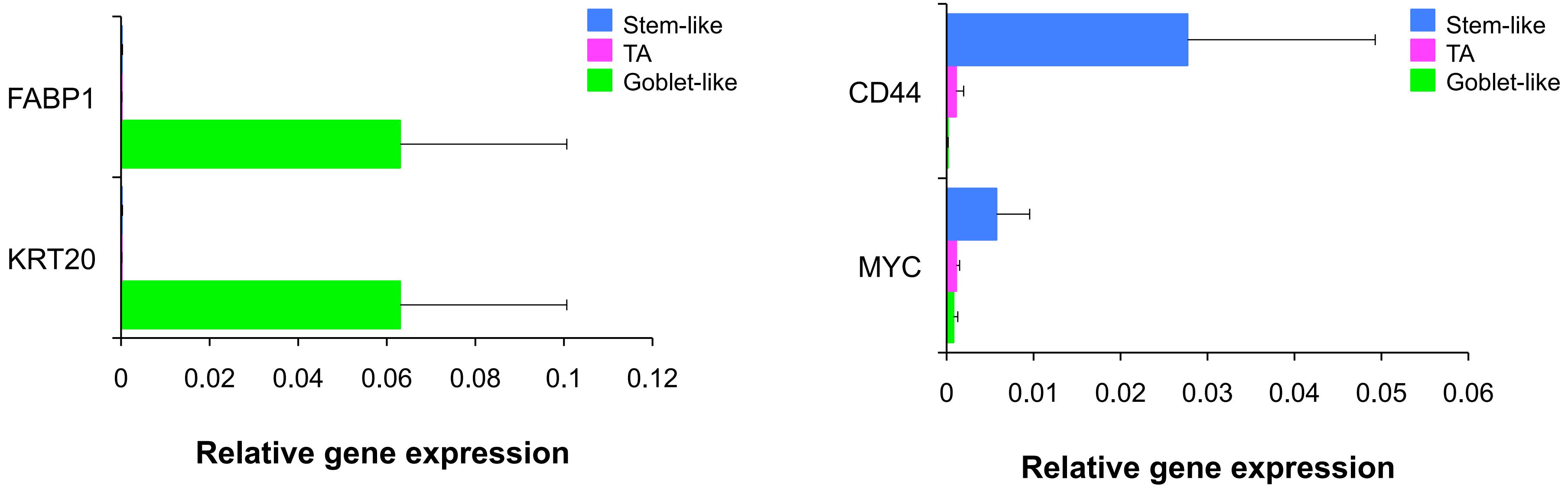


a. NMF consensus clusters for combined cell lines and core datasets

b. Cell line subtypes

c. qRT-PCR for differentiated and stem cell marker expression in SW620 cell line

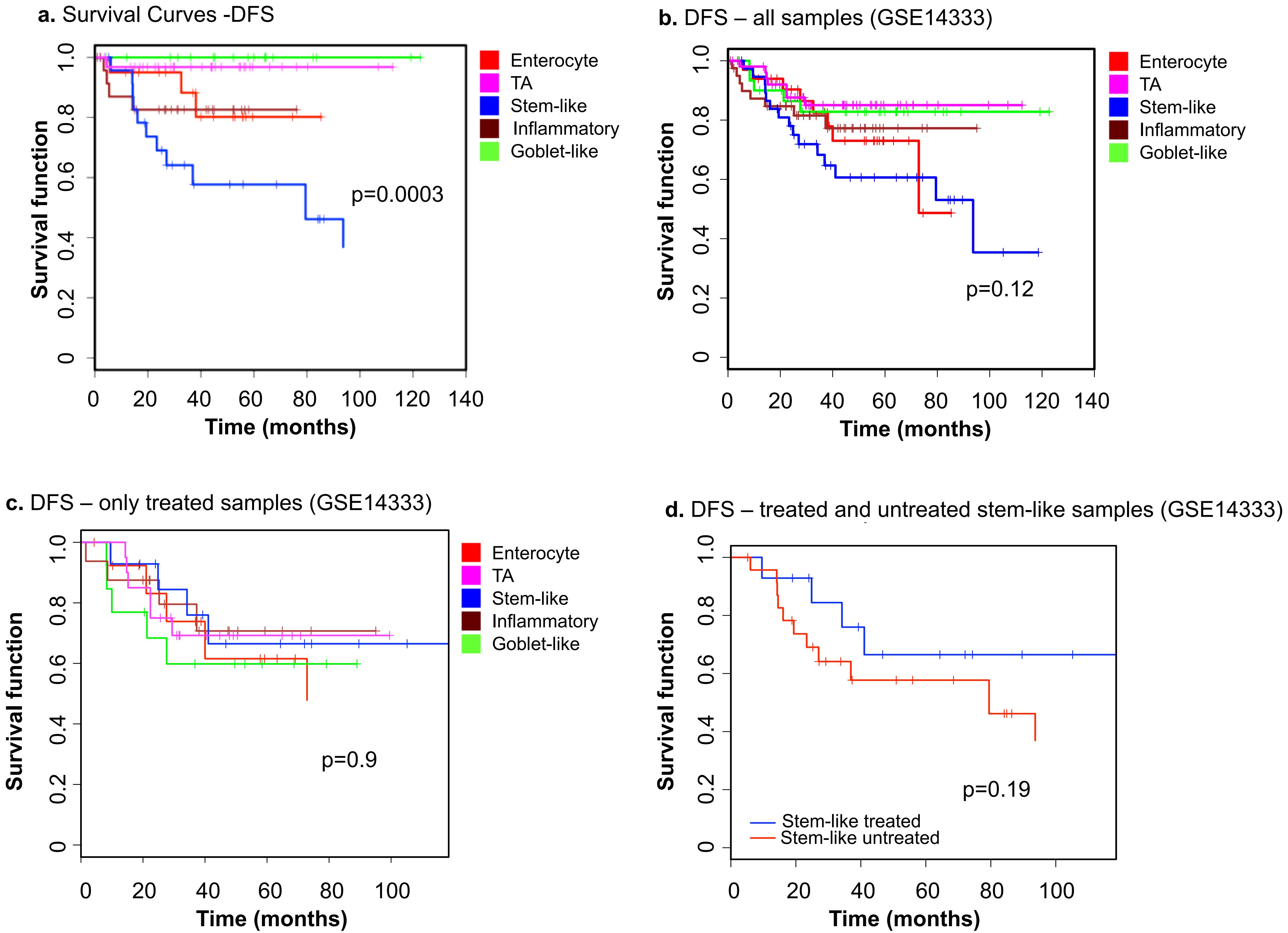
Supplementary Fig. 4, cont’d



d. Differentiated marker expression

e. Stem cell marker expression

Supplementary Fig. 5



a. Survival Curves -DFS

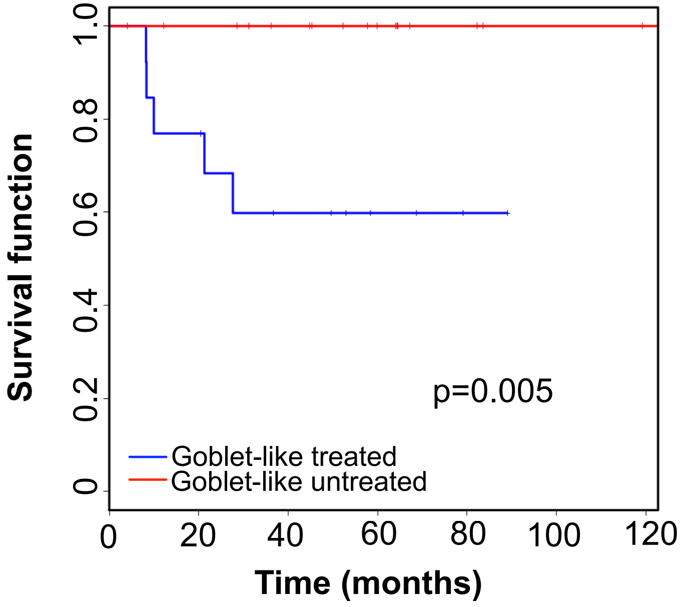
b. DFS – all samples (GSE14333)

c. DFS – only treated samples (GSE14333)

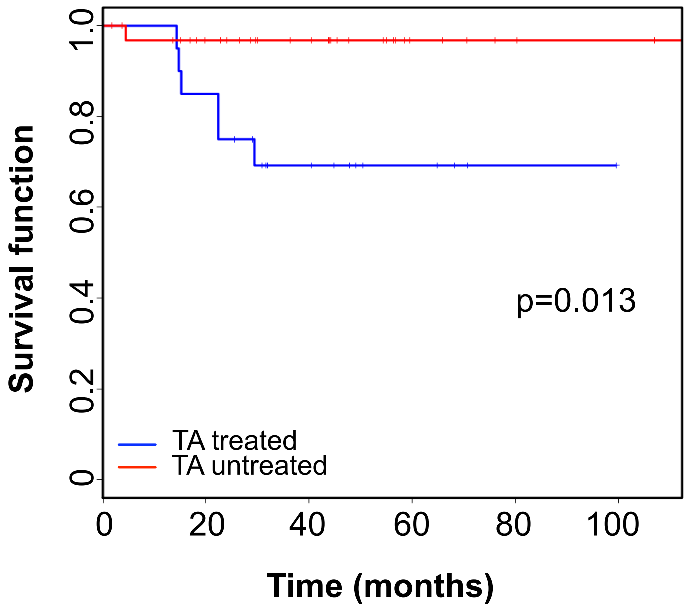
d. DFS – treated and untreated stem-like samples (GSE14333)

Supplementary Fig. 5, cont'd

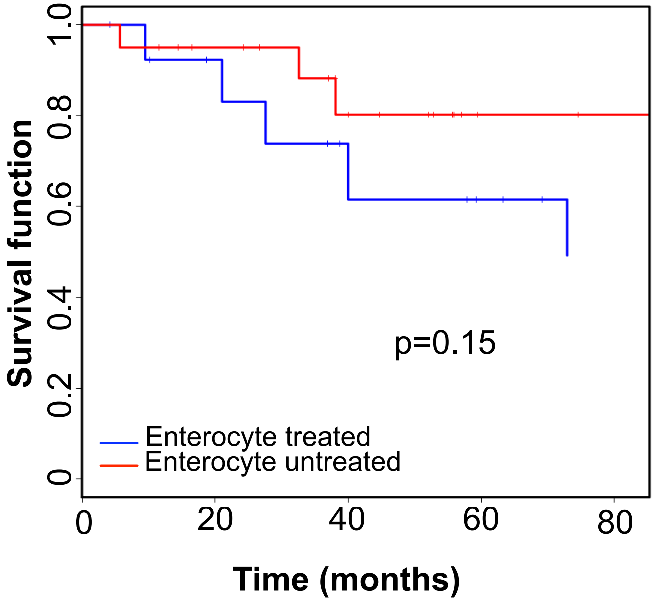
e. DFS – treated and untreated Goblet-like samples (GSE14333)



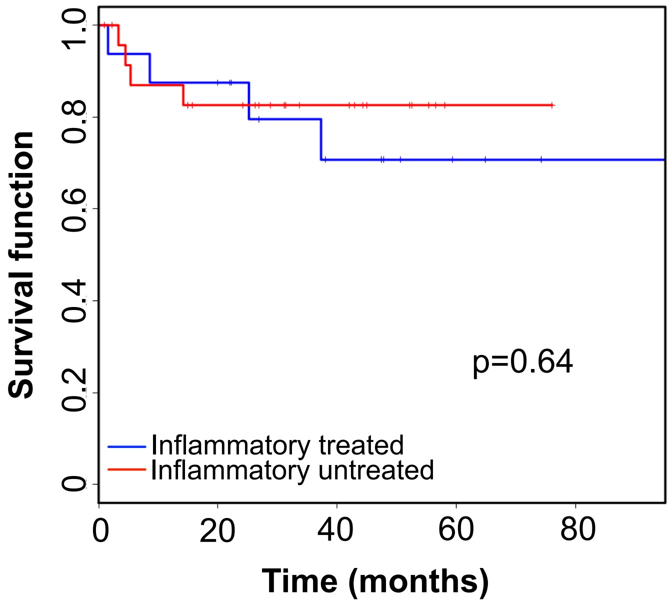
f. DFS – treated and untreated TA samples (GSE14333)



