A Composite Gene Expression Signature Optimizes Prediction of Colorectal Cancer Metastasis and Outcome

Michael J. Schell1, Mingli Yang2, Edoardo Missiaglia3, Mauro Delorenzi3,4,5, Charlotte Soneson3, Binglin Yue1, Michael V. Nebozhyn6, Andrey Loboda6, Gregory Bloom1, and Timothy J. Yeatman2

Abstract

Purpose: We previously found that an epithelial-to-mesenchymal transition (EMT)–based gene expression signature was highly correlated with the first principal component (PC1) of 326 colorectal cancer tumors and was prognostic. This study was designed to improve these signatures for better prediction of metastasis and outcome.

Experimental Design: A total of 468 colorectal cancer tumors including all stages (I–IV) and metastatic lesions were used to develop a new prognostic score (ΔPC1.EMT) by subtracting the EMT signature score from its correlated PC1 signature score. The score was validated on six other independent datasets with a total of 3,697 tumors.

Results: ΔPC1.EMT was found to be far more predictive of metastasis and outcome than its parent scores. It performed well in stages I to III, among microsatellite instability subtypes, and across multiple mutation-based subclasses, demonstrating a refined capacity to predict distant metastatic potential even in tumors with a "good" prognosis. For example, in the PETACC-3 clinical trial dataset, it predicted worse overall survival in an adjusted multivariable model for stage III patients (HR standardized by interquartile range [IQR] = 1.50; 95% confidence interval, 1.25–1.81; P = 0.000016, N = 644). The improved performance of ΔPC1.EMT was related to its propensity to identify epithelial-like subpopulations as well as mesenchymal-like subpopulations. Biologically, the signature was correlated positively with RAS signaling but negatively with mitochondrial metabolism. ΔPC1.EMT was a "best of assessed" prognostic score when compared with 10 other known prognostic signatures.

Conclusions: The study developed a prognostic signature score with a propensity to detect non-EMT features, including epithelial cancer stem cell–related properties, thereby improving its potential to predict metastasis and poorer outcome in stage I–III patients.

Introduction

The heterogeneity of colorectal cancer makes it difficult to determine which patients will benefit from adjuvant therapy and which patients do not require further therapy beyond surgical resection. Thus, there is an urgent need for objective molecular classification to stratify adjuvant therapy for colorectal cancer patients (1–3). One major challenge is the identification of factors to evaluate the potential of distant metastasis that has contributed to most of colorectal cancer mortality. Metastasis is a complex series of steps, including the tumor cell invasion and dissemination, survival in circulation, organ-specific targeting, tumor dormancy, and reactivation for colonization at distant sites (4, 5). The epithelial-to-mesenchymal transition (EMT) has been intensely studied in various types of cancers (especially breast cancer) as a major mechanism promoting invasion and metastasis (5, 6). EMT has also been reported as one of the mechanisms contributing to the resistance to cetuximab (anti-EGFR) therapy (7). However, the biology for most steps of metastasis, especially after the early steps of invasion and dissemination, is still poorly understood (4, 5). This has greatly restricted our ability to understand and predict metastatic potential in cancer patients and has led to generalized “one size fits all” approaches to the administration of adjuvant therapy.

We have previously shown that EMT gene expression signatures can predict poor outcome in colorectal cancer and breast cancer (3, 8, 9). In an unsupervised analysis, our past work yielded a list of top-ranked genes bearing positive and negative correlation with the first principal component (PC1) of colorectal cancer expression dataset of 326 tumors (8). Of many signatures tested, our "EMT signature", derived from a gene expression analysis of 93 lung cancer cell lines sorted (based on their expression of CDH1 or VIM) into epithelial or mesenchymal groups, showed a very strong correlation (Pearson R = 0.92; P < 10⁻¹³⁵) with PC1. This
Translational Relevance

The heterogeneity of colorectal cancer results in an urgent need for objective molecular classification to stratify adjuvant therapy for colorectal cancer patients. One major challenge is to evaluate the potential of distant metastasis, which has contributed the most to colorectal cancer mortality. The new prognostic signature score developed and validated in this study was shown to be a “best of assessed” prognostic signature score demonstrating a refined capacity to predict metastasis and outcomes in colorectal cancer, with a non-EMT propensity including epithelial cancer stem cell–related properties. The finding has clinical utility in determining which patients will have metastases and which will not, in otherwise good and poor prognosis lesions (all stages, microsatellite instability, and within mutational subgroups). The signature may have the potential to identify which patients may or may not benefit from adjuvant chemotherapy—a problem for which there is no current solution.

