

Multi-Cancer Computational Analysis Reveals Metastasis-Associated Variant of Desmoplastic Reaction Involving INHBA and THBS2

Hoon Kim¹, John Watkinson¹, Vinay Varadan² and Dimitris Anastassiou^{1*}

¹ Center for Computational Biology and Bioinformatics and Department of Electrical Engineering, Columbia University, New York, NY, USA, ² Philips Research North America, Briarcliff Manor, NY, USA

* Email: anastas@ee.columbia.edu

Abstract

Despite extensive research, the details of the metastasis-associated biological mechanisms are largely unknown. Here, we analyze data from multiple cancers using a novel computational method identifying sets of genes whose coordinated overexpression indicates the presence of a particular phenotype. We conclude that there is one shared “core” metastasis-associated gene expression signature corresponding to a specific variant of desmoplastic reaction, present in a large subset of samples that have exceeded a threshold of invasive transition specific to each cancer, indicating that the biological mechanism is triggered at that point. For example this threshold is reached at stage IIIc in ovarian cancer and at stage II in colorectal cancer. It has several features, such as coordinated expression of particular collagens, mainly *COL11A1* and other genes, mainly *THBS2* and *INHBA*. The universally prominent presence of *INHBA* in all cancers strongly suggests a biological mechanism centered on activin A induced TGF- β signaling, because activin A is a member of the TGF- β superfamily consisting of an *INHBA* homodimer. It is accompanied by the expression of several transcription factors related to epithelial-mesenchymal transition, but not of *SNAI1*, and expression of E-cadherin is not downregulated. It is reversible, as evidenced by its absence in many matched metastasized samples, but its presence indicates that metastasis has occurred. Therefore, these results can be used for developing high-specificity biomarkers, as well as potential multi-cancer metastasis-inhibiting therapeutics targeting the corresponding biological mechanism.

Introduction

It has been hypothesized [1,2] that the activation of some form of epithelial-mesenchymal transition (EMT) is a critical mechanism for the acquisition of motility and invasiveness in epithelial cancers. The version of EMT associated with cancer progression and metastasis has recently been labeled “Type 3 EMT” [2]. The details of the mechanisms have remained unclear, but it is believed that it involves activated fibroblasts in the desmoplastic stroma of tumors, referred to as “carcinoma associated fibroblasts” (CAFs), activation of some type of TGF- β signaling and significant involvement of the tumor microenvironment [3]. There is currently great interest in identifying the precise metastasis-associated features of type 3 EMT.

A study [4] of serous papillary ovarian carcinomas, comparing the gene expression profiles of primary vs. omental metastatic tumors, identified 156 differentially expressed genes. To investigate the significance of these genes in an independent rich data set we performed hierarchical clustering, using only these genes, on The Cancer Genome Atlas (TCGA) gene expression data set consisting of 377 ovarian cancer samples containing staging information. The resulting heat map revealed a prominent block of about 100 highly overexpressed genes in 94 samples (Figure S1, shown in www.ee.columbia.edu/~anastas/Figure_S1.pdf). Remarkably, we found that none of the 41 samples from tumors of stages IIIb and below were among the 94 samples. This cannot be due to chance ($P = 4 \times 10^{-6}$), leading to the hypothesis that coordinated overexpression of these genes implies that the tumour has metastasized into at least stage IIIc.

To further validate this hypothesis and test if similar versions apply to other cancers, we developed a computational technique (see Materials and Methods), which identifies, in an unbiased manner, clusters of coordinately overexpressed genes associated with a phenotype (such as transition to a particular metastatic stage). Our results consistently rediscover the same “core” signature of overexpressed genes, confirming the hypothesis. We found that this phenomenon occurs in multiple cancers, each of which has its own features involving additional genes, but the core signature is common. This common signature points to one particular variant of metastasis-associated desmoplastic reaction and to a corresponding particular variant of carcinoma associated fibroblasts to which we refer as “metastasis associated fibroblasts” (MAFs). Accordingly, in the following we refer to the corresponding gene expression signature and biological mechanism as “the MAF signature” and “the MAF mechanism,” respectively.

Results

Since we focus on the cluster of genes associated with the metastasis binary (“low stage” versus “high stage”) phenotype when the genes have their extreme (in most cases, largest) values, but not otherwise, we developed a special measure of association between the gene and the phenotype, which we call “extreme value association” (EVA), described in Materials and Methods. We performed the above algorithm on four rich gene expression data sets, two from ovarian cancer, the one from TCGA and another one [5], and two from colorectal cancer [6,7] accompanied by staging information. Using various staging transitions, it became clear that, in all cases, the MAF threshold can only be defined as exceeding stage IIIb in each of the ovarian data sets and stage I in each of the colorectal data sets. Interestingly, the “metastasis-associated

genes” identified in [4] as present in omental metastasis of ovarian cancer were also largely identified in [5] as belonging to a subtype of ovarian cancer characterized by extensive desmoplasia, which contains the MAF signature.

Remarkably, we found that there were many genes, each of which had Bonferroni-corrected $P < 10^{-3}$ in *all* four data sets. Table 1 shows a list of these genes with average log fold change greater than 2. The top ranked gene was *COL11A1* (probe 37892_at), followed by *COL10A1*, *POSTN*, *ASPN*, *THBS2*, and *FAP*. Again, these genes were found purely as a result of their association with the staging phenotype in all four cancers. Gene Ontology enrichment testing of these genes identified cell adhesion, extracellular matrix and collagen fibril organization.