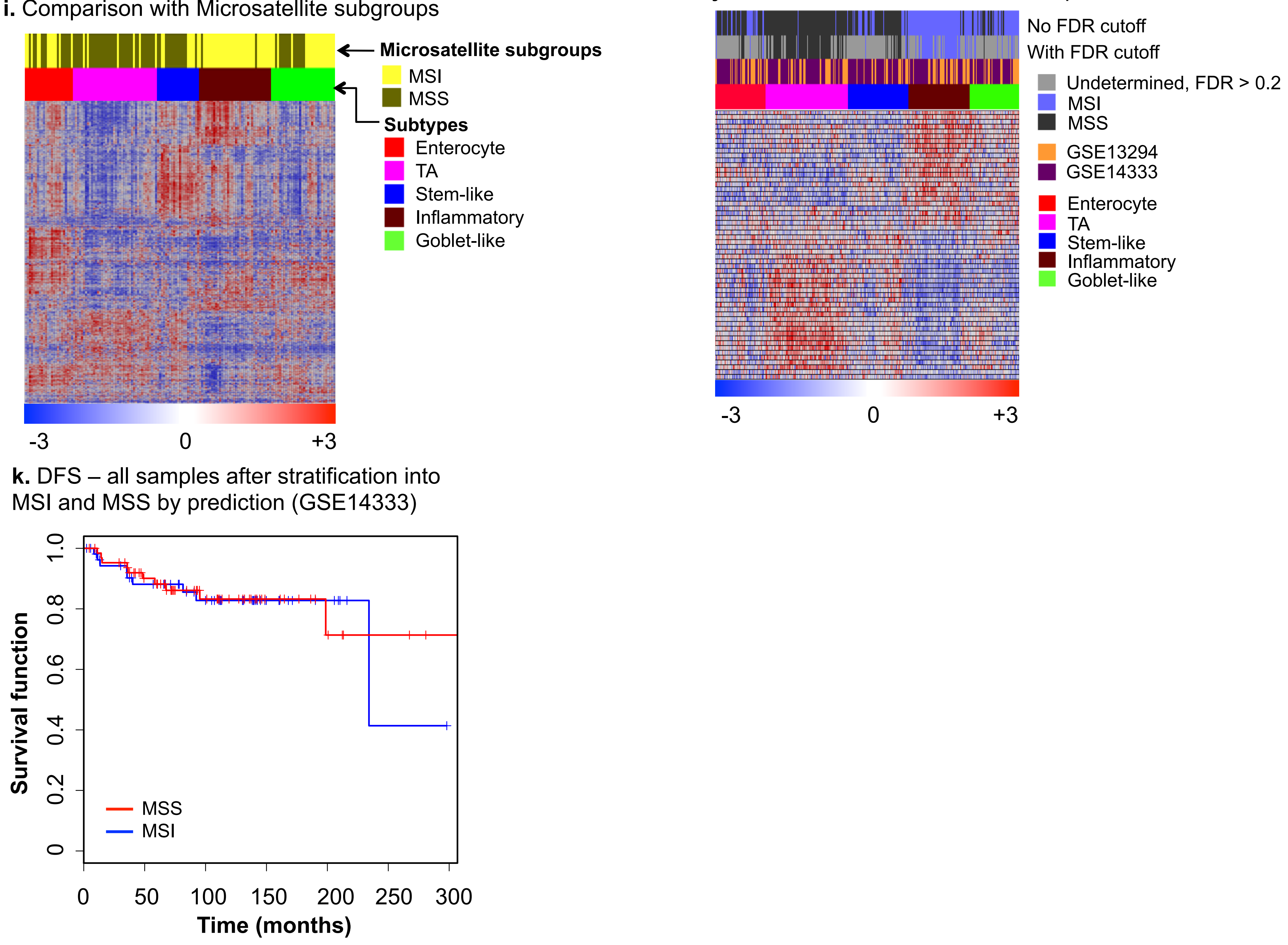
g. DFS – treated and untreated Enterocyte samples (GSE14333)



h. DFS – treated and untreated Inflammatory samples (GSE14333)



Supplementary Fig. 5, cont'd

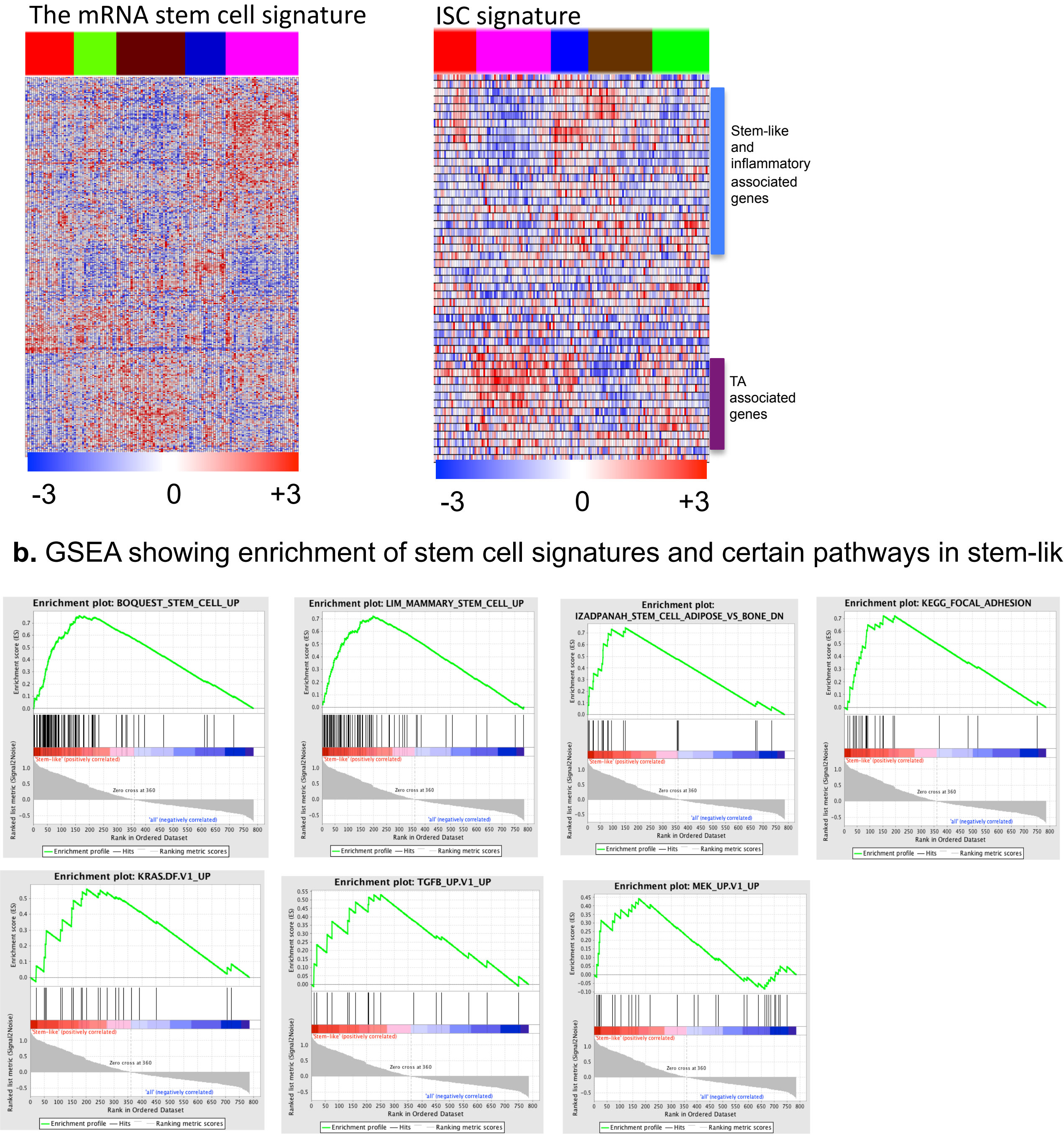


i. Comparison with Microsatellite subgroups

j. Microsatellite stable and instable prediction

k. DFS – all samples after stratification into MSI and MSS by prediction (GSE14333)

Supplementary Fig. 6

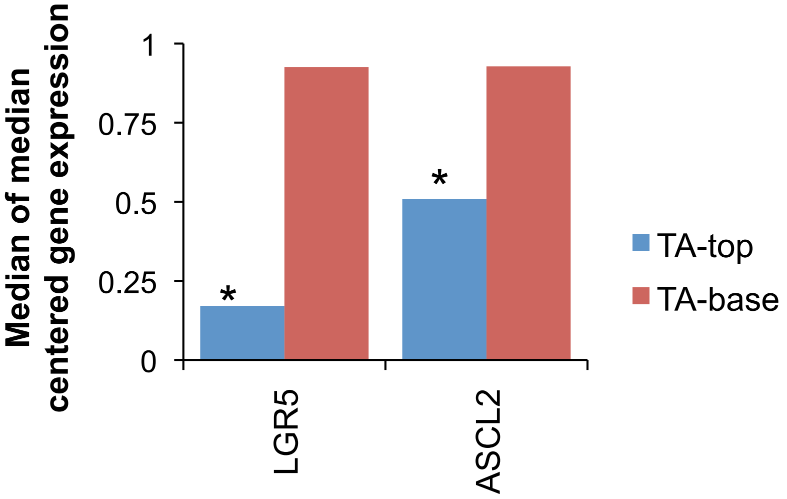


a. Association of stem cell signatures with CRC subtypes

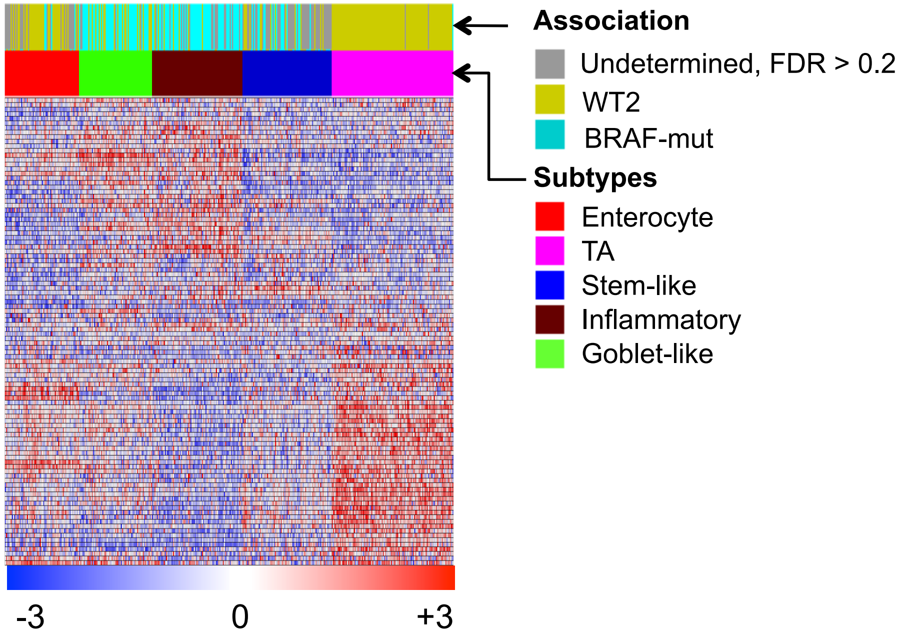
b. GSEA showing enrichment of stem cell signatures and certain pathways in stem-like subtype CRC samples