PC1 and EMT association was confirmed in 38 colorectal cancer cell lines, and was also verified by assessment of other known EMT-related genes and miRNAs in colorectal cancer tumors (8). Both PC1 and EMT signatures were found to predict recurrence (indicating metastasis; ref. 8).

To further assess the respective prognostic values of the PC1 and EMT signatures, we recently evaluated the outcomes on a new set of 468 colorectal cancer tumors (Moffitt468). The improvement in prognostic power noted in a bivariable survival model, when the two signatures were put in competition with each other, prompted us to generate a composite signature (ΔPC1.EMT) by subtracting EMT from PC1. Consequently, ΔPC1.EMT emerges as a new prognostic score for colorectal cancer prognosis, which could predominantly capture the non-EMT biologic features to optimize prediction of metastasis and outcome.

Materials and Methods

Tumor samples

The cohort of 468 colorectal adenocarcinoma patients (Moffitt468 dataset) from 468 distinct patients, including 367 primary lesions (306 stage I–III and 61 stage IV) and 101 metastatic lesions (49 from stage IV patients), with global gene expression analysis data from the surgical specimen, microsatellite instability (MSI) status, and targeted gene sequencing (Supplementary Table S1), with samples obtained between October 2006 and September 2010, was used to develop the “difference score” ΔPC1.EMT. Metastatic samples were included only for patients for whom primary samples were not sequenced. ΔPC1.EMT was then validated on 1,544 independent primary and metastatic tumors. In all cases, tissue and clinical data were collected on patients under institutional review board approval as part of the Total Cancer Care (TCC) project (10).

We assessed/selected five additional large, independent colorectal cancer datasets from public resources (GEO and ArrayExpress) for cohorts of colorectal cancer patients with more than 100 samples, gene expression profile as well as relevant clinical information (including stage and follow-up) to be used to validate prognostic value and to determine biological significance. These include PETACC3, ALMAC, LNCC, GEO41258, and GSE14333 (refs. 3, 11–14; Supplementary Table S2). Notably, PETACC3 was selected because it is one of the largest gene expression profile set derived from stage II & III patients recruited in a single clinical trial, while other datasets were retrospective collections of patients. Moreover, the TCGA adenocarcinoma dataset (15) was also used for biologic interpretation.

ΔPC1.EMT score computation

Probe intensities were preprocessed using RMA. PC1 and EMT scores were calculated as previously described (8). Briefly, for each of the datasets, a score was computed for each of the 4 signatures (EMT.UP.score, EMT.DOWN.score, PC1.UP.score and PC1.DOWN.score) as the arithmetic mean of all probesets corresponding to gene symbols present in the corresponding gene signature (Supplementary Table S3). EMT and PC1 scores were then obtained as follows:

EMT.score = EMT.UP.score − EMT.DOWN.score.

PC1.score = PC1.UP.score − PC1.DOWN.score.

The ΔPC1.EMT score was computed as follows:

ΔPC1.EMT score = PC1.score − EMT.score.

Scores were standardized by subtracting the score median and dividing by the score IQR. For Moffitt 468 dataset, the score median (interquartile range) for the PC1, EMT, and ΔPC1.EMT scores are −0.29 (−0.42 to −0.18), −0.38 (−0.55 to 0.21), and −0.09 (−0.17 to 0.01), respectively.

Correlation analysis

Pearson product moment correlation coefficient was used to quantify the association between the scores, MSI status, and mutation status for various driver genes.

GO Process analysis

Pathway analysis of the nonoverlapped genes of PC1 (i.e., PC1 genelist minus PC1 and EMT overlapped gene list) by GO Process was performed using the MetaCore package. A P value cut-off of 0.05 unadjusted for multiplicity resulted in 35 significant dysregulated pathways.

Hierarchical cluster analysis

We performed hierarchical clustering of five datasets (PETACC3, ALMAC, LNCC, GEO41258, and GSE14333) to visualize how the genes included in the signature grouped across different cohorts and platforms. To do that across all the datasets, we had to collapse the expression at the gene level selecting the probeset which showed the higher variability (measured by median absolute deviation). We also tested the association of each gene to overall survival (OS) and relapse-free survival (RFS) using a meta-analytical approach.