We then did an extensive literature search aimed at identifying other studies in which at least some of these genes were identified as differentially expressed in various stages of other cancers. We even scrutinized studies in which none of the genes were mentioned in the main text, by looking at their supplementary data and re-ranking particular columns of genes in terms of their fold changes, from genes containing numerous genes. Although most of our results were negative, we were able to produce cancer gene lists with striking similarity (Table 2) to our own list (Table 1) in three studies of breast [8], gastric [9] and pancreatic [10] cancer.

Specifically, a breast cancer study [8] comparing ductal carcinomas in situ (DCIS) with invasive ductal carcinoma (IDC) had a list of genes upregulated in IDC (Table 2) that had similarities to those we had identified, and the top-ranked gene was again *COL11A1* (probe 37892_at) with log fold change of 6.50, while the next highest (4.08) corresponded to another probe of *COL11A1*, followed by a probe of *COL10A1*. Second, a study [9] comparing early gastric cancer (EGC) with advanced gastric cancer (AGC) – roughly separating stages I and II – also identified a similar differentially expressed gene list (Table 2) of which again *COL11A1* (probe 37892_at) was at the top (log fold change: 4.26) followed by *COL10A1* and *FAP*. Third, a study of pancreatic ductal adenocarcinoma [10] identified a list (Table 2) of gene overexpressed in whole tumor tissue versus normal pancreatic tissue, in which *COL11A1* (probe 37892_at) is again prominent and the top entry (log fold change 5.15) was *INHBA*, supportive of our hypothesis of activin induced TGF- β signaling. The presence of the MAF signature in the latter study indicates that pancreatic cancer had reached the MAF staging threshold in most cases before the biopsy. The prominent desmoplastic reaction in pancreatic cancers (which contains the MAF signature) has recently been increasingly recognized as a “foe” [11] that could lead to new therapeutic strategies targeting stromal cells to inhibit cancer. Finally, we realized that *COL11A1* has been identified as a potential metastasis-associated gene in other types of cancer as well, such as in lung [12], and oral cavity [13], suggesting that the MAF signature may be present in a subset of high stage samples of most if not all epithelial cancers.

In those cases as well as in our own findings, there was prominent presence of *COL11A1* (probe 37892_at). This remarkable consistent strong association of *COL11A1* with the staging phenotype (specific to each cancer type) suggests that it could be used as a “proxy” of the MAF signature. This, in turn, allowed us to improve on the gene list of Table 1 by making use of numerous publicly available gene expression data sets of cancers of many types, even without any staging information, as long as the MAF signature is present in a sizeable subset of them, aiming at finding the “intersection” of the associated factors in these sets, revealing the “core” of the MAF biological mechanism.

As a first step for this task, we identified the few genes that are consistently highest associated with *COL11A1*. Table 3 shows a listing of genes in 22 cancer data sets, as well as an aggregate list of genes ranked in terms of their association with *COL11A1*. The list is very similar to the phenotype-based gene ranking (Table 1). In addition to *COL10A1* and a few other collagens, the top ranked genes are thrombospondin-2 (*THBS2*), inhibin beta A (*INHBA*), fibroblast activation protein (*FAP*), leucine rich repeat containing 15 (*LRRC15*), periostin (*POSTN*), and a disintegrin and metalloproteinase domain-containing protein 12 (*ADAM12*). The presence of *FAP* indicates a general desmoplastic reaction and is not, by itself, sufficient for inferring the MAF signature. Indeed, *FAP* is occasionally co-expressed with several other EMT-related genes even in healthy tissues. However, *COL11A1* was *not* associated with any of these genes in neither healthy nor low-stage cancerous tissues, further supporting the hypothesis that it can be used as a proxy for the MAF signature. These results indicate that *THBS2* and *INHBA*, top ranked in Table 3 except for collagens, are the most important players in the MAF mechanism.

As a second step, we identified gene pairs that are highest associated with *COL11A1* jointly, but not individually, and therefore they would not appear in the previous list. For this task we ranked gene pairs according to their synergy [14] with *COL11A1*, using the computational method in [15], which could further facilitate biological discovery. For example, the scatter plots in Figure 1 show that genes *ECM2* and *TCF21* are jointly, but not individually, strongly associated with *COL11A1* ($P < 10^{-6}$, see Materials and Methods) in the two ovarian cancer data sets. Such findings are useful for developing biological hypotheses, e.g. in this particular case they suggest that the extracellular matrix protein 2 is associated with the MAF signature only when the *TCF21* gene (a known mesenchymal-epithelial transition mediator) is downregulated.

We only had miRNA and methylation data available for the TCGA ovarian data set. Using as measure the mutual information with *COL11A1*, we found many statistically significant miRNAs, among them hsa-miR-22 and hsa-miR-152, as well as differentially methylated genes, such as *SNAI1* and *PRAME*, suggesting a particularly complex biological mechanism (correlation with the MAF phenotype led to essentially the same lists with lower significance). Table 4 contains a list of the miRNAs, while Table 5 contains a list of the methylated genes (multiple test corrected $P < 10^{-16}$ in both cases, see Materials and Methods). *SNAI1* (*snail*) methylation is particularly important as the gene is known as one of the most important EMT-related transcription factors. Instead, the strongest MAF-associated transcription factor is *AEBP1*, making it a particularly interesting potential target. Many of the other EMT-related transcription factors, such as *SNAI2*, *TWIST1*, and *ZEB1* are often overexpressed in the MAF signature, but *SNAI1* is not (and, at least in ovarian carcinoma in which we have methylation data, this is due to its differentially methylated status). We believe that the lack of *SNAI1* expression is a key distinguishing feature of the MAF signature, in which we observed neither *SNAI1* overexpression nor *CDH1* (E-cadherin) downregulation.