Supplementary Fig. 6, cont'd

c. Expression of Wnt targets in TA subtype samples from colon crypt top and base

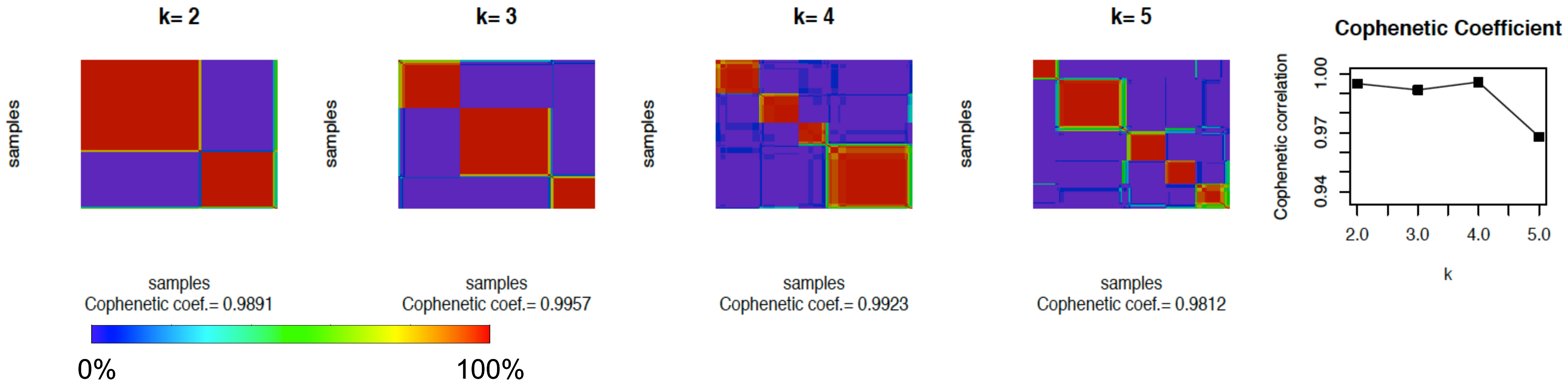


d. Association of BRAF-mut signature with subtypes



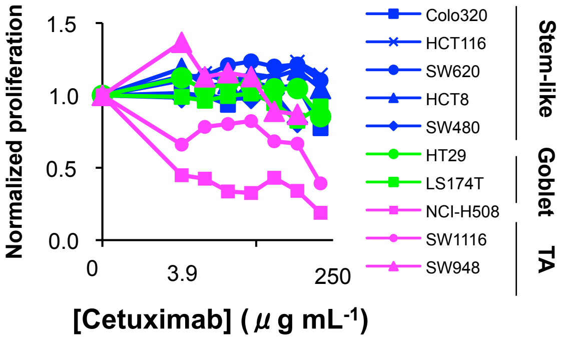
Supplementary Fig. 7

a. NMF consensus clustering Khambata-Ford liver metastases from CRC dataset



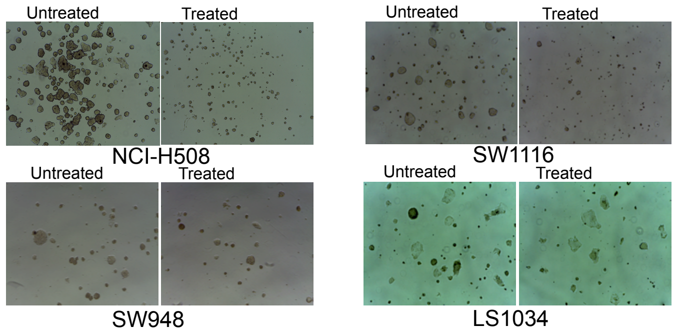
b. Subtype-specific Cetuximab sensitivity

Proliferation assay



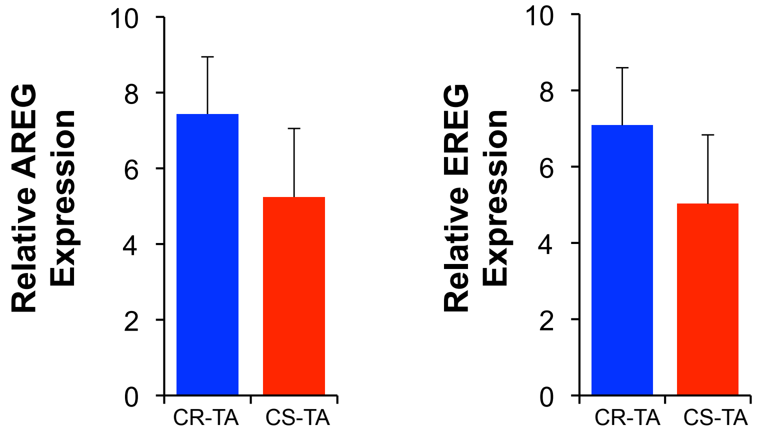
c. Subtype-specific Cetuximab sensitivity

Clonogenic assay

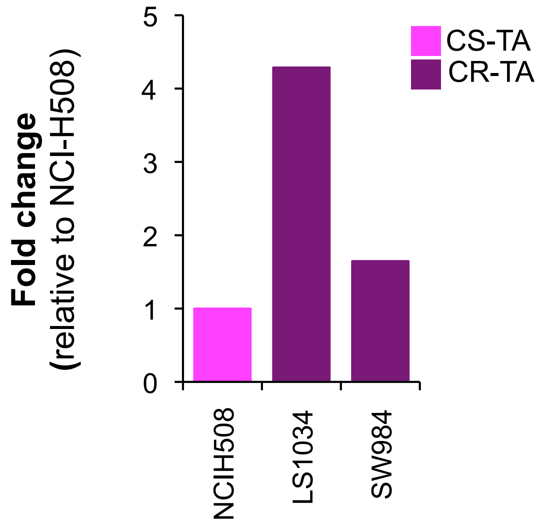


Supplementary Fig. 7, cont'd

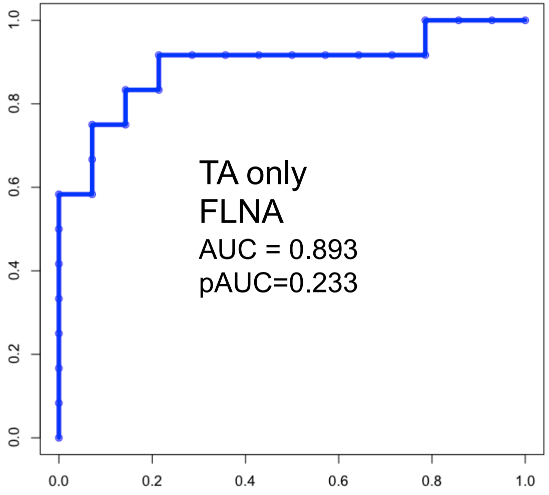
d. Expression among TA tumors in Cetuximab-treated Khambata-Ford data



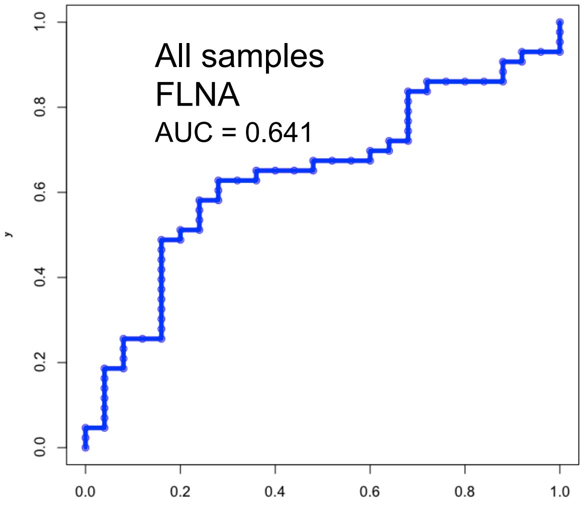
e. FLNA expression in TA cell lines



f. ROC curve for FLNA marker in TA samples only

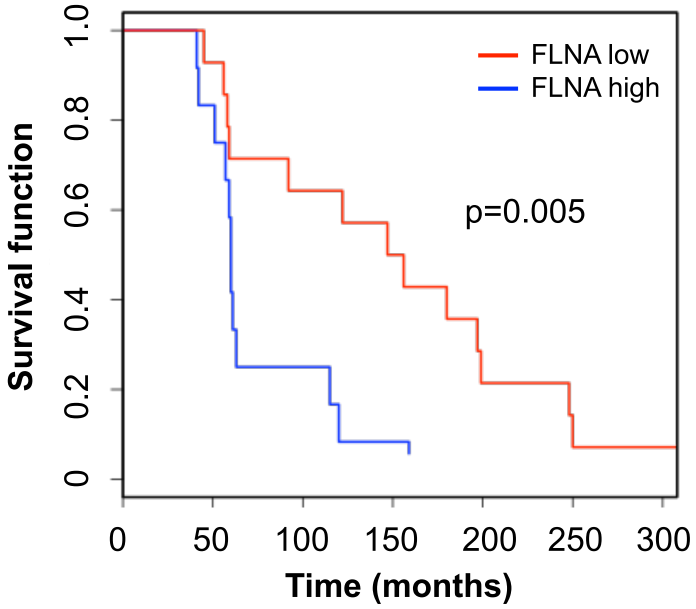


g. ROC curve for FLNA marker in all samples

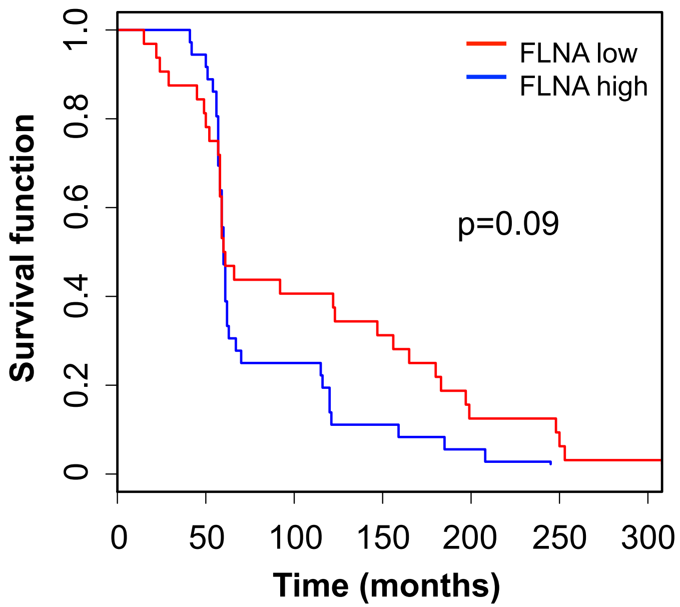


Supplementary Fig. 7, cont'd

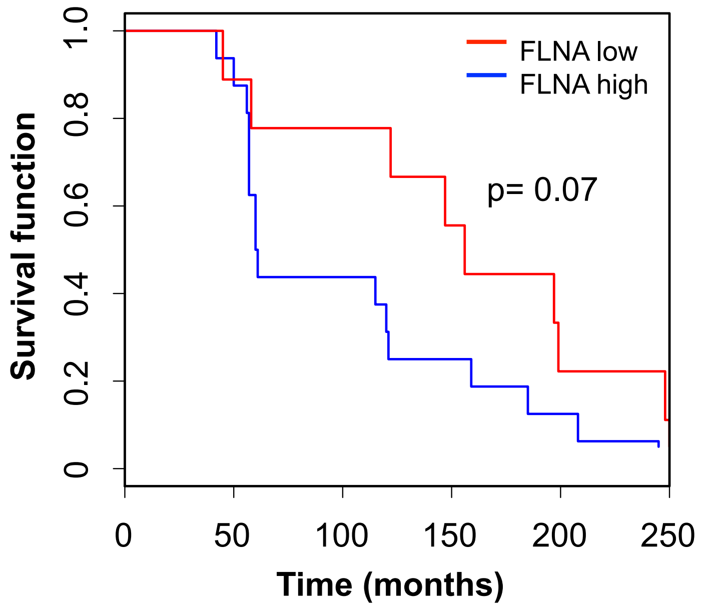
h. Survival curve among TA subtypes tumors based on FLNA expression



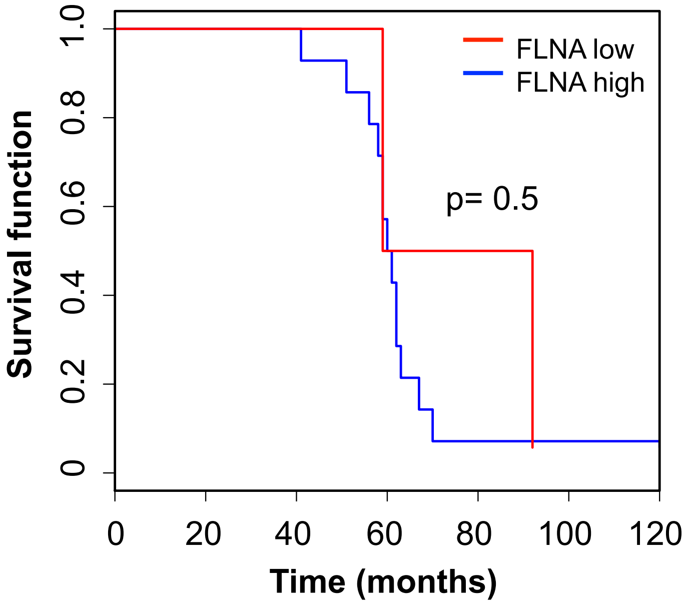
i. Survival curve, all samples



j. Survival curve, KRAS wild-type



k. Survival curve, KRAS Mutant



Supplementary Fig. 7, cont'd

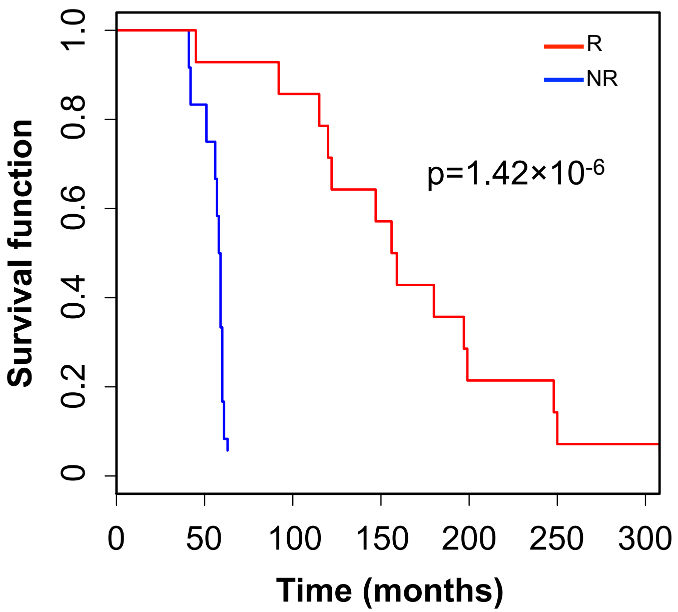
l. Quantitative RT-PCR assays for the identification of subtypes using patient CRC samples