Weight contributions of individual signature genes

To characterize the three signatures (PC1, EMT, and ΔPC1.EMT), we estimated the average contribution of each gene to each of the signatures across the five datasets. Within each dataset, we first calculated a weight for each probe set in the PC1 and EMT signatures, respectively. The weight was defined as 1/P⁺ for probe sets with positive weight and −1/P⁻ for probe sets with negative weight in the signature. Here, P⁺ and P⁻ are the total number of probe sets with positive and negative weights, respectively.
The contribution of a probe set to a signature in a given dataset was then defined as the product between the weight of the probe set and its average expression level across the dataset. By summing contributions for all probe sets corresponding to a given gene, we estimated gene-wise contributions to each signature. The contributions to the ΔPC1.EMT signature were obtained as the difference between the contributions to the PC1 and the EMT signatures. The final estimates of gene contributions to the three signatures were obtained as weighted averages of the gene contributions across all five datasets to obtain final estimates of the gene contributions to the three signatures. The weight for a dataset in this sum was inversely proportional to the Euclidean norm of the vector of gene contributions to the PC1 and EMT signatures in
Association of gene expression with ΔPC1.EMT score

We tested the association of gene expression with the ΔPC1.EMT score within each of the five datasets plus TCGA colorectal cancer dataset (15) by a linear regression model with the score as the explanatory variable using the "limma" R package (version 3.16.3; ref. 16), adjusting SEs estimates by an empirical Bayes approach. P values were combined across datasets using Fisher method (MADAM R package version 1.2.2). A Bonferroni correction was applied to control for false positive results introduced by multiple testing. Genes showing an adjusted P value <0.0001 were split in two groups: those positively (N = 2,983) and those negatively (N = 2,221) correlated with the ΔPC1.EMT score.

Gene set enrichment analysis (GSEA) was performed to interpret the list of genes found to be correlating with ΔPC1.EMT score. The functional tool DAVID (http://david.abcc.ncifcrf.gov/) was employed to identify annotation terms enriched within each of the groups. We also performed GSEA using gene sets obtained from the MSig database (DB; ref. 12; MSigDB) which includes C2 (curated gene sets - Chemical and Genetic Perturbations, Biocarta and KEGG), C3 (transcription factors), C5 (GO biological process terms), C6 (oncogenic signature), and C7 (immunologic signatures). The analysis was done using the "Romer" algorithm (similar to GSEA; ref. 12) and the same linear model used to identify genes correlating with ΔPC1.EMT score. The P values obtained across the datasets were merged using Fisher method.

Survival analysis

We performed the Kaplan–Meier survival analysis on the Mofitt468 dataset and used Cox proportional hazards regression models in the R package "survival" (version 2.37-7) to assess association of tumor scores with OS, RFS, and/or survival after relapse (SAR) on the other five datasets.

Univariate analysis (OS and RFS) of 10 other known prognostic signatures

We selected a set of gene signatures known to be prognostic in colorectal cancer and could be computed from gene expression profiles. We computed the scores from 10 signatures (RAS Merck [17], RAS Astrazeneca [18], OncotypedX colon [19], VerideX [20], MD Anderson [21], Decorin [9], MED12 [22], BRAF score [23], and ALM [12]) on the five datasets as described in the original studies.

Results

ΔPC1.EMT outperformed both PC1 and EMT in predicting metastasis and survival

Our analysis of the Mofitt468 dataset showed that PC1 and EMT were highly correlated (Fig. 1A, top, Pearson R = 0.90, P < 0.0001). The EMT score can be used to separate tumors with epithelial (<0) versus mesenchymal (>0) features (8). The majority of metastatic tumors (who have poor overall survival) appear to be epithelial-like (EMT scores < 0, Fig. 1A, top). Notably, in the PC1 versus EMT plot, tumors from metastatic patients or stage IV primaries (with synchronous metastasis; open and filled red cycles) appeared to cluster above the blue regression line (Stage I-III primary tumors), suggesting that metastatic tumors were more associated with PC1 than EMT.