Discussion

A direct clinical application of these findings is the development of a high-specificity metastasis-sensing biomarker product detecting coordinated overexpression of a few top-ranked genes, such as *THBS2*, *INHBA* and the collagens *COL11A1*, *COL10A1*. A positive result in

seemingly low-stage primary tumors will indicate that the disease has actually already reached a higher stage.

Remarkably, the same product can also be used to predict drug response. Indeed, at least in breast cancer, the MAF signature is associated with resistance to neoadjuvant chemotherapy. This is demonstrated in [16] where a stromal “metagene” signature of 50 genes was defined based on *DCN* (decorin). Although some of our key genes (such as *COL11A1*, *THBS2*) were not among these 50, the metagene signature used in that study has a strong intersection with the MAF signature. The stromal signature was resistant to neoadjuvant chemotherapy.

Of course, the most significant clinical application would be to develop metastasis-inhibiting therapeutics using targets deduced from the biological knowledge provided by the MAF signature. Our top ranked genes strongly suggest that a key feature of the MAF signature is fibroblast activation based on an activin A induced version of TGF- β signaling resulting in partial EMT lacking *SNAIL* expression, and leading to some form of altered proteolysis [17], which results in an environment rich in particular collagens, mainly *COL11A1* and *COL10A1*. Supporting this hypothesis are the facts that activin A (INHBA homodimer) is a TGF- β superfamily member (ligand), *THBS2* inhibits activation of TGF- β by *THBS1*, *POSTN* is highly homologous to the TGF- β induced gene *TGFBI*, and *ADAM12* is known to have both protease activity and to contribute to TGF- β signaling [18]. The role of gene *LRRC15* (aka *LIB*) appears important but unclear, though it has already been recognized as promoting migration through the extracellular matrix [19]. Other related genes often present in the MAF signature that appear to be significant players in the mechanism are tissue inhibitor of metalloproteinases-3 (*TIMP3*), stromelysin-3 (*MMP11*), and cadherin-11 (*CDH11*). Overexpression of *INHBA* has been known to occur in cancers, occasionally accompanied by concomitant overexpression of the activin A receptors [20]. In one of the ovarian cancer data sets [5] we found that the MAF signature was accompanied ($P < 10^{-5}$) by concomitant overexpression of genes *ACVRL1* aka *ALK1* (Activin A type “II-like” 1 receptor), *ACVR1* aka *ALK2* (Activin A Type 1 receptor), and *ACVR2A* (Activin A Type 2A receptor). *ACVRL1* also consistently appears overexpressed and associated with *INHBA* expression in all cancers, suggesting that it plays a key role in the corresponding signaling mechanism. Remarkably, activin A is already known to facilitate fibroblast-mediated collagen gel contraction [21].

Although each of the MAF signature molecules could serve as a potential therapeutic target, alone or in combination, including miRNAs and methylated genes such as *SNAIL*, the hypothesis that activin A induced TGF- β signaling is at the heart of the MAF mechanism immediately suggests that follistatin (activin-binding protein) could serve as a metastasis inhibitor, which is exactly what recent research [22] indicates. Specifically, lung cancer cell lines transfected with follistatin and injected intravenously into immunodeficient mice markedly inhibited metastasis compared with non-transfected cell lines, but the authors of the study recognize that the role of follistatin in cancer metastasis is totally unknown [23]. Our work provides an explanation and suggests that the same could be true for other cancers as well. Further support is provided by the fact that follistatin virtually abolished the fibroblast-mediated collagen gel contraction mentioned earlier [21].

There are several reasons that the core MAF signature has not yet been discovered as a multi-cancer metastasis-associated signature. First, it is essential to define a precise phenotypic threshold recognizing that the signature only exists in a subset of tumors that exceed a particular

stage. Indeed, if the threshold in breast cancer was put between stage I and stage II, or between stage II and stage III, rather than between *in situ* and stage I, the signature would not be apparent, or it could even be reversed (see below). Second, each cancer type has its own additional features accompanying the MAF signature. For example, in ovarian cancer it is accompanied by sharp downregulation of genes *COLEC11*, *PEG3* and *TSPAN8*, which is not the case in other cancers. Indeed, the main contribution of our work is the identification of the common multi-cancer “core” signature, from which a universal metastasis-associated biological mechanism can be identified. Third and most importantly, the MAF signature is sharply reversible through a mesenchymal-epithelial transition (MET). For example [24], in a comparison of metastatic lymph node samples with their corresponding matched primary breast cancer samples, it was found that *COL11A1* had a much higher expression in the *primary* tumor samples. Such reverse results can be particularly confusing.