Samples/ Markers	MUC2	TFF3	SFRP2	RARRES3	CFTR	Subtypes
CR560671	Neg	Neg	Neg	Neg	Pos	TA
CR560974	Neg	Neg	Neg	Neg	Pos	TA
CR560973	Neg	Neg	Neg	Neg	Pos	TA
CR560603	Pos	Neg	Neg	Neg	Neg	Enterocyte
CR561060	Pos	Pos	Neg	Neg	Neg	Goblet-like
CR559521	Pos	Pos	Neg	Neg	Neg	Goblet-like
CR560367	Neg	Neg	Neg	Pos	Neg	Inflammatory
CR559251	Neg	Neg	Pos	Neg	Neg	Stem-like
CR560476	Neg	Neg	Pos	Neg	Neg	Stem-like
CR560080	Neg	Neg	Pos	Neg	Neg	Stem-like

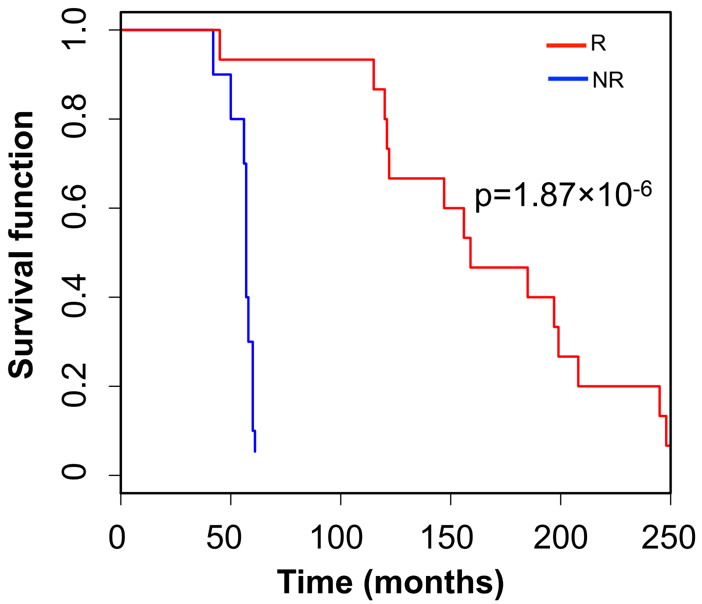
m. Quantitative RT-PCR assays for the identification of subtypes including sub-subtypes of TA

Samples	MUC2	TFF3	SFRP2	RARRES3	CFTR	FLNA	Subtypes
CR560671	Neg	Neg	Neg	Neg	Pos	Pos	CR-TA
CR560974	Neg	Neg	Neg	Neg	Pos	Pos	CR-TA
CR560973	Neg	Neg	Neg	Neg	Pos	Neg	CS-TA

n. Survival curve, TA subtype

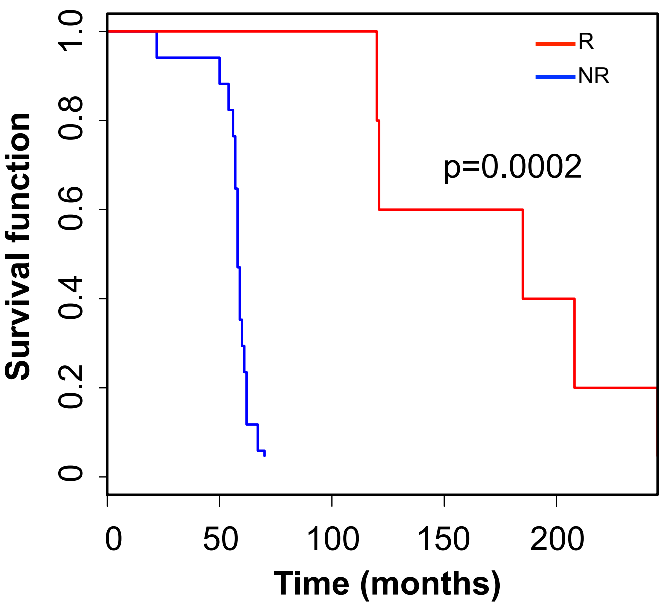


o. Survival curve, KRAS WT (except unknown)

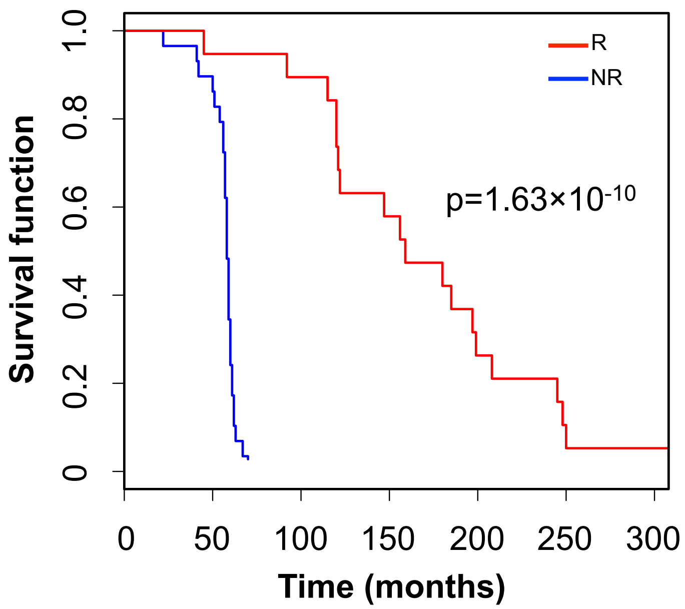


Supplementary Fig. 7, cont'd

p. Survival curve, non-TA samples (except unknown)

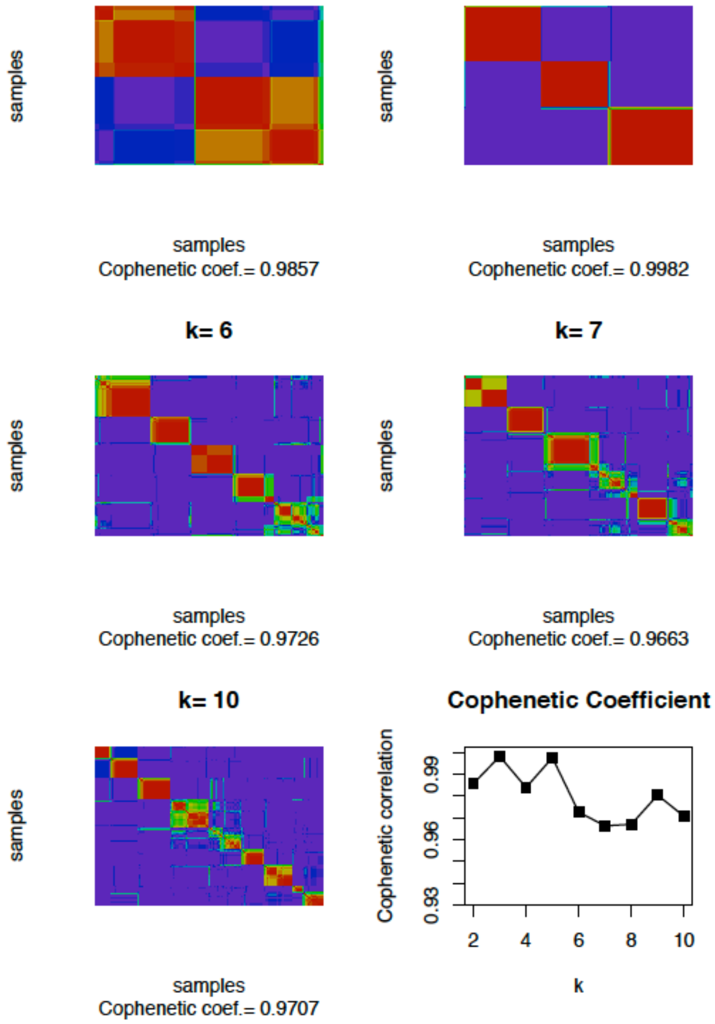


q. Survival curve, all samples (except unknown)