Consistent with this figure, survival analysis using the univariate Cox proportional hazard regression model for overall survival (OS) on Moffitt468 indicates that the PC1 score was predictive of OS [HR, 1.40; 95% confidence interval (CI), 1.18–1.66, P = 0.0001], while the EMT score fell short of statistical significance (HR, 1.13; 95% CI, 0.96–1.34, P = 0.14). Interestingly, when the scores were used in a multivariable Cox survival model, the coefficients (logarithms of the HR) for PC1 and EMT were both highly significant, but of roughly equal magnitude and opposite numeric sign [i.e., for PC1, HR, 3.75 (worse survival); 95% CI, 2.51–5.61, P < 0.0001; for EMT, HR, 0.36 (better survival); 95% CI, 0.24–0.53, P < 0.0001, with log HRs = 1.32 and −1.02]. The statistical interpretation of this result is that survival is best explained not by PC1 or EMT alone but by a score obtained by combining them into a new score to which the PC1 score contributes positively and the EMT score negatively, with roughly equal magnitudes. Thus, we elected to subtract the EMT score from the PC1 score to produce a "difference" score (ΔPC1.EMT). Subsequent univariate OS analysis indeed demonstrated that ΔPC1.EMT (HR, 1.82; 95% CI, 1.51–2.18, P < 0.0001) clearly outperformed not only EMT, but also PC1, which is supported by a significantly stronger association with metastatic tumors (Fig. 1A, middle and bottom). The ΔPC1.EMT score had a good association with EMT (Pearson R = 0.38, P < 0.0001), but displayed an even stronger correlation with PC1 (Pearson R = 0.74, P < 0.0001), suggesting that PC1 includes a non-EMT biologic component (presuming that the EMT score captures the EMT component fairly completely). In support of this notion, higher ΔPC1.EMT scores better separate the metastatic and nonmetastatic tumor tissues, most of which have EMT score < 0 indicating epithelial-like tumors (Fig. 1A, bottom, highlighted by red box). Moreover, it was clear that PC1, and especially ΔPC1.EMT, outperformed EMT in progressively deciphering the degree of tumor progression of primary colorectal cancers (from increasing primary stage to metastatic lesions (Fig. 1B)).

Furthermore, the Kaplan–Meier survival analysis shows that a higher ΔPC1.EMT score could better predict poorer OS for all patients (log-rank trend, P < 0.0001, Fig. 2A, left) than PC1 (log-rank trend, P = 0.0006) and EMT (log-rank trend, P = 0.1571; Supplementary Figs. S1A and S2A). Notably, ΔPC1.
EMT predicted poorer OS for MSS ($P < 0.0001$) and tended toward statistical significance for MSI ($P = 0.085$) patients (Supplementary Fig. S3A). Moreover, when limited to the 306 stage I–III primary tumors, ΔPC1.EMT clearly outperformed its parental scores ($P = 0.0005$ for ΔPC1.EMT, Fig. 2A, right, as compared with $P = 0.1437$ for PC1 and $P = 0.3313$ for EMT, Supplementary Figs. S1B and S2B). In contrast, for metastatic tumors, like its parental scores (Supplementary Figs. S1C and S2C), ΔPC1.EMT did not predict poorer OS (Supplementary Fig. S3B).

Validation of ΔPC1.EMT’s prognostic value

The prognostic value of ΔPC1.EMT was also tested and confirmed by a univariate Cox regression analysis in a Moffitt dataset with 1,544 independent cases, showing that ΔPC1.EMT robustly predicted worse OS (HR, 1.49; 95% CI, 1.36–1.64, $P = 2.2 \times 10^{-16}$), or when restricted to 981 stage I–III primary tumors (HR, 1.43; 95% CI, 1.26–1.63, $P = 3.6 \times 10^{-8}$).

These findings were validated when ΔPC1.EMT was further tested for OS, relapse-free survival (RFS), and survival after relapse (SAR) on the PETACC3 dataset ($n = 752$; ref. 3; Table 1A). As
observed on Moffitt468, ΔPC1.EMT outperformed both PC1 and EMT scores. For instance, in a univariate model for OS with the stage III patients (n = 644, Table 1B), ΔPC1.EMT had the most significant P value of the three signatures, with an HR of 1.69 (P = 8.22 × 10^{-5}) compared with that of 1.41 (P = 5.13 × 10^{-5}) and 1.28 (P = 8.21 × 10^{-5}) for PC1 and EMT, respectively. In the multivariable modeling including PC1 and EMT on the same dataset (Table 1C), the HR for PC1 was 3.22 while the HR for EMT was 0.37 (coefficients 1.17 and -0.99). The statistical meaning is that also in this cohort a contrast of these two scores is significantly better than either of them alone, and quantitatively similarly as in the Moffitt data the simple difference (coefficients -1 and -1) is close to the optimally fitting combination.

The validation was then expanded to include additional independent datasets (n = 1,401 colorectal cancer tumors from the other 4 datasets; Supplementary Table S2) along with various clinicopathologic and molecular variables, including age, T and N stages, number of examined lymph nodes, tumor site (left and right), MSI status, BRAF mutation, EMT score, and/or KRAS mutation. Generally, in univariate models, ΔPC1.EMT outperformed PC1, which performed better than EMT (this ordering held in 11 of 15 models; Fig. 2B; Supplementary Table S4). Furthermore, the independent prognostic value of ΔPC1.EMT was confirmed in 3 of 5 datasets when analyzed in multivariate models including other clinicopathologic and molecular parameters, such as MSI, BRAF, and/or KRAS mutations (Fig. 2C; Supplementary Fig. S4; Supplementary Table S5). The signature performance was further verified by additional analyses, as shown by the survival versus score curves (Supplementary Figs. S5 and S6) as well as the observed versus predicted survival probability curves (Supplementary Figs. S7 and S8).