In fact, there are occasions in which the presence of the MAF signature in high-stage tumor samples (not having yet being reversed) can be an indicator of better prognosis, because many of the top-ranked genes in the MAF signature (such as thrombospondins, decorin, INHBA itself) are known to be potent anti-angiogenesis mediators. The reversal of the MAF signature would thus facilitate angiogenesis and further metastatic dissemination to distant sites. In other words, (a) the desmoplastic MAF signature and (b) angiogenesis, are two independent biological events. The former appears to be based on activin A – induced TGF- β signaling (note that the same proteins mentioned above: thrombospondins, decorin, INHBA, etc, are also known inhibitors of the “standard” TGF- β ligand such as TGF β 1. So, the reversal of the MAF signature would allow the standard ligand to take over in TGF beta signaling, and may thus facilitate angiogenesis. These observations provide explanations for the seemingly contradictory observed roles of TGF- β signaling in cancer and metastasis.

The reversibility of the MAF signature leads to the intriguing hypothesis that it is part of a dynamic process and perhaps all metastases have, at some point temporarily been there, which explains why we only observe it in a subset of them. It has already been recognized that “it is plausible, though hardly proven, that all types of carcinoma cells must undergo a partial or complete EMT to become motile and invasive [25] p. 600.” This would be particularly exciting, because any metastasis-inhibiting therapeutic intervention targeting the MAF mechanism would be widely applicable to low-stage tumors.

In conclusion, we have shown that, using computational analysis of publicly available biological information, systems biology has revealed the core of a multi-cancer metastasis-associated gene expression signature. In the near future, a vast amount of additional information will become available, including next generation sequencing, miRNA and methylation information for many cancers, which will allow additional computational research building on this work and clarifying the details of the underlying complex biological process.

Materials and Methods

The gene expression data sets used in this paper are described in Table 6.

Extreme Value Association (EVA)

The EVA metric is the minimum P value of biased partitions over all subsets of samples with highest expression values of the gene. In other words, suppose that there are totally M samples, out of which N are “low stage” and $M - N$ are “high stage,” and we select the m samples with the highest gene expression values. Under the assumption that gene expression values are uncorrelated with the phenotype, the probability that there will be at most n “low stage” samples among the selected m samples is given by the cumulative hypergeometric probability $h(x \leq n; M, N, m)$. The EVA metric is then equal to $-\log_{10}$ of the minimum of these probabilities over all possible values of n . For example, assume that there are 250 high-stage samples and 50 low-stage sample for a total of 300 samples. Furthermore, assume that the 100 samples with the highest values of a particular gene contain 99 high-stage samples and one low-stage sample. In that case, $h(x \leq 1; 300, 50, 100)$ can be evaluated using the MATLAB function `hypercdf(1,300,50,100) = 5 \times 10^{-9}, resulting in the EVA metric for that gene of at least $-\log_{10}(5 \times 10^{-9}) = 8.3$, e.g. if the 101th sample is also high-stage, then the EVA metric of the gene will be even higher. Note that, once the highest value is reached, the sorting arrangement of the remaining samples is irrelevant, reflecting the hypothesis that only the extreme values are associated with the phenotype. Figure 2 shows the values of the cumulative hypergeometric probability for the COL11A1 gene using the TCGA ovarian cancer data set and the staging threshold between IIIb and IIIc: The maximum (8.31) occurs when $m = 133$. In fact, all 133 samples with the highest COL11A1 expression are at stage IIIc or IV.`

We then developed a mechanistic unbiased (only dependent on the phenotype) algorithm, which, when given a gene expression data set for a number of samples labelled “high stage” or “low stage,” leads to a selection of genes that are coordinately overexpressed only in high-stage samples. We first select the top 100 genes that rank highest according to the EVA metric criterion. Using this set of genes only, we perform k-means clustering with gap statistic [26]. At that step, if indeed the genes are coordinately overexpressed, they will align well in the heat map. This leads to the selection of the samples belonging to the cluster most associated with the high/low stage phenotype – call this the set of “EVA-based samples.” Nearly all samples in that cluster have exceeded the MAF staging threshold, and the very few exceptions could be due to misdiagnosis. Next, we define a “clean” MAF phenotype, contrasting the samples that are: (a) both “EVA-based” and “high-stage” against (b) the samples that are both “non EVA-based” and “low stage.” If the number of samples is sufficiently large, this “clean” phenotype provides the sharpest way by which we can identify the genes that are most associated with the observed phenomenon of metastasis-associated coordinated overexpression. We then rank the genes and compute their multiple-test-corrected P values using a heteroscedastic t-test using the “clean” phenotype and select the genes for which $P < 10^{-3}$ after Bonferroni correction. Finally, we find the intersection of these selected gene sets over all cancer expression data sets and rank them in terms of fold change.

For a data set with n samples and m probe sets, The EVA algorithm computes $n \times m$ cumulative hypergeometric distribution probabilities. This can be quite computationally intensive, so we

devised a low-complexity implementation algorithm to dynamically “build” the cumulative hypergeometric distribution for each probe set as the EVA algorithm progresses, as follows:

Given a data set with a high-stage samples and b -low stage samples, the idea is to construct an $(a + 1) \times (b + 1)$ table of the hypergeometric probabilities corresponding to all possible subsets of the samples. Then, for each probe set, the samples are sorted according to the expression value of the probe set. This ordering results in a path through the table from the bottom left corner to the top right corner, moving either up or to the right for each sample. At each step in the path, the cumulative probability of encountering the observed number of high stage samples or more is computed by summing the entries diagonally down and to the right of the current cell, including the current cell itself. The algorithm is best demonstrated with a visual example shown in Figure 3, in which the data set has three low stage samples and five high stage samples in total. Each probe set results in a path through this table, and an example path is displayed here in gray. Letting 1 correspond to a high stage sample and 0 correspond to a low stage sample, this example probe set results in the path 111001011 . For the cell in blue, corresponding to the sub-path 111001 , the probability of encountering this many high stage samples or more is computed by summing the three probabilities diagonally down and to the right of the blue cell (including itself). In this case, the probability is quite high (82.2%). This cumulative probability is computed for every step along the path, and the minimum of these is the output of the EVA algorithm.