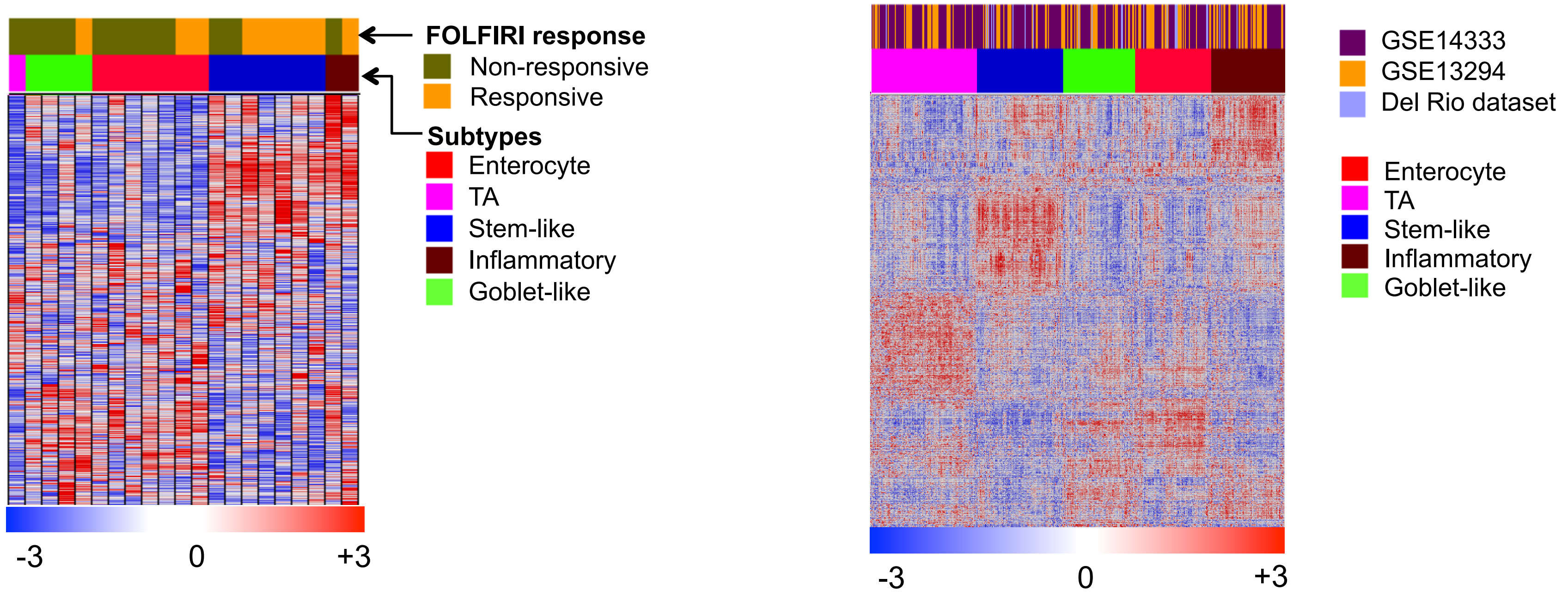


Supplementary Fig. 8

a. NMF consensus clusters for merged Del Rio (n = 21) and core CRC datasets



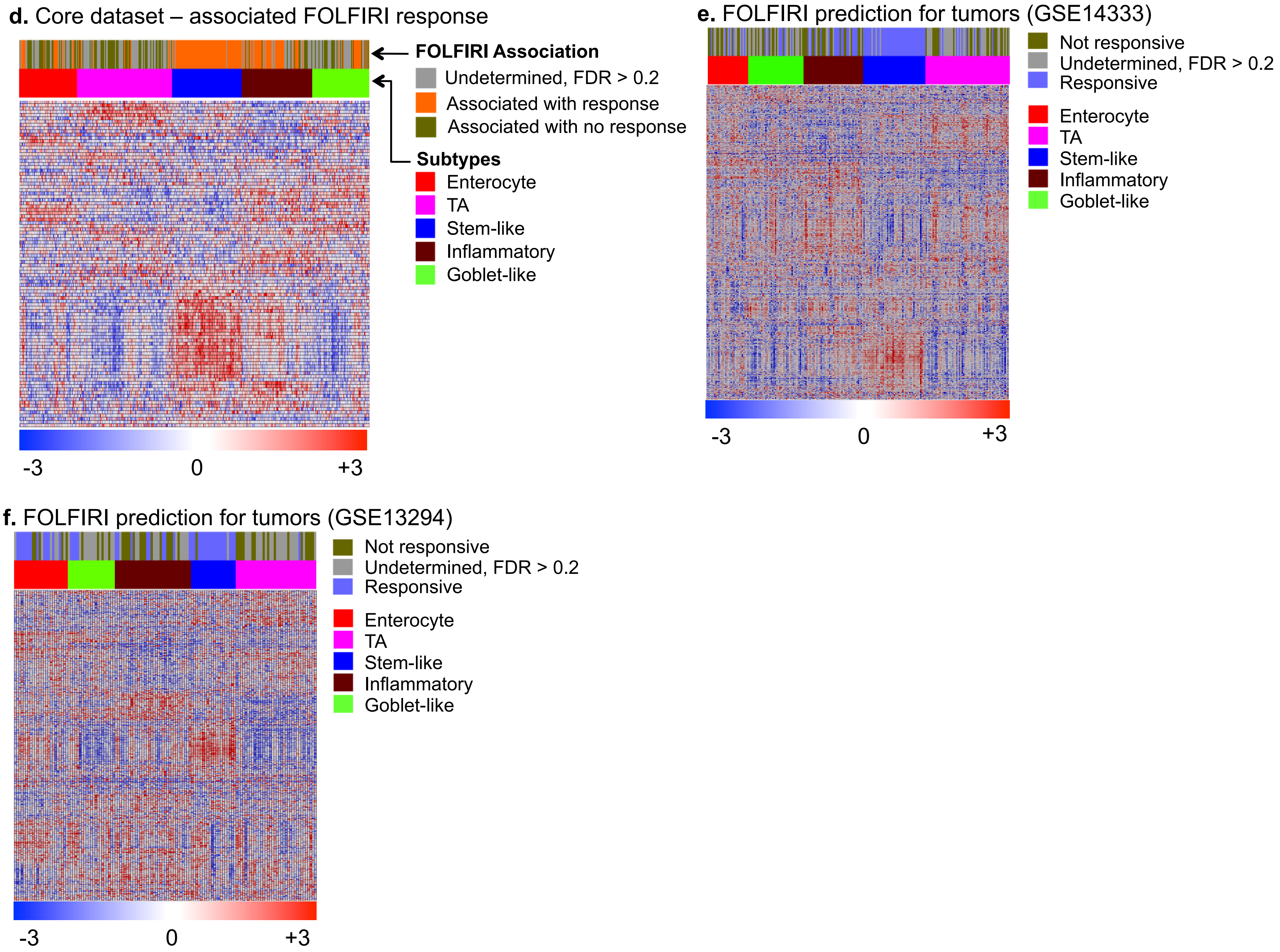
Supplementary Fig. 8, con't



b. Subtypes and FOLFIRI response in Del Rio dataset

c. Combined Del Rio and core tumor datasets and subtypes

Supplementary Fig. 8, con't



d. Core dataset – associated FOLFIRI response

e. FOLFIRI prediction for tumors (GSE14333)

f. FOLFIRI prediction for tumors (GSE13294)

A colorectal cancer classification system that associates cellular phenotype and responses to therapy

Anguraj Sadanandam[1,2], Costas A Lyssiotis[3,4,14,15], Krisztian Homicsko[2,5,15], Eric A Collisson[6], William J Gibb[7], Stephan Wullschleger[2], Liliane C Gonzalez Ostos[2], William A Lannon[3,14], Carsten Grotzinger[8], Maguy Del Rio[9], Benoit Lhermitte[10], Adam B Olshen[11,12], Bertram Wiedenmann[8], Lewis C Cantley[3,4,14], Joe W Gray[13] & Douglas Hanahan[2]

[1]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [2]Swiss Institute of Experimental Cancer Research, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. [3]Department of Medicine, Division of Signal Transduction, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [5]Service of Medical Oncology, Department of Oncology, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland. [6] Division of Hematology and Oncology, University of California–San Francisco, San Francisco, California, USA. [7]Genomic Health, Redwood City, California, USA. [8]Department of Hepatology and Gastroenterology, Charité, Campus Virchow-Klinikum, University Medicine Berlin, Berlin, Germany. [9]Institut de Recherche en Cancérologie de Montpellier, Institut National de la Santé et de la Recherche Médicale, U896, Université Montpellier, Centre Régional de Lutte contre le Cancer, Val d'Aurelle Paul Lamarque, Montpellier, France. [10]University Institute of Pathology, CHUV, Lausanne, Switzerland. [11]Department of Epidemiology and Biostatistics, University of California–San Francisco, San Francisco, California, USA and [12]Helen Diller Comprehensive Cancer Center, University of California–San Francisco, San Francisco, California, USA. [13]Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon, USA. [14]Present address: Department of Medicine, Weill Cornell Medical College, New York, New York 10065, USA. [15]These authors contributed equally to this work.

Correspondence should be addressed to J.W.G. (grayjo@oshu.edu) or D.H. (douglas.hanahan@epfl.ch).

Supplementary Results and Discussions

Overview of CRC subtype stratification and biomarker identification. A primary goal of this study was to determine whether we could detect novel molecular subtypes of CRC based on published microarray data, and if so, to identify biomarkers to effectively and practically classify samples into the detected subtypes. As has been shown in studies of other types of cancers[1-7], novel subtypes can be identified using unsupervised clustering methods. In order to detect *multiple* subtypes (some of which may represent relatively small fractions of the patient population), the clustering methods require moderately large numbers of samples – more than contained in any one of the individual CRC data sets published to date. With that in mind, we began our analysis by identifying suitable and comparable microarray data sets (**Supplementary Table 2**) and selecting only those data sets that were described in Dalerba, *et. al*[8] as not having redundant samples. Once the data sets were selected, the raw gene expression readouts were either normalized using robust multiarray averaging (RMA)[9] or obtained as processed data from the authors (see **Supplementary Table 2a**) and then pooled using distance weighted discrimination (DWD)[10] method after normalizing each data set to N(0,1). Consensus-based non-negative matrix factorization (NMF)[11] was applied to the pooled data to cluster the samples into the initial set of three and then five CRC subtypes. Although consensus-based clustering algorithms can be used to detect robust clusters (*i.e.* clusters that tolerate a moderate degree of outlier contamination in the training set), the identification of genes (or markers) specific to each cluster is somewhat more sensitive to samples representing rare subtypes or samples of indeterminate origin. Therefore, once the clusters (subtypes) were identified using NMF, we used silhouette width[1,12] (similar to that used by TCGA for classifying glioblastoma[2]) to screen out those samples residing on the periphery of the NMF-identified clusters. From there, we applied well-established methods (Significance Analysis of Microarrays, SAM[13]; and Prediction Analysis for Microarrays, PAM[14]) to extract biomarkers associated with the screened subtypes. The summary of the methodology is shown in **Supplementary Fig. 1a** and the details of our analysis are described in the subsections below.

1.1. Pooling data sets using DWD. When pooling microarray data, one of the main challenges is to pool the microarray data sets in such a way as to compensate for systematic biases (*e.g.* batch effects) without distorting or collapsing biologically informative and subtype-discriminating structures in the gene expression space. In this respect, Benito *et. al*[10] applied a method known as DWD to pool microarray data and showed that DWD demonstrates superior pooling characteristics when compared to alternative methods such as singular value decomposition (SVD) and Fisher linear discrimination, especially for high-throughput gene expression data in which we must contend with small numbers of samples relative to the number of gene expression readouts (*i.e.* a high dimensional features space). As a variation on the support vector machine (SVM) approach, DWD is suitable for high dimensional features spaces, but it has the added benefit of minimizing the effects that data artifacts and outliers can have on the batch effect adjustments.

1.2. Unsupervised clustering using consensus-based NMF. By itself, NMF[11] is a dimensionality reduction method in which we can attempt to capture the salient functional properties of a high-dimensional gene expression profile using a relatively small number of "metagenes" (defined to be non-negative linear combinations of the expression of individual genes – *i.e.* a weighted average of gene expression, with each metagene having its own set of weighting coefficients). As with principal component analysis, the familiar gene expression table (samples x genes) is factored into two lower-dimensional matrices except that for NMF the matrix factors are constrained to be purely non-negative values. This 'non-negativity' constraint is believed to more realistically represent the nature of gene expression[11], in that gene expression is either zero- or positive-valued. In contrast, principal component analysis (PCA) matrix factors can be either positive- or negative-valued.

Given an arbitrary gene expression table (profile), it is not generally possible to analytically factor the table into two matrix factors. As a consequence, numerical algorithms have been developed[15] to accomplish this by first initializing the two matrices to random values and then iteratively updating the matrices using a search algorithm. There is no guarantee that this search algorithm will converge to a globally optimal factorization, hence one re-runs the algorithm using multiple random initial conditions to see whether the algorithm provides a consistent factorization. At the end of the factorization algorithm, one obtains two lower-dimensional matrices, which when multiplied together will yield an approximation to the original gene expression table. The metagenes correspond to functional properties represented in the original gene expression table and can be viewed as 'anchors' for clustering the samples into subtypes. Specifically, each sample is assigned to a subtype by finding which metagene is most closely aligned with the sample's gene expression profile. Hence each sample is assigned to one and only one cluster.

As explained above, the robustness of clustering can be gauged by repeating the factorization process several times using different random initial conditions for the factorization algorithm. If the factorization is insensitive to the initial conditions of the search algorithm, then any pair of samples will tend to co-cluster irrespective of the initial condition. By keeping track of the pairwise co-clustering, we can graphically represent the clustering "consensus" by plotting the frequency with which two samples co-cluster. This is the basis of the consensus plots illustrated in **Supplementary Fig. 1b, 3**, where we have color-coded consensus as ranging from "always co-clustering" (red) to "never co-clustering" (blue), with intermediate colors representing "occasional" co-clustering. When there is high consensus clustering, the boundaries separating red from blue regions will be sharp. In this case, we see consistent pairwise co-clustering irrespective of the initial condition on the factorization algorithm. When consensus clustering is poor, the boundaries will be 'fuzzy' with bands of intermediate colors separating the red and blue regions as shown in **Supplementary Fig. 1b**. Given k clusters, consensus can be quantitatively summarized using a single scalar-valued function known as the cophenetic coefficient, ρ_k (defined in Brunet, et al[11]) which ranges in value from 0.0 to 1.0. When the consensus clustering is sharp, ρ_k will be close to 1.0. When the consensus is blurry, ρ_k will be closer to 0.0. We generally seek high consensus clusters, *i.e.* those for which ρ_k is close to 1.0, to be confident that the chosen number of clusters is robustly supported by the data. When computing our NMF consensus plots (and associated cophenetic coefficients), we used 20 different initial conditions for the factorization algorithm for each value of k. For the core data sets, the values of k were varied from 2 to 10 hence we obtain a consensus plot (and cophenetic coefficient) for each k=2,…10 and seek values of k for which ρ_k is close to 1.0.

In the NMF consensus analysis of the core data sets, we found good consensus for both k=3 and k=5 clusters, suggesting that there was evidence for 5 consensus clusters and hence 5 functional properties in the core data sets.

1.3. Removing outliers using silhouette width. For the purposes of identifying subtype-specific markers, our analysis includes only those samples that belong statistically to the core of each of the clusters. Excluding samples with negative silhouette width[12] has been shown to minimize the impact of sample outliers on the identification of subtype markers, as described in TCGA glioblastoma classification[1]. Accordingly, 58 samples from the original 445 were identified as having negative silhouette width and were therefore excluded from the marker identification phase of the analysis (**Supplementary Fig. 1f**).