In addition, in agreement with the univariate results, overall, ΔPC1.EMT significantly outperformed both EMT and PC1 scores in multivariate OS and RFS analyses when they were compared with each other by individually (Table 1D, Supplementary Tables S8 [ΔPC1.EMT vs. PC1] and S9 [ΔPC1.EMT vs. EMT]). For example, when compared individually for OS on PETACC stage III (n = 642), HR (95% CI) = 1.50 (1.25–1.81), P = 1.61 × 10^{-5} (ΔPC1.EMT) versus 1.32 (1.11–1.58), P = 2.12 × 10^{-5} (PC1) versus 1.21 (1.00–1.46), P = 4.97 × 10^{-2} (EMT), whereas for RFS on the same dataset, HR (95% CI) = 1.41 (1.20–1.65), P = 2.84 × 10^{-5} (ΔPC1.EMT) versus 1.28 (1.09–1.49), P = 1.92 × 10^{-3} (PC1) versus 1.19 (1.01–1.40), P = 4.04 × 10^{-2} (EMT).

It is noteworthy that currently only few colorectal cancer datasets exist and are accessible where survival and expression profiles having >100 patients. For example, GEO41258 is our smallest dataset in which we could not also find correlation with survival, neither in univariate nor in multivariate models, for well-known variables such as MSI status, tumor side, T-stage, and stage. This may suggest that this population is not representative of colorectal cancer patients. However, we decided to include it for an unbiased report of our results.

ΔPC1.EMT identified metastatic tumors with non-EMT features

To explore the molecular basis for the observed prognostic improvement of ΔPC1.EMT from its parent PC1 and EMT scores, we examined quartile trends of these three scores versus the number of tumors harboring observed mutations of several known "driver" genes on Moffitt468, as this may provide insights into the mechanisms underlying the signature. The ΔPC1.EMT score had stronger trends (relative to PC1 and EMT) with tumors harboring APC-truncated mutations (negative) and BRAF (V600E) mutations (positive), as well as tumors identified as MSI-H (positive) or stage IV (positive; Fig. 3A; for stage I–III patients, see Supplementary Fig. S9). Notably, while percentage of distant metastatic tumors overall increased across the quartiles for all three scores, for some subgroups of combined mutations (KRAS and TP53 or BRAF and TP53), as well as in MSI-H and stage I cases, the positive trend was more pronounced for ΔPC1.EMT in contrast to the negative trend for EMT (Fig. 3B). Further supporting the notion that ΔPC1.EMT might be measuring non-EMT components of metastasis.

The ΔPC1.EMT score was found to be associated with several clinicopathologic and molecular variables using the The Cancer Genome Atlas (TCGA) dataset (ref. 15; Supplementary Fig. S10), with BRAF mutation, MSI status, and mucinous tumors showing the strongest positive associations (P < 0.001). It is noteworthy that for the Moffitt468 data, MSI was positively correlated with ΔPC1.EMT, but uncorrelated with PC1 and negatively correlated with EMT (Supplementary Table S10).

Hierarchical clustering and contribution analyses of the signature genes

To better understand the molecular underpinnings of ΔPC1.EMT, gene expression clustering analysis was performed on the five datasets. Data show areas of strong overlap of PC1 and EMT genes, especially in the middle of the OS and RFS heatmaps, accounting for their high correlation, but also show isolated, nonoverlapping genes (Supplementary Fig. S11), providing the potential for ΔPC1.EMT to improve outcome. Notably, the high correlation between these two signatures was shown in Supplementary Fig. S12. As the contributions of VM (a mesenchymal gene used to create the EMT signature) and other overlapped genes were effectively diminished in ΔPC1.EMT, we suspected that ΔPC1.EMT might better measure non-EMT features of colorectal cancer. An analysis of the GO Process of those nonoverlapping genes indicates that a number of the pathways were related to cell adhesion and cellular remodeling, which are frequently associated with metastasis (Supplementary Table S11). To further address this issue, we analyzed respective weighted contributions of individual signature genes on the five datasets to identify the genes whose contributions changed the most from PC1 or EMT to ΔPC1.EMT (Fig. 3C). ΔPC1.EMT was represented by more epithelial and less mesenchymal gene contributions as evidenced by the increased contribution of the epithelial marker CDH1, whereas the mesenchymal marker VM and other EMT-related genes including SPARC, TCF4, COL1A2, and COL3A1 decreased.