The pseudo-code for this algorithm is given below:

Input:

Let n be the number of samples.
 Let a be the number of high stage samples.
 Let b be the number of low stage samples.
 Let s be the array of phenotype labels sorted by this probe set's expression level
 (Note: $a + b = n$)

Algorithm:

```
// Step 1 - Build the table of hypergeometric probabilities.
// This step need only be run once for the entire data set.
Define  $c$  as an array with  $(a+1)$  rows and  $(b+1)$  columns.
For  $x$  from 0 to  $a$ 
  For  $y$  from 0 to  $b$ 
    If  $x = 0$  and  $y = 0$ 
       $c[x][y] = 0$ 
    Else
      If  $(y > 0)$ 
         $c[x][y] = c[x][y] + c[x][y - 1] * (b - y + 1) / ((b - y + 1) + (a - x))$ 
      End if
      If  $(x > 0)$ 
         $c[x][y] = c[x][y] + c[x - 1][y] * (a - x + 1) / ((a - x + 1) + (b - y))$ 
      End If
    End If
  End For
End For
// Step 2 - Compute the cumulative hypergeometric probability for the given sequence.
Define  $x = 0$ 
Define  $y = 0$ 
Define  $bestP = 1$ 
For  $i$  from 1 to  $n$ 
  If  $(s[i] = 1)$ 
     $x = x + 1$ 
  Else
     $y = y + 1$ 
  End If
End For
```



```

End If
Define p = 0
For j from 0 to y
  If (x + y - j <= a)
    p = p + c[x + y - j][j]
  End If
End For
If (p < bestP)
  bestP = p
End If
End For

```

Output:

$-\log_{10}(\text{bestP})$

Mutual Information and Synergy

Assuming that two variables, such as the expression levels of two genes G_1 and G_2 are governed by a joint probability density p_{12} with corresponding marginals p_1 and p_2 and using simplified notation, the mutual information $I(G_1; G_2)$ is a general measure of correlation and is defined as the expected value $E\left\{\log\frac{p_{12}}{p_1p_2}\right\}$. The synergy of two variables G_1, G_2 with respect to a third variable G_3 is [14] equal to $I(G_1, G_2; G_3) - [I(G_1; G_3) + I(G_2; G_3)]$, i.e., the part of the association of the pair G_1, G_2 with G_3 that is purely due to a synergistic cooperation between G_1 and G_2 (the “whole” minus the sum of the “parts”).

P* value evaluation for the significance of miRNA and methylation sites, and for synergistic pair *ECM2* and *TCF21

We applied a permutation-based approach accounting for multiple test correction: We did 100 permutation experiments of the class labels, saving the corresponding 100 highest values after doing exhaustive search in each permutation experiment. Using the set of these 100 highest-value scores, we obtained the maximum likelihood estimates of the location parameter and the scale parameter of the Gumbel (type-I extreme value) distribution, resulting in a cumulative density function F . The P value of an actual score x_0 is then $1-F(x_0)$ under the null hypothesis of no association with phenotype. Similarly, for the synergistic pair, we found the top-scoring synergy in 100 data sets that were identical to the original except that the COL11A1 probe values were randomly permuted on each, and the top permuted synergy scores were modelled, as above, with the Gumbel distribution.