1.4. Identification of subtype-specific biomarkers using SAM and PAM. We used a two-step process to identify subtype-specific biomarkers. The first step identifies the differentially expressed genes and the second step finds those genes that are associated with specific subtypes. For the first step, we used SAM[13] to identify genes significantly differentially expressed across the 5 subtypes. This is a well-established method that looks for large differential gene expression relative to the spread of expression across all genes. Sample permutation is used to estimate false discovery rates (FDR) associated with sets of genes identified as differentially expressed. By adjusting a sensitivity threshold, Δ_{SAM}, users can control the estimated FDR associated with the gene sets. For our analysis, we selected Δ_{SAM} = 12.2, which yielded 786 differentially expressed genes and an FDR of zero (**Supplementary Table 1a**). The second step in the process was to match the differentially expressed genes to specific subtypes. For this step, we used PAM[14], which is similar in nature to the centroid method recently applied by the TCGA consortium to glioblastoma data[1], except that PAM eliminates the contribution of genes that differentially express below a specific threshold, Δ_{PAM}, relative to the subtype-specific centroids. A threshold parameter or scale of Δ_{PAM} = 2 was chosen after evaluating various Δ_{PAM} values and misclassification errors. Leave out cross validation (LOCV) analysis was then performed to identify a set of genes that had the lowest prediction error. We identified all of the 786 SAM selected genes that had the lowest prediction error of 5.4% after PAM and LOCV analysis. The resulting subtype-specific markers (CRCassigner) are listed in **Supplementary Table 1b**. With this the overall description of the computational methods and analysis used in this study to identify gene expression subtypes has been completed.

Reason for choosing five over three CRC subtypes. Although the cophenetic coefficient score (greater than 0.99) from the NMF consensus clustering of the merged core CRC data set is highest for the cluster k=3 (three subtypes), we chose k=5 (five subtypes, **Supplementary Fig. 1b**). As illustrated in **Supplementary Fig. 1e** for k=3, the subtype 1 and subtype 2 in the heatmap show heterogeneity in gene expression patterns as revealed by subtype-specific signatures 1A, 1B, 2A and 2B in the side bars. In addition, using all the three different SD cut-offs, we found consistent support for 3 to 5 subtypes (**Supplementary Fig. 1c)**. This demonstrates that the consensus support for 3 to 5 clusters is fairly insensitive to the SD threshold across the range of SD thresholds flanking SD=0.8. Taken together with 3 subtypes, these distinct signatures suggest a total of 5 CRC subtypes, and this is similar to that discussed elsewhere in Brunet, et al[11]. This decision has been validated by the subsequent analysis documenting clear differences amongst the 5 subtypes.

Association of CRC subtypes to colon crypt top/base using NTP. To associate CRC subtypes to colon crypt top/base, we used a previously published gene signature[16] of the colon crypt base (see **Fig. 2a**) together with NTP[17]. The analysis confirmed that majority of the samples from the NMF-identified stem-like subtype were associated with the crypt base signature. Hoshida *et al.*[17] proposed the NTP method as a way of associating individual samples to known gene signatures even if the published gene signature was derived from data acquired from a different gene expression platform. This is accomplished by splitting the up- and down-regulated signature genes into two groups to form a dichotomized gene expression template. The similarity of a sample's gene expression profile to the template is computed using a nearest neighbor approach. By random sub-sampling the gene space, NTP estimates a null distribution of similarity coefficients. Then the similarity coefficient obtained using the published gene signature can be compared to the null distribution so as to compute a p-value. The same approach was followed for the association of CRC subtypes to Wnt signaling (**Fig. 2a**) and FOLFIRI response (**Fig. 4b,c**) using specific signatures as described in the main text.

Statistics for association of colon-crypt top/base to CRC subtypes. After performing the NTP algorithm[17] based prediction for association of colon-crypt top/base to each sample using a published gene signature that discriminates between the normal colon crypt top and the normal crypt base[16], we observed statistically significant (only for samples with FDR<0.2) associations as reported in the main text (**Fig. 2a**). Here, we are reporting the statistics for all the samples irrespective of the FDR cut-off. We observed that 55% (n=77) of the stem-like subtype is associated with the crypt base whereas 33% (n=105) of TA, 43% (n=63) of goblet-like and 75% (n=64) of enterocyte subtypes are associated with the crypt top. On the other hand, we observed that more than 80% (n=78) of the inflammatory subtypes have no significant association with either the crypt base or top.

Validation of subtypes in additional data sets. In order to validate the five subtypes in additional data sets, including TCGA CRC data set, we mapped the SAM and PAM genes specific to each subtype onto each of the preprocessed data sets (RMA in the case of Affymetrix arrays and directly from authors in case of other microarray platforms). Later, we normalized for N(0,1) followed by consensus-based NMF analysis to identify the number of classes. We used metagenes (defined above) to assign subtype identity to each cluster defined by NMF using additional data sets as described in our previous publication[4]. Furthermore, a heatmap was generated using NMF class and CRCassigner-786 genes. The data sets, their processed information and classification results for each data set have been provided in **Supplementary Fig. 3** and **Supplementary Table 2**.

Association of subtypes with TCGA subtypes and BRAF-mutant-like signature. We have compared our five subtypes with the three CRC gene expression subtypes recently reported by the TCGA[18]. We found that only 17% (171 out of 1020) of the subtype-defining signature genes overlap between the two studies. Indeed, when we audited the underlying expression data from the TCGA-subtyped tumors for the 171 overlapping genes using NMF analysis, we found evidence for our subtypes within their subtypes, and clearly demonstrate that each of the TCGA CRC subtypes can be further refined and subdivided (**Supplementary Fig. 3k,l**). In our view an explanation for subdivision of TCGA subtypes into our subtypes lies in the different methodologies employed. We identified our subtypes using algorithms (NMF, SAM and PAM, as described in **Supplementary Fig. 1a**) that were not employed in the TCGA CRC study. Moreover, the TCGA subtypes of CRC were not delineated with functional and therapeutic parameters, in contrast to their incorporation into our experimental design.

In addition, we also compared our CRC classification with a recent report of a “BRAF-mutant-like” gene signature[19] in CRC with poor prognosis using NTP algorithm. We found that the BRAF-mutant-like signature incorporates tumors from several of our subtypes, being significantly associated ($p<0.05$, chi-square test) with the goblet-like, inflammatory and stem-like subtypes, while the complimentary *BRAF* and *KRAS* double wild type signature is associated with the enterocyte and TA subtypes (**Supplementary Fig. 6d**).

Certainly the TCGA and the BRAF-mutant-like subtype classifiers may be informative about certain facets of CRC tumor biology. There is, however, reason to suggest that the distinctive subtype classification system we describe will have particular utility in guiding treatment decisions. The potential for therapeutic applicability of our subtype classification is based on the features of each subtype, which are distinctive in their putative cells-of-origin, in DFS, and in response to chemo- and targeted therapies.

Combination of CRC cell line data sets. We used DWD[10] to merge gene expression profile data sets for CRC cell lines from two different sources[20,21] for the purpose of increasing the total number of CRC cell lines. Prior to our analysis 14 repeated cell lines between the two data sets were removed. Overall, we obtained data for 51 unique CRC cell lines. The merged cell lines data set was later merged again with the CRC core data sets (after selecting silhouette positive samples from the CRC core data sets) using the DWD[10] based method. The merging of cell lines and core data sets was performed after selecting CRCassigner genes (786 SAM and PAM selected genes) from each data set. Next, we performed NMF based consensus clustering of the merged CRC cell lines and core data sets, seeking to identify subtypes amongst the cell lines (**Supplementary Fig. 4a,b**). We identified maximum cophentic coefficients at k=3 and k=5. We again selected k=5 for the reason explained above under "*Reason for choosing five over three CRC subtypes*". We determined that this collection of CRC cell lines represented only 4 subtypes: there was no single cell line that belonged to the enterocyte subtype. A few of the duplicate cell lines from different sources showed different subtype identity (probably due to variation in cell culture between different laboratories) after NMF consensus clustering. We tested the subtype of the SW620 cell line using RT-PCR analysis for markers of differentiation and stemness, since this cell line was used for various experiments. We found that SW620 had higher expression of stem cell markers and lower expression of differentiated marker, confirming its stem-like subtype identity (**Supplementary Fig. 4c**).

Intra- and inter-tumoral heterogeneity in CRC. Although intra-tumoral heterogeneity exists in CRC[22], our analyses using additional independent data sets containing samples from both microdissected and whole tumors, and from tumor RNAs profiled on different microarray platforms (**Supplementary Fig. 3** and **Supplementary Table 2),** consistently identified samples as having a particular CRC subtype. This result, in addition to the distinctive gene expression patterns in the five CRC subtypes reflective of different cell types in the colon-crypt (**Supplementary Fig. 2a-f),** suggests that individual tumors are dominated by cancer cells with characteristics of a particular subtype. This is similar to what has been suggested in breast cancer[23], where subtypes are routinely identified despite possible intra-tumoral heterogeneity.

Clinical/histopathologic analysis for the GSE14333 data set: We examined the relationship between DFS and other histopathological information such as Dukes' stage, age, location of tumors (left or right of colon or rectum) and adjuvant treatment in the GSE14333[10] data set; see **Supplementary Table 3**. We censored those patients who were alive without tumor recurrence or dead at last contact. In this data set, the median follow up among patients without events (tumor recurrence) was 45.1 months. As explained in the main text, we first evaluated DFS for all the samples irrespective of the type of the treatment (adjuvant chemotherapy or chemoradiotherapy), the specific components consisting the treatment (standard chemotherapy of either single agent 5-fluouracil (5-FU) or capecitabine or a combination of 5-FU and oxaliplatin) or Dukes' stage (for analysis, we considered Dukes' stage A and B patients with lymph node negativity together whereas Dukes' stage C patients with lymph node positivity separately). Dukes' stage is known to correlate with CRC survival[24]. We did not find a significant association between subtype and DFS (p=0.12; log-rank test; **Supplementary Fig. 5b and Supplementary Table 3**) in all the patient samples irrespective of stage or treatment. As previously known[24], we also observed in the current set of samples that treatment (p=0.03; log-rank test) and Dukes' stage (p=0.0009; log-rank test) were significantly associated with DFS. Similarly, we also observed that treatment was significantly associated with Dukes' stage ($p=1.98x10^{-14}$, Fisher's exact test). Since treatment and Dukes' stage were associated with DFS, we examined whether subtype was associated with DFS within subsets defined by these variables. In untreated patients, there was a significant association between subtypes and DFS (p=0.0003; n=120; log-rank test), with stem-like subtype tumors having the shortest DFS, and inflammatory and enterocyte subtypes having intermediate DFS (**Fig. 1e**). On the other hand, there was no significant association between subtype and DFS (p=0.9; n=77; log-rank test) in treated patients (**Supplementary Fig. 5c**). Similarly, we did not find significant association between subtype and DFS in Dukes' stages A and B (p=0.13; n=119; log-rank test) or in Dukes' stage C (p=0.7; n=98; log-rank test) patients. We also observed that treatment preferentially improved DFS in stem-like subtype patients (though not in a statistically significant manner, **Supplementary Fig. 5d**). This is because there is an interaction between subtypes and treatment. Typically, we would fit a Cox Propotional Hazard model for DFS that includes subtype, treatment, and their interaction. But due to the low number of events (n=43), this model did not converge. When we fit a Cox model of DFS on subtype alone for the untreated patients, the model again did not converge. Hence we cannot use the Cox model in our analysis. As mentioned earlier, we used a log-rank test to find the association between DFS and subtype in the untreated patients. To determine if this association changed in the presence of Dukes' stage in the untreated patients, we used a stratified log rank test, with Dukes' stage used as the stratification variable. In this model, the association was still significant with a p-value of 0.004. As such, subtype is associated with DFS even after adjusting for Dukes' stage in untreated patients. Since there were only 43 events of tumor recurrence amongst the treated and untreated samples (and only 15 in the stem-like subtype), additional patient samples will be needed to fully elucidate the relationships between subtype, treatment and DFS (at present, sufficiently large CRC sample microarray data sets annotated for kind of treatment, or none, are unavailable).