Identification of ΔPC1.EMT-correlated genes and pathways

To further explore the biologic implication of ΔPC1.EMT, we performed another association analysis and identified a list of top-ranked genes whose expression was either positively or negatively correlated with ΔPC1.EMT (Table 2) in a linear model on the five datasets plus the TCGA colorectal cancer dataset (15). Many of the identified genes have been reported to have biological functions related to metastasis and cancer stem cell-like properties, as discussed below. Notably, 13 of 20 of them belong to PC1 and/or EMT signature genes. To interpret
the biologic meaning of identified ΔPC1.EMT-correlated genes, we also carried out extensive GSEA and identified a variety of biologic processes correlated with ΔPC1.EMT, including negatively correlated mitochondrial metabolism (Supplementary Tables S12 and S13).

Comparison of ΔPC1.EMT with other known prognostic signatures
Finally, we compared the ΔPC1.EMT score with an expanded set of other known prognostic signatures on the five datasets in a univariate analysis. Results show that overall, ΔPC1.EMT was
among the best prognostic signature scores for OS and RFS analyses when compared with 10 other known prognostic signatures across eight comparisons, with a higher HR more often associated with a “good” prognosis (Fig. 3). Our first evidence using human tissues reveals that although EMT is a dominant molecular program of colorectal cancer (30), although the related biology is not clear yet. It is of interest to mention that ΔPC1.EMT showed a partial correlation with the OncotypeDX colon signature (GH1) which had exploited cell proliferation as a potential prognostic marker (ref. 19; Supplementary Table S14).

Discussion

Here we present the first evidence using human tissues revealing that although EMT is a dominant molecular program of colorectal cancer (8), the non-EMT features captured by ΔPC1.EMT appear to be necessary to optimally predict distant metastasis. In support of this notion, the ΔPC1.EMT score demonstrated a strong non-EMT signature propensity in predicting distant metastasis (Fig. 1). It also displayed a refined capacity to detect non–EMT-related metastatic potential in tumors harboring subgroups of combined mutations (KRAS and TP53 or BRAF and TP53) with abnormal RAS activation as well as in MSI-H and stage II cases generally classified with a “good” prognosis (Fig. 3). Our findings are in agreement with the recent notion that the epithelial phenotype may be critical for the successful seeding and propagation of cancer cells at distant sites (4, 24–29). For instance, from a clinicopathologic point of view, cohesive epithelial migration was often observed as the predominant pattern in colorectal cancer (30), although the related biology is not clear yet.

The result of analyzing contributions of the signature genes sheds light on the molecular underpinnings of ΔPC1.EMT. Compared with EMT, the gene with the greatest contribution increase in ΔPC1.EMT was CD24 (Fig. 3C), previously reported as a metastasis-associated gene (31), and a marker of colon cancer stem cells (CSC) whose properties are thought to contribute to metastatic traits and therapeutic resistance (5, 32). Thus, ΔPC1.EMT captures both epithelial and CSC features, which are supported by a recent report demonstrating that in breast cancer, CDH1 and CD24 were highly enriched in the epithelial CSCs (ALDH1-positive), while their expression was downregulated in the mesenchymal CSCs (CD44+CD24−; ref. 33). ERBB3, a member of the EGF family (34), was also identified as one of the genes whose contribution was increased in ΔPC1.EMT (Fig. 3C). In agreement with this, we observed that ΔPC1.EMT, but not EMT, was associated with activation of the RAS/MAPK pathway, evidenced by its positive correlation with various RAS signature scores (Supplementary Table S10). Thus, we speculated that ΔPC1.EMT-associated poor prognosis might, in part, reflect RAS/MAPK activation-mediated drug resistance (34) in epithelial-like colorectal cancer.

Moreover, many of identified most strongly ΔPC1.EMT-correlated genes (Table 2) have been reported to relate to metastasis and/or CSCs. For instance, CD109 (the top positively correlated gene) has recently identified by proteomic analyses as a metastasis-associated protein marker (35) and was highly expressed in ALDH1-characterized epithelioid sarcoma CSCs (36). Meanwhile, CDX1 and CDX2 (the two most negatively correlated genes) were reported as putative tumor suppressor genes whose expression was epigenetically repressed in colorectal cancer, and reduced expression of CDX1 inhibited CSC stem cell differentiation and thus promoted CSC renewal (35). In support of this, HCT116, an epithelial, MSI colorectal cancer cell line that lacks expression of CDX1 was recently classified as a colon CSC cell line (2). In addition, reduced expression of EPHB2 was associated with metastasis (37) while its overexpression induced EMT (38), whereas the cell-cycle gene MYB, when ectopically expressed, contributed to cell migration and invasion but to also prevent metastasis (39). Thus, identification of EPHB2 and MYB as strong negatively correlated genes of ΔPC1.EMT further supports the