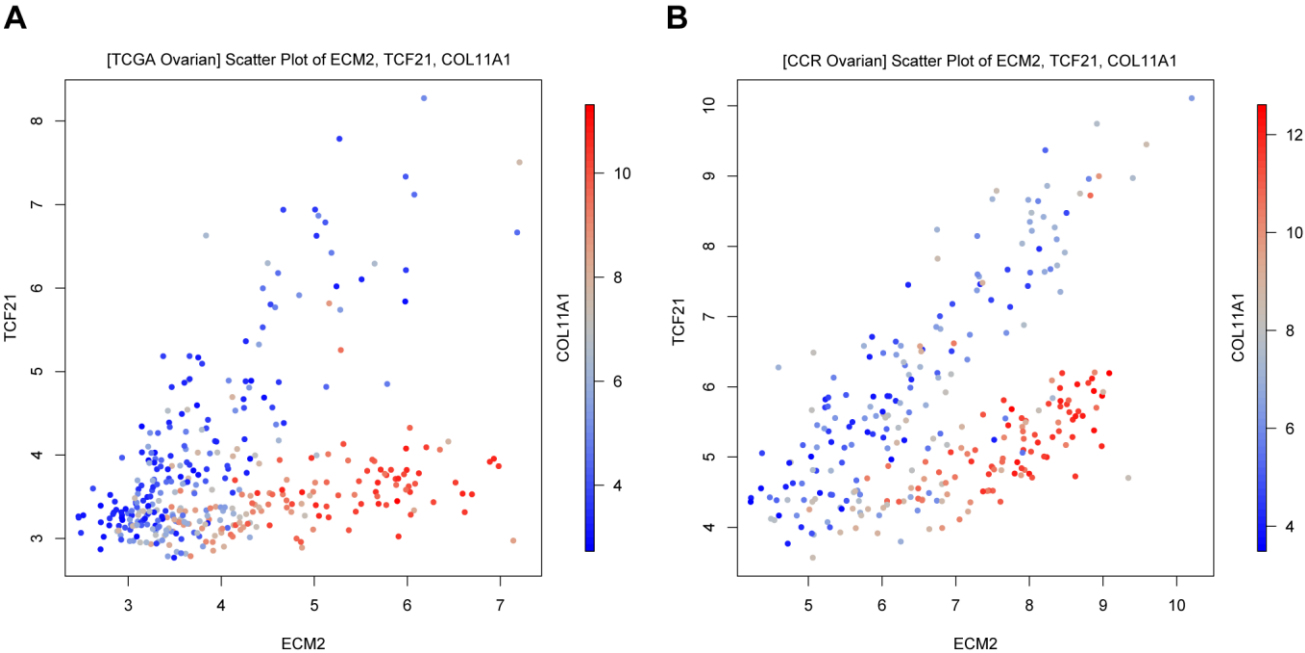
References

1. Thiery JP (2002) Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer* 2: 442-454.
2. Kalluri R, Weinberg RA (2009) The basics of epithelial-mesenchymal transition. *J Clin Invest* 119: 1420-1428.
3. Stover DG, Bierie B, Moses HL (2007) A delicate balance: TGF-beta and the tumor microenvironment. *J Cell Biochem* 101: 851-861.
4. Bignotti E, Tassi RA, Calza S, Ravaggi A, Bandiera E, et al. (2007) Gene expression profile of ovarian serous papillary carcinomas: identification of metastasis-associated genes. *Am J Obstet Gynecol* 196: 245 e241-211.
5. Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. (2008) Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 14: 5198-5208.
6. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, et al. (2009) Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clin Cancer Res* 15: 7642-7651.
7. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 138: 958-968.
8. Schuetz CS, Bonin M, Clare SE, Nieselt K, Sotlar K, et al. (2006) Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. *Cancer Res* 66: 5278-5286.
9. Vecchi M, Nuciforo P, Romagnoli S, Confalonieri S, Pellegrini C, et al. (2007) Gene expression analysis of early and advanced gastric cancers. *Oncogene* 26: 4284-4294.
10. Badea L, Herlea V, Dima SO, Dumitrascu T, Popescu I (2008) Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology* 55: 2016-2027.
11. Apte MV, Wilson JS (2007) The Desmoplastic Reaction in Pancreatic Cancer: An Increasingly Recognised Foe. *Pancreatology* 7: 378-379.
12. Chong IW, Chang MY, Chang HC, Yu YP, Sheu CC, et al. (2006) Great potential of a panel of multiple hMTH1, SPD, ITGA11 and COL11A1 markers for diagnosis of patients with non-small cell lung cancer. *Oncol Rep* 16: 981-988.
13. Schmalbach CE, Chepeha DB, Giordano TJ, Rubin MA, Teknos TN, et al. (2004) Molecular profiling and the identification of genes associated with metastatic oral cavity/pharynx squamous cell carcinoma. *Arch Otolaryngol Head Neck Surg* 130: 295-302.
14. Anastassiou D (2007) Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol* 3: 83.
15. Watkinson J, Liang KC, Wang X, Zheng T, Anastassiou D (2009) Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann N Y Acad Sci* 1158: 302-313.

16. Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, et al. (2009) A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. *Nat Med* 15: 68-74.
17. Kessenbrock K, Plaks V, Werb Z (2010) Matrix metalloproteinases: Regulators of the tumor microenvironment. *Cell* 141: 52-67.
18. Atfi A, Dumont E, Colland F, Bonnier D, L'Helgoualc'h A, et al. (2007) The disintegrin and metalloproteinase ADAM12 contributes to TGF-beta signaling through interaction with the type II receptor. *J Cell Biol* 178: 201-208.
19. Satoh K, Hata M, Shimizu T, Yokota H, Akatsu H, et al. (2005) Lib, transcriptionally induced in senile plaque-associated astrocytes, promotes glial migration through extracellular matrix. *Biochem Biophys Res Commun* 335: 631-636.
20. Kleeff J, Ishiwata T, Friess H, Buchler MW, Korc M (1998) Concomitant over-expression of activin/inhibin beta subunits and their receptors in human pancreatic cancer. *Int J Cancer* 77: 860-868.
21. Ohga E, Matsuse T, Teramoto S, Ouchi Y (2000) Activin receptors are expressed on human lung fibroblast and activin A facilitates fibroblast-mediated collagen gel contraction. *Life Sci* 66: 1603-1613.
22. Talmadge JE (2008) Follistatin as an inhibitor of experimental metastasis. *Clin Cancer Res* 14: 624-626.
23. Ogino H, Yano S, Kakiuchi S, Muguruma H, Ikuta K, et al. (2008) Follistatin suppresses the production of experimental multiple-organ metastasis by small cell lung cancer cells in natural killer cell-depleted SCID mice. *Clin Cancer Res* 14: 660-667.
24. Ellsworth RE, Seebach J, Field LA, Heckman C, Kane J, et al. (2009) A gene expression signature that defines breast cancer metastases. *Clin Exp Metastasis* 26: 205-213.
25. Weinberg RA (2007) *The biology of cancer*. New York: Garland Science. 1 v. (various pagings) p.
26. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the Gap statistic. *J R Statist Soc B* 63: 411-423.