Clinical/histopathologic analysis for the Khambata-Ford data set: We identified only 3 CRC subtypes (goblet-like, TA and stem-like) in Khambata-Ford data set that had liver metastases samples from colorectal cancer patients. Here, we discuss more information that we obtained after analyzing the DFS from this data set. **Supplementary Fig. 7n-o** illustrate comparable differential responses to cetuximab treatment when restricting the analysis to the TA subtype ($p=1.4x10^{-6}$; log-rank test; n=26; **Supplementary Fig. 7n**) versus KRAS WT patients ($p=1.9x10^{-6}$; log-rank test; n=39; **Supplementary Fig. 7o**) using the Khambata-Ford[25] data set. For **Supplementary Figs. 7n-q** of these Kaplan-Meier plots, we excluded samples falling into the "unknown" subtype, which we suspect to have been contaminated by liver metastases, based on comparison to normal liver-specific gene expression signature (**Fig. 3a**). Survival statistics for responders (R), evaluated based on modified WHO criteria[26], were differentiated from non-responders (NR) using a log-rank test.

Supplementary Methods

Processing of microarrays. The processing of microarrays from CEL files was performed as described[4]. Published microarray data were obtained from GEO Omnibus[27]. The robust multiarray average (RMA)[9] preprocessing and normalization of raw CEL files from Affymetrix GeneChip® arrays were performed using R-based Bioconductor[28]. The patient characteristics for the published microarray data sets were obtained from GEO Omnibus using Bioconductor package, GEOquery[29].

Graphically illustrating gene expression profiles using hierarchical clustering. Median centering of genes and clustering of samples and/or genes from the microarray data sets were performed using Gene Cluster 3.0[30]. The clustering results were viewed using GenePattern based Hierarchical Clustering Viewer[31].

Survival statistics. Kaplan-Meier Survival curves were plotted and log-rank tests were performed using GenePattern based Survival Curve and Survival Difference programs[31]. Multivariate Cox Regression analysis was performed using R based survival package[32].

Patient samples. Tissue microarray (TMA; COC1021; n=120; only 53 of these samples were useful for analyses) slides were purchased from Pantomics (Hiddenhausen, Germany). RNA from colorectal cancer samples (n=19) was purchased from Origene (Rockville, MD, USA).

Cell lines. Colon cancer cell lines were grown in DMEM (Life Technologies, Grand Island, NY, USA) plus 10% FBS (Life Technologies) without antibiotics/antimycotics. All the cell lines were confirmed to be negative for mycoplamsa by PCR (VenorGeM kit, Sigma-Aldrich, St. Louis, MO, USA) prior to use. SW1116 was purchased from LGC Standards (France). HT29 cell line was gift from Dr. Renaud A. Du Pasquier (Centre Hospitalier Universitaire Vaudois; CHUV, Lausanne, Switzerland), SW480, SW48, HCT8, LS174T and SW948 cell lines were gift from Dr. Philippe Depeille (University of California at San Francisco, San Francisco, USA) and NCI-H508, LS1034, SW620, COLO320, SW1417, HCT116, RKO and DLD1 cell lines Dr. Haoqiang Ying (MD Anderson Medical Center, TX, USA).

Clonogenic assay. Single cells (10^4) were plated in 6-well dishes and treated with cetuximab (15.6 μg ml^{-1}) or media alone (untreated control) the following day. Once colonies with 30–50 cells were formed, they were counted at 50x power and pictures were taken using Leica ICC50 HD microscope (Leica Microsystems, Heerbrugg, Switzerland). The experiments were performed in duplicates and in the presence of serum. The number of clones from the cetuximab treated cells was normalized to the vehicle (media-alone) control.

Gene set enrichment analysis (GSEA). GSEA[33] was performed using javaGSEA Desktop Application using GenePattern software[31].

Xenograft Studies

Subcutaneous injection and drug pre-clinical trials. Swiss male nu/nu mice (5-6 weeks old, Charles-Rivers) were implanted with 10^6 or $2×10^6$ tumor cells mixed with an equal volume of Matrigel^{TM} (BD Biosiences) in a total volume of 100 µL on the flank of the animal. Mice were followed until tumors reached between 50 and 100 mm^3. The animals were then randomly assigned to treatment groups. Tumors were measured by calipers using the formula: length × ($width^2$)/2. The weight of animals was measured twice a week to detect toxicity. Cetuximab (Erbitux, Merck-Serono, Geneva, Switzerland) was administered by intraperitoneal (i.p.) injection every 3 days (q3D) at a dose of 1 mg/mouse for up to 5 consecutive injections (a modified protocol from Wild *et al.*[34]). For FOLFIRI treatment (which clinically consists of a combination of 5-FU, irinotecan and leukovorin), irinotecan (Actavis, Regensdorf-Zurich, Switzerland) was given at 20 mg/kg as a single i.p. injection at day 1. 5-FU (Sigma-Aldrich) was given at 100 mg/kg as an i.p. dose on days 1 and 2 followed by leucovorin (Sigma-Aldrich) at 40 mg/kg i.p. dose on days 1 and 2 as described[35]. Animals were followed until tumors reached a maximum of 1000 mm^3 in size or upto 35 days depending on the potential of the individual cell lines to grow as xenograft tumors. Growth curves and significance was defined using Prizm (GraphPad Software, La Jolla, CA, USA) and paired two-tailed student t-test, respectively. The cetuximab trials using CR-TA subtype cell line (LS1034) and CS-TA subtype cell line (NCI-H508) were repeated twice in our laboratory (we report only a representative experiment) and were also repeated at Charité, Universitätsmedizin Berlin, Berlin, Germany along with stem-like subtype cell line (HCT1116). In addition, we observed that cetuximab and vehicle treated tumors from HCT116 did not show significant difference in tumor volume (data not shown). All the animal procedures were performed after approval of protocol (authorization number 2263) and as per guidelines from Experience sur animaux (EXPANIM) – Service de la consommation et des Affaires veterinaries (SCAV) in Switzerland and as per institutional guidelines from Charité Universitätsmedizin Berlin, Germany and the experiments were performed after approval from the Berlin animal research authority LAGeSo (registration number G0068/10) in Germany

*Orthotopic implantation of CRC cell lines into mice and RNA isolation***.** NMRI nu/nu mice (6-8 week old females) were anesthetized with Ketamine and Xylazin, additionally receiving buprenorphin before surgery. The animals were placed on a heated operation table. A midline incision was performed and the descending colon was identified. A polyethylene catheter was inserted rectally and the descending colon was bedded extra-abdominally. To obtain a transplant tumor, human CRC cell lines ($2×10^6$ cells per site) were injected into the wall of the descending colon. Care was taken not to puncture the thin wall and inject the cells into the lumen of the colon. Presence of growing tumors at the site of injection was detected by colonoscopy or laparatomy 21 days after the initial surgery. The animals were sacrificed and tumors were explanted and immediately frozen in liquid nitrogen, and tumor samples were stored at -80°C. The animals were cared for per institutional guidelines from Charité Universitätsmedizin Berlin, Germany and the experiments were performed after approval from the Berlin animal research authority LAGeSo (registration number G0068/10).

RNA isolation and qRT-PCR. Snap-frozen tissue samples were embedded in Tissue-Tek® OCT^{TM} (Sakura, Alphen aan den Rijn, The Netherlands) and cut into 20 µm sections. Sections corresponding to 5-10 mg of tissue were collected in a microtube. RNA from these samples was prepared using the miRNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. RNA concentration and purity were determined using spectrophotometric measurement at 260 and 280 nm, and integrity of the RNA was evaluated using a total RNA nano/microfluidic cartridge on the Bioanalyzer 2100 (Agilent, Böblingen, Germany).

Total RNA from cell lines was extracted using miReasy kit (Qiagen, Hombrechtikon, Switzerland) and cDNA was generated using Superscript reverse transcriptase (Life Technologies, Grand Island, NY, USA). The concentration of cDNA was determined (Nanodrop 2000, Thermo Scientific, Asheville, NC, USA) and 25 ng of total cDNA was subjected to quantitative PCR using QIAgility (automated PCR setup, Qiagen), QuantiTect SYBR Green PCR kit (Qiagen), gene specific primers (**Supplementary Table 4**) and Rotor-Gene Q (Qiagen) real-time PCR machine. A standard curve ranging from 60 to 6×10^6 copies of linearized plasmid containing the target sequence was created for each gene. The gene-specific copy number was calculated according to the standard curve and normalized to the amount of cDNA (ng) in the reaction. The average (mean) copy number per ng of cDNA for each gene was calculated, and each sample was mean-centered for that particular gene.

Immunostaining of subtype-specific markers in cell lines. Colon cancer cell lines were plated, and allowed to set overnight, on gelatin-coated (0.1% solution in PBS) cover slides in 24-well dishes. The following day, the cells were fixed with 4% paraformaldehyde in PBS (20 minutes, room temperature) and washed twice. Immunofluorescent analysis was performed as described[36]. Antibody dilutions are as follows: MUC2 (1:100, SC7314; Santa Cruz, USA) and KRT20 (1:50, M7019; DAKO, USA. Slides were imaged using Olympus BX51 fluorescence microscope and images were taken at 400x power.

Immunohistochemistry. IHC for TMA was performed as described[37]. The following antibodies and dilutions were used: anti-MUC2 and anti-Zeb1 (1:100, Santa Cruz Biotechnology Inc., Santa Cruz, CA, USA), anti-TFF3 (1:100, R&D Systems, Minneapolis, MN, USA) and anti-CFTR (1:750, Abcam, Cambridge, MA, USA). Anti-mouse/anti-rabbit ImmPRESS^{TM} (Vector Laboratories, Burlingame, CA, USA) or horseradish peroxidase (HRP) conjugated goat anti-mouse antibodies (Jackson ImmunoResearch Laboratories Inc., Suffolk, UK) were used for visualization. Slides were imaged using Nikon Y-THF (Nikon AG, Egg/ZH, Switzerland) microscope and images were taken at 200x power.

IHC scoring system. The immunolabeled TMA sections were evaluated by a pathologist (B.L.). Each core was assessed semi-quantitatively, using a scoring scale based on both the extent and intensity of the stainings. For staining extent, scores 1, 2, 3, 4 and 5 corresponded to positivity in 1-5%, 6-25%, 26-50%, 51-75% and 76-100% of tumor cells, respectively. Scores +, ++ and +++ denoted weak, moderate and strong staining intensities, respectively; undetectable staining was scored 0. Only the intensity of the staining was used for subtype identification. Those samples that were missing or of low IHC staining quality were removed from the analysis. Out of 120 samples, only 53 were useful for the analyses. Those samples with intensity +++ for MUC2 and ++ or + for other proteins were considered as enterocyte CRC subtype, those with intensity +++ for MUC2 and TFF3 and ++ or + for other proteins were considered as globlet-like CRC subtype, those with intensity +++ for CFTR and ++ or + for other proteins was considered as TA CRC subtype and those with intensity +++ for ZEB1 and ++ or + for other proteins was considered as stem-like CRC subtype.

Legends

Supplementary Fig. 1. O*verview of methodology, consensus clusters and cophenetic coefficient plots from NMF for CRC core data sets and subtype determination.* **(a)** Summary of the methodology used to identify combined gene expression and drug response CRC subtypes. A detailed explanation of the methodology is provided above. **(b)** NMF consensus clustering analysis and cophenetic coefficient for cluster k=2 to k=5 of a DWD merged CRC core dataset (GSE13294 and GSE14333[38,39]). Maximum cophenetic coefficient occurred for k=3 or k=5. **(c)** GSE13294 and GSE14333 using different SD cutoffs. We filtered the genes using a SD cut-off of 0.8 individually from each of the core CRC data set, then merged the data sets using distance weighted discrimination (DWD) and performed NMF based consensus clustering on the gene sets. We performed similar analyses for SD cut-offs: 0.5 and 1. Using all the three SD cut-offs, we found consistent support for 3 to 5 subtypes. This demonstrates that the consensus support for 3 to 5 clusters is fairly insensitive to the SD threshold across the range of SD thresholds flanking SD=0.8. **(d)** Cophenetic coefficient plots from NMF based clustering of CRC core data sets (GSE13294 and GSE14333) where genes were selected with fold change greater than 2 in at least 3 samples followed by DWD based merging of data sets and NMF analysis, and found evidence for 3 to 5 subtypes, with the highest cophentic coefficient for k=5. (**e**) Heatmap showing 3 subtypes (k=3) from NMF consensus clustering of the CRC core data sets. Subtypes 1 and 2 each have 2 distinct signatures indicating heterogeneity in these subtypes. For this reason, and in accordance with the cophenetic coefficient, we chose 5 subtypes instead of 3 subtypes. (**f**) Silhouette plot for the DWD merged CRC core data sets showing samples from different subtypes and those with positive and negative silhouette score.