Table 2. Genes most correlated with ΔPC1.EMT score

| Gene symbol | EntrezID | S | num.p | P_adj | Sum t statistics | Signature genes |
|-------------|---------|---|-------|-------|-----------------|-----------------
| CD109       | 1044    | 860.61 | 6     | <0.0001 | −80.16          | PC1 Down        |
| CD2X        | 1045    | 845.27 | 6     | <0.0001 | −79.41          | PC1 Down        |
| CD10sr99    | 387695  | 767.33 | 5     | <0.0001 | −67.82          | PC1 Down        |
| DDC         | 1644    | 752.19 | 6     | <0.0001 | −73.57          | PC1 Down        |
| GPA33       | 10223   | 726.29 | 6     | <0.0001 | −72.98          | PC1 Down        |
| FAM84A      | 151354  | 720.55 | 5     | <0.0001 | −72.98          | PC1 Down        |
| NR12        | 8856    | 697.98 | 6     | <0.0001 | −70.24          | PC1 Down        |
| MYB         | 4602    | 630.56 | 6     | <0.0001 | −68.13          | PC1 Down        |
| C2orf89     | 129293  | 616.89 | 5     | <0.0001 | −60.62          | PC1 and EMT Down|
| EPBHB2      | 2048    | 597.82 | 6     | <0.0001 | −66.42          | PC1 Down        |

Identified in a linear model on the six datasets.

P is a meta-analytic statistic to combine P values across datasets.

num.p denotes the number of P values used.

The genes correlated with ΔPC1.EMT that are overlapped with the PC1 and EMT signature genes (also see Supplementary Table S3).
Figure 4.
Univariate analysis (OS and RFS) of ΔPC1.EMT and 10 other known prognostic signatures on five datasets (PETACC3, ALMAC, LNCC, GEO41258, and GSE14333). We computed the scores from 10 signatures [RAS Merck, RAS Astrazeneca, OncotypeDX colon, Veridex, MD Anderson (MDA), Decorin (DCN), EMT, MED12, BRAF score and ALM] on the five datasets as described in the original studies. ΔPC1.EMT is colored in red, whereas signatures showing relatively higher HR are colored in blue. Note, the solid lines represent 95% CI and prognostic signature scores were standardized by IQR.
notion of non-EMT contributions to metastasis. In agreement with this, the ΔPC1.EMT-correlated GH prognostic signature was negatively correlated with cell-cycle genes such as MYBL2 (19), although how GH may be potentially related to metastasis is unknown.

In further support of the negative association of ΔPC1.EMT with cell proliferation, pathway analyses show that ΔPC1.EMT was clearly correlated with negative regulation of mitochondrial metabolism (Supplementary Tables S12 and S13). It is noteworthy that the metastasis suppressor gene KISS1 was recently reported to promote normal mitochondrial metabolism, an anti-metastasis mechanism (40). Moreover, it has recently been reported that the mitochondrial pyruvate carrier (MPC) played a repressor role in the Warburg effect in colorectal cancer and results indicated that inhibition of mitochondrial metabolism was connected to the maintenance and fate of cancer stem cells (41).

In the "cell cooperativity" model (27), Tsuji and colleagues demonstrated that primary tumors were heterogeneous and contained both mesenchymal and epithelial cell types (with mesenchymal cells populating the invasive front), but metastatic tumors contained only the cells originating from the epithelial type. This model postulates that the canonical epithelial-to-mesenchymal transition (EMT) does not fully explain metastatic potential, and should have strong epithelial features. Accordingly, the ΔPC1.EMT score reported here captures predominantly non-EMT features. Although ΔPC1.EMT is certainly distinct from EMT, it still retains some correlation with it (Supplementary Table S10 and Supplementary Fig. S12). Indeed, ΔPC1.EMT was positively correlated with the EMT-related pathways associated with response to wounding, cell motility, extracellular matrix remodeling, activation of TGFβ signaling, and angiogenesis (Supplementary Tables S12 and S13), as well as three important EMT-related pathways centered around SLUG1 (Supplementary Table S11). It is noteworthy that SLUG1 was reported to cooperate with SOX9 to convert differentiated mammary epithelial cells to stem cells (42) and stromal gene expression (EMT-related) was also recently reported to define poor prognosis subtypes in colorectal cancer (43, 44). The reason for a significant EMT correlation for ΔPC1.EMT is not yet clear. According to the recent notion of epithelial plasticity (25), EMT is not a "black and white" program in human cancers, and there likely exist a variety of "gray" EMT states in most tumors especially colorectal cancer, which may be a part of the intrinsic heterogeneous nature of the disease. PC1 and EMT scores, which have quite different lists of signature genes, may differ in their abilities to measure the degree of "gray." Thus, the "EMT" components might be only partially canceled out by subtracting the EMT score from the PC1 score, resulting in the significant "residual" correlation between ΔPC1.EMT and EMT.