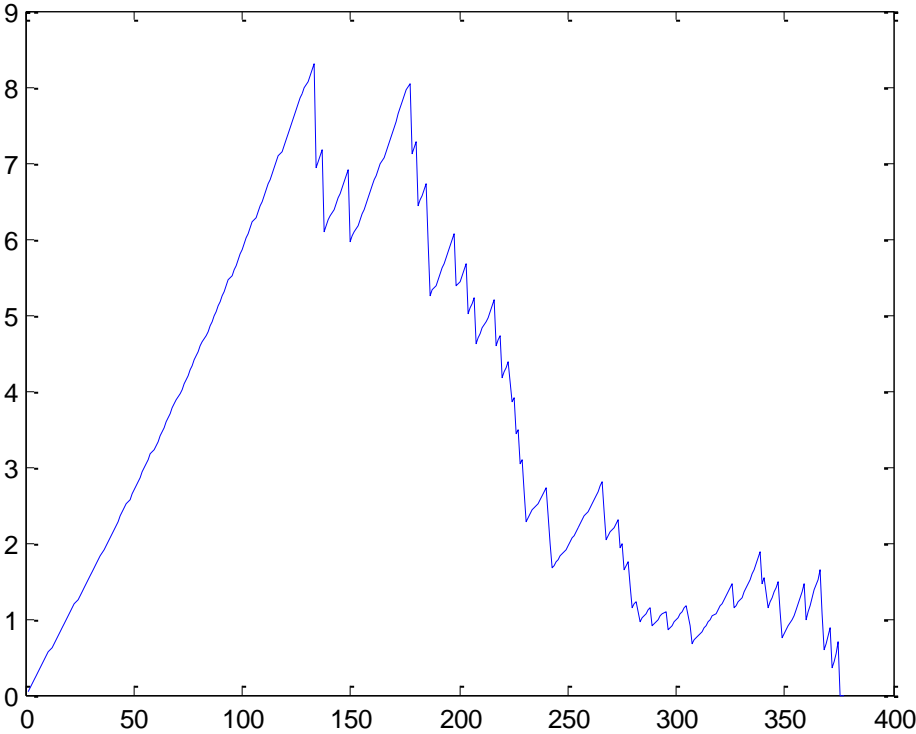
Figure Legends and Figures

Figure 1: Example of a synergistic pair of genes in two ovarian cancer datasets



Nature Precedings : hdl:10101/npre.2010.4503.1 : Posted 28 May 2010

Figure 2: Evaluation of the EVA metric for gene *COL11A1* in the TCGA ovarian cancer dataset using phenotypic staging threshold the transition to stage IIIc.



Nature Precedings : hdl:10101/npre.2010.4503.1 : Posted 28 May 2010

Figure 3: Illustration for the low-complexity implementation of the EVA algorithm.

| | | | | | | | |
|------------------------|-------------------------|----------|----------|----------|----------|----------|-------|
| <i>Low Stage Count</i> | 3 | 0.018 | 0.071 | 0.179 | 0.357 | 0.625 | 1.000 |
| | 2 | 0.107 | 0.268 | 0.429 | 0.536 | 0.536 | 0.375 |
| | 1 | 0.375 | 0.536 | 0.536 | 0.429 | 0.268 | 0.107 |
| | 0 | 1.000 | 0.625 | 0.357 | 0.179 | 0.071 | 0.018 |
| | 0 | 1 | 2 | 3 | 4 | 5 | |
| | <i>High Stage Count</i> | | | | | | |

Tables

Table 1: Top-ranked genes associated with high carcinoma stage in ovarian and colorectal cancers according to the EVA-based algorithm with Bonferroni corrected $P < 10^{-3}$ in all four data sets

| Probe Set^a | Gene | Log FC |
|------------------------------|-------------|---------------|
| 37892_at | COL11A1 | 3.94 |
| 217428_s_at | COL10A1 | 3.55 |
| 204320_at | COL11A1 | 3.39 |
| 210809_s_at | POSTN | 3.14 |
| 219087_at | ASPN | 2.99 |
| 205941_s_at | COL10A1 | 2.88 |
| 203083_at | THBS2 | 2.81 |
| 209955_s_at | FAP | 2.73 |
| 215446_s_at | LOX | 2.63 |
| 213764_s_at | MFAP5 | 2.61 |
| 210511_s_at | INHBA | 2.52 |
| 215646_s_at | VCAN | 2.5 |
| 209758_s_at | MFAP5 | 2.42 |
| 221730_at | COL5A2 | 2.34 |
| 211571_s_at | VCAN | 2.33 |
| 205713_s_at | COMP | 2.31 |
| 213765_at | MFAP5 | 2.27 |
| 201150_s_at | TIMP3 | 2.25 |
| 221729_at | COL5A2 | 2.24 |
| 212354_at | SULF1 | 2.23 |
| 212489_at | COL5A1 | 2.22 |
| 213790_at | ADAM12 | 2.21 |
| 212488_at | COL5A1 | 2.2 |
| 201147_s_at | TIMP3 | 2.19 |
| 204457_s_at | GAS1 | 2.17 |
| 202952_s_at | ADAM12 | 2.12 |
| 202766_s_at | FBN1 | 2.08 |
| 212344_at | SULF1 | 2.07 |