Supplementary Fig. 2 *Mapping the cellular phenotypes of each subtype and RT-PCR assays.* (**a**) Goblet specific markers (MUC2 and TFF3) show high median expression only in CRC goblet-like subtype; (**b**) enterocyte markers (CA1, CA2, KRT20, SLC26A3, AQP8 and MS4A12) show high median expression only in CRC enterocyte subtype; (**c**) Wnt target genes (SFRP2 and SFRP4), (**d**) myoepithelial genes (FN1 and TAGLN) and (**e**) epithelial-mesenchymal (EMT) markers (ZEB1, ZEB2, TWIST1 and SNAI2) show high median expression only in CRC stem-like subtype; and (**f**) chemokine and interferon-related genes (CXCL9, CXCL10, CXCL11, CXCL13, IFIT3) show high median expression only in CRC inflammatory subtype. The gene expression data are presented as the median of median-centered data from DWD merged CRC core microarray data sets. **(g)** qRT-PCR (positive, Pos; negative, neg) that classifies CRC patient samples into any of the five gene expression subtypes (see Supplementary Methods for the details of the assay). 'Pos' indicates positive expression value that is above the average (mean) copy number per ng of cDNA for that gene whereas 'Neg' indicates negative expression value that is below the mean copy number per ng of cDNA for that gene.

Supplementary Fig. 3 *Subtype validation in additional data sets.* (**a**) NMF consensus clustering analysis and cophenetic coefficient for cluster k=2 to k=5 and (**b**) heatmap with top bar representing subtypes for GSE12945 (microdissected tumor samples, n=62); (**c**) and (**d**) for GSE16125 (Affymetrix Human Exon 1.0 ST array, n=36); (**e**) and (**f**) for GSE20916 (whole tumor, n=101); (**g**) and (**h**) for GSE20842 (Whole tumor, Agilent-014850 Whole Human Genome Microarray 4x44k, n=65); (**i**) and (**j**) for GSE21510 (laser capture microdissection and whole tumor, n=123)[37] and (**k**) and (**l**) for TCGA (whole tumor data set, n=220)[12] CRC data sets. (**m**) Heatmap showing subtypes in GSE28722 (n=125) samples and their associated metastasis information. All data are presented after CRCassigner genes had been mapped on to the data sets individually. More information about each data set is available in **Supplementary Table 2**.

Supplementary Fig. 4 *Subtypes in CRC cell lines and subtype-specific gene expression in CRC xenograft tumors.* (**a**) NMF consensus clustering analysis and cophenetic coefficient for cluster k=2 to k=5 from combining CRC cell line data sets with the core primary tumor data sets; the maximum cophenetic coefficient occurred for k=5. However, CRC cell lines representing only 4 of the 5 subtypes were identified; no cell line for the enterocyte subtype was found. The cell lines data set is presented after CRCassigner genes had been mapped. (**b**) Heatmap showing CRC subtypes represented amongst a set of CRC cell lines as identified by merging core tumor data set and cell lines as in **Fig. 1b**. (**c**) Quantitative (q)RT-PCR analysis of SW1116 cell line using stem cell and differentiated markers. **d-e**) qRT-PCR analysis of xenograft tumors derived from the cell lines HCT116 (stem-like subtype), COLO205 (TA subtype) and HT29 (goblet-like subtype) for (**d**) differentiated and (**e**) stem cell markers. The expression is relative to the house-keeping gene, *RPL13A*. Error bars represent standard deviation (SD; technical triplicates).

Supplementary Fig. 5 *DFS comparison of CRC subtypes versus MSI/MSS.* Kaplan-Meier Survival curve depicting differential survival for data set GSE14333, which includes (**a**) untreated patients, (**b**) both treated (adjuvant chemotherapy or chemoradiation therapy) and untreated patients, (**c**) only treated patients. Kaplan-Meier Survival curve depicting differential survival for data set GSE14333, which includes treated and untreated patients only from (**d**) stem-like, (**e**) goblet-like, (**f**) TA, (**g**) enterocyte and (**h**) inflammatory subtypes. (**i**) Heatmap depicting known MSI or MSS status for each of the colorectal tumor subtype samples from the data set GSE13294. (**j**) Predicted MSI status for core data sets (GSE13294 and GSE14333) samples using publicly available gene signatures with the NTP algorithm. Predicted MSI status with FDR<0.2 or no FDR cutoff are shown. (**k**) Kaplan-Meier Survival curve depicting differential DFS for samples from data set GSE14333 that were predicted to be MSI or MSS.

Supplementary Fig. 6 *Gene enrichment analysis and differential Wnt target gene expression in two different sub-populations of TA subtype tumor samples, association of subtypes with BRAF-mutant-like signature.* (**a**) Heatmap showing the association of published stem cell signatures – the mRNA stem cell signature and intestinal stem cell (ISC) signature - with CRC subtypes. (**b**) gene set enrichment analysis (GSEA) analysis showing enrichment of published stem cell signatures and other pathways. **(c)** Bar graph showing median of median centered gene expression of the Wnt signaling targets LGR5 and ASCL2 in the core CRC microarray data for TA subtype tumors that are either predicted to be crypt top- or base-like. **(d)** Heatmap showing association of BRAF-mutant signature[40] with CRC subtypes (BRAF-mut indicate *BRAF*-mutant-like and WT2 – *BRAF* and *KRAS* double wild type).

Supplementary Fig. 7 *Cetuximab response and progression-free survival (PFS) in subtype-specific CRC tumors and cetuximab response for cell lines.* (**a**) NMF consensus clustering analysis and cophenetic coefficient for cluster k=2 to k=5 of Khambata-Ford data set. The data set is presented after PAM colorectal subtype-specific genes had been mapped. (**b**,**c**) Cetuximab response in cell lines from different CRC subtypes **(b)** proliferation assay (3.9 to 250 μg mL^{-1} of cetuximab) and **(c)** clonogenic assay (15.6 μg mL^{-1} of cetuximab). Data were normalized to vehicle-treatment (images were taken at 50x power). (**d**) Differential expression of AREG and EREG gene predictors between CR-TA and CS-TA, as measured by qRT-PCR analysis (data from Khambata-Ford, *et al*[25]). (**e**) qRT-PCR data showing fold change in FLNA expression. Gene expression was normalized to the house-keeping gene, RPL13A. The NCI-H508 is presented as a control. Receiver operating curve analysis for FLNA as a marker for cetuximab response (**f**) within TA samples only and (**g**) all the samples in Khambata-Ford data set. Kaplan-Meier Survival curve (Khambata-Ford data set) comparing FLNA expression in (**h**) only TA samples, (**i**) all samples, (**j**) KRAS wild-type samples or (**k**) KRAS mutant samples. qRT-PCR (Positive expression value (Pos) represents that above the average (mean) copy number per ng of cDNA and negative expression value (Neg) represents that below the mean copy number per ng of cDNA for that gene) that classifies (**l**) CRC patient samples into any of the five gene expression subtypes and (**m**) only TA subtype samples from (**l**) into CR-TA and CS-TA sub-subtypes using cystic fibrosis transmembrane conductance regulator (CFTR) and FLNA expression. Kaplan-Meier Survival curve for patients (Khambata-Ford data set) that belong to cetuximab responders (R) and non-responders (NR) based on: (**n**) only TA subtype samples; (**o**) only KRAS wild type samples; (**p**) all samples except those from the TA subtype and unknown (liver contamination); and (**q**) all samples except those that were unknown. Responders to cetuximab include those patients with complete or partial response or stable disease whereas non-responders include those with progressive disease. A more detailed explanation about the survival analysis and the results are available in the **Supplementary Information**.

Supplementary Fig. 8 *Subtype-specific FOLFIRI response.* (**a**) NMF consensus clustering analysis and cophenetic coefficient for cluster k=2 to k=10 from DWD combined Del Rio data set (with FOLFIRI response from patients) and the core primary tumor data sets; the maximum cophenetic coefficient occurred for k=5. Heatmap showing subtypes in (**b**) Del Rio data set with individual responses of primary CRC patients (Del Rio data set, n=21) to FOLFIRI treatment and their association with subtypes and (**c**) DWD combined Del Rio data set (with FOLFIRI response from patients) and the core primary tumor data sets. Association of response to FOLFIRI in individual patient samples from the data sets (**d**) DWD combined GSE14333 and GSE13294, (**e**) only GSE14333 or (**f**) only GSE13294 by applying specific signatures using the NTP algorithm.

Supplementary Table 1. *Results from SAM and PAM analysis, the list of genes associated with each subtype and qRT-PCR and IHC assays.*

Supplementary Table 2. *Summary of gene expression profile data sets used*. Subtype identity for each sample from all the data sets used including the cell lines and their associations with different gene signatures. The information regarding GEO Omnibus ID, the nature of samples, processing methods and the PubMed reference numbers are provided.

Supplementary Table 3. *Clinical/histopathological, subtype and statistical information for GSE14333 samples.*

Supplementary Table 4. *Khambata-Ford data set liver genes and PCR primers.* List of genes from “Unknown” subtype of Khambata-Ford data set that we identified to contain the liver-specific genes. List of primers used and their sequence and annealing temperatures.

References

1. Verhaak, R.G.*, et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell* **17**, 98-110 (2009).
2. Markert, E.K., Mizuno, H., Vazquez, A. & Levine, A.J. Molecular classification of prostate cancer using curated expression signatures. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 21276-21281 (2011).
3. Perou, C.M.*, et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).
4. Collisson, E.A.*, et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine* **17**, 500-503 (2011).
5. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).

6. Alizadeh, A.A., *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).

7. Tothill, R.W., *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* **14**, 5198-5208 (2008).

8. Dalerba, P., *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature biotechnology* **29**, 1120-1127 (2011).

9. Irizarry, R.A., *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* **4**, 249-264 (2003).

10. Benito, M., *et al.* Adjustment of systematic microarray data biases. *Bioinformatics (Oxford, England)* **20**, 105-114 (2004).

11. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4164-4169 (2004).

12. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65 (1987).

13. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121 (2001).

14. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567-6572 (2002).

15. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999).

16. Kosinski, C., *et al.* Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 15418-15423 (2007).

17. Hoshida, Y. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PloS one* **5**, e15543 (2010).

18. TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337 (2012).

19. Popovici, V., *et al.* Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J Clin Oncol* **30**, 1288-1295 (2012).

20. Barretina, J., *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).

21. Greshock, J., *et al.* Molecular target class is predictive of in vitro response profile. *Cancer research* **70**, 3677-3686 (2010).

22. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767 (1990).

23. Polyak, K. Heterogeneity in breast cancer. *The Journal of clinical investigation* **121**, 3786-3788 (2011).

24. Van Cutsem, E. & Oliveira, J. Primary colon cancer: ESMO clinical recommendations for diagnosis, adjuvant treatment and follow-up. *Ann Oncol* **20 Suppl 4**, 49-50 (2009).

25. Khambata-Ford, S.*, et al.* Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J Clin Oncol* **25**, 3230-3237 (2007).
26. Miller, A.B., Hoogstraten, B., Staquet, M. & Winkler, A. Reporting results of cancer treatment. *Cancer* **47**, 207-214 (1981).
27. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210 (2002).
28. Gentleman, R.C.*, et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).
29. Sean, D. & Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)* **23**, 1846-1847 (2007).
30. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863-14868 (1998).
31. Reich, M.*, et al.* GenePattern 2.0. *Nature genetics* **38**, 500-501 (2006).
32. Therneau, T. A package for survival analysis in S. R package version 2.36-14. (2012).
33. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).
34. Wild, R.*, et al.* Cetuximab preclinical antitumor activity (monotherapy and combination based) is not predicted by relative total or activated epidermal growth factor receptor tumor expression levels. *Molecular cancer therapeutics* **5**, 104-113 (2006).
35. Kim, H.S.*, et al.* Dendritic cell vaccine in addition to FOLFIRI regimen improve antitumor effects through the inhibition of immunosuppressive cells in murine colorectal cancer model. *Vaccine* **28**, 7787-7796 (2010).
36. Lyssiotis, C.A.*, et al.* Inhibition of histone deacetylase activity induces developmental plasticity in oligodendrocyte precursor cells. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 14982-14987 (2007).
37. Sadanandam, A.*, et al.* High gene expression of semaphorin 5A in pancreatic cancer is associated with tumor growth, invasion and metastasis. *International journal of cancer* **127**, 1373-1383 (2010).
38. 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).
39. Jorissen, R.N.*, et al.* DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* **14**, 8061-8069 (2008).