The results of this study are compelling and suggest that ΔPC1.EMT may be strongly predictive of adverse outcomes (metastasis and diminished survival), which should help determine which patients may need adjuvant chemotherapy in stage II/III disease (45). However, few high quality datasets exist where molecular data have been collected with clinical data in patients with identified adjuvant chemotherapy history. Probably for this reason, we observed that the ΔPC1.EMT score was found significantly correlated with survival in the majority but not all the test sets. This also happened for 10 other known prognostic signatures when tested on the same datasets (Fig. 4); but overall, ΔPC1.EMT appeared to be a "best of assessed" prognostic score with an "optimized" capacity to predict metastasis. However, excluding the PETACC dataset, the analyzed cohorts were derived from retrospective collections of patients, hampering the generalization of our findings. Therefore, there is a clear need for further investigation of ΔPC1.EMT in a prospective clinical trial to determine its prognostic value in predicting which patients will metastasize and thus possibly benefit from adjuvant chemotherapy.

In conclusion, our findings suggest that poor RFS and OS can be predicted by a robust gene expression signature, ΔPC1.EMT, preferentially based on non-EMT stem cell biology. ΔPC1.EMT had a refined capacity to detect poorer overall, and in various subgroups of colorectal cancer using preferentially non-EMT features (including epithelial cancer stem cell-related properties), thereby potentially providing new targets for therapy of distant disease. The score may have utility in identifying stage II and III patients with high risk of metastasis. Thus, we believe that there should be considerable enthusiasm about further examination of this signature in a prospective clinical trial.

Disclosure of Potential Conflicts of Interest

M. Delorenzi has ownership interest (including patents) in Novartis and Roche. No potential conflicts of interest were disclosed by the other authors.

Disclaimer

G. Bloom was not available to confirm coauthorship, but the corresponding author, T.J. Yeatman, affirms that G. Bloom contributed to the article and thus confirms his coauthorship status.

Authors’ Contributions

Conception and design: M.J. Schell, M. Yang, M. Delorenzi, T.J. Yeatman
Development of methodology: M.J. Schell, M. Yang, E. Missiaglia
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Yang, M. Delorenzi
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M.J. Schell, M. Yang, E. Missiaglia, M. Delorenzi, C. Sonesson, B. Yue, M.V. Nebzrhyn, A. Loboda, G. Bloom
Writing, review, and/or revision of the manuscript: M.J. Schell, M. Yang, E. Missiaglia, M. Delorenzi, C. Sonesson, M.V. Nebzrhyn
Study supervision: M.J. Schell, M. Delorenzi, T.J. Yeatman

Acknowledgments

Total Cancer Care is enabled, in part, by the support of the DeBartolo Family, and the authors thank the patients who provided data and tissue to the Total Cancer Care Consortium. The authors also thank the Fondation Medic for financial support.

Grant Support

This work was supported by NIH grant U01CA157960 (to T.J. Yeatman). This study also received assistance from the Biostatistics and Cancer Informatics Core Facilities at the H. Lee Moffitt Cancer Center & Research Institute, an NCI-designated Comprehensive Cancer Center, supported under NIH grant P30-CA76292. This work was also supported by a grant of the Swiss State Secretariat for Education, Research and Innovation to the SIB Swiss Institute of Bioinformatics (group M.D.) for Service and infrastructure resources.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 17, 2015; revised June 8, 2015; accepted June 30, 2015; published OnlineFirst October 7, 2015.

www.aacrjournals.org

Clin Cancer Res; 2015

Published OnlineFirst October 7, 2015; DOI: 10.1158/1078-0432.CCR-15-0143

Downloaded from clincancerres.aacrjournals.org on May 1, 2017. © 2015 American Association for Cancer Research.
References


A Composite Gene Expression Signature Optimizes Prediction of Colorectal Cancer Metastasis and Outcome

Michael J. Schell, Mingli Yang, Edoardo Missiaglia, et al.

Clin Cancer Res  Published OnlineFirst October 7, 2015.

Updated version  Access the most recent version of this article at:  doi:10.1158/1078-0432.CCR-15-0143

Supplementary Material  Access the most recent supplemental material at:  http://clincancerres.aacrjournals.org/content/suppl/2015/10/07/1078-0432.CCR-15-0143.DC1

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.