^aAffymetrix probe sets

Table 2: Gene lists produced from information provided in the corresponding papers for breast, gastric and pancreatic cancer.

| Breast Cancer, Shuetz et al ^a | | | Gastric cancer, Vecchi et al ^b | | | Pancreatic cancer, Badea et al ^c | | |
|--|-------------|--------|---|-------------|--------|---|-------------|--------|
| Probe Set ^d | Gene Symbol | Log FC | Probe Set ^d | Gene Symbol | Log FC | Probe Set ^d | Gene Symbol | Log FC |
| 37892_at | COL11A1 | 6.50 | 37892_at | COL11A1 | 4.26 | 227140_at | INHBA | 5.15 |
| 204320_at | COL11A1 | 4.08 | 217428_s_at | COL10A1 | 4.15 | 217428_s_at | COL10A1 | 5.00 |
| 217428_s_at | COL10A1 | 4.07 | 209955_s_at | FAP | 3.40 | 1555778_a_at | POSTN | 4.92 |
| 213764_s_at | MFAP5 | 3.73 | 235458_at | HAVCR2 | 3.30 | 212353_at | SULF1 | 4.63 |
| 213909_at | LRRRC15 | 3.61 | 204320_at | COL11A1 | 3.28 | 226237_at | COL8A1 | 4.60 |
| 205941_s_at | COL10A1 | 3.52 | 205941_s_at | COL10A1 | 3.21 | 37892_at | COL11A1 | 4.40 |
| 210511_s_at | INHBA | 3.44 | 204052_s_at | SFRP4 | 2.90 | 225681_at | CTHRC1 | 4.38 |
| 202766_s_at | FBN1 | 3.43 | 226930_at | FNDC1 | 2.85 | 202311_s_at | COL1A1 | 4.12 |
| 212353_at | SULF1 | 3.35 | 227140_at | INHBA | 2.77 | 203083_at | THBS2 | 3.97 |
| 218468_s_at | GREM1 | 3.35 | 209875_s_at | SPP1 | 2.77 | 227566_at | HNT | 3.90 |
| 215446_s_at | LOX | 3.22 | 205422_s_at | ITGBL1 | 2.63 | 204619_s_at | CSPG2 | 3.87 |
| 221730_at | COL5A2 | 3.22 | 226311_at | --- | 2.63 | 229802_at | WISP1 | 3.80 |
| 218469_at | GREM1 | 3.20 | 222288_at | --- | 2.62 | 212464_s_at | FN1 | 3.69 |
| 212489_at | COL5A1 | 3.08 | 231993_at | --- | 2.50 | 205713_s_at | COMP | 3.53 |
| 203083_at | THBS2 | 2.99 | 226237_at | COL8A1 | 2.48 | 221729_at | COL5A2 | 3.38 |
| 201505_at | LAMB1 | 2.97 | 223122_s_at | SFRP2 | 2.47 | 209955_s_at | FAP | 3.37 |
| 209955_s_at | FAP | 2.96 | 210511_s_at | INHBA | 2.43 | 229218_at | COL1A2 | 3.16 |
| 209758_s_at | MFAP5 | 2.92 | 203819_s_at | IMP-3 | 2.39 | 209016_s_at | KRT7 | 3.13 |
| 202363_at | SPOCK | 2.91 | 212464_s_at | FN1 | 2.36 | 210004_at | OLR1 | 3.03 |
| 213241_at | NY-REN-58 | 2.90 | 212353_at | SULF1 | 2.35 | 219773_at | NOX4 | 3.02 |
| 205479_s_at | PLAU | 2.89 | 227995_at | --- | 2.34 | 218804_at | TMEM16A | 2.90 |
| 206584_at | LY96 | 2.88 | 225681_at | CTHRC1 | 2.30 | 238617_at | --- | 2.87 |
| 204475_at | MMP1 | 2.83 | 204457_s_at | GAS1 | 2.27 | 224694_at | ANTXR1 | 2.82 |
| 202952_s_at | ADAM12 | 2.83 | 216442_x_at | FN1 | 2.25 | 228481_at | COX7A1 | 2.77 |
| 201792_at | AEBP1 | 2.81 | 223121_s_at | SFRP2 | 2.23 | 226311_at | ADAMTS2 | 2.76 |
| 204114_at | NID2 | 2.81 | 211719_x_at | FN1 | 2.23 | 201792_at | AEBP1 | 2.68 |
| 213790_at | ADAM12 | 2.80 | 204776_at | THBS4 | 2.18 | 203021_at | SLPI | 2.65 |
| 209156_s_at | COL6A2 | 2.77 | 210495_x_at | FN1 | 2.15 | 227314_at | ITGA2 | 2.58 |
| 219179_at | DACT1 | 2.74 | 202800_at | SLC1A3 | 2.13 | 205499_at | SRPX2 | 2.44 |
| 212488_at | COL5A1 | 2.73 | 214927_at | --- | 2.11 | 226997_at | --- | 2.41 |
| 219087_at | ASPN | 2.73 | 212354_at | SULF1 | 2.09 | 219179_at | DACT1 | 2.36 |
| 204619_s_at | CSPG2 | 2.70 | 238654_at | LOC147645 | 2.06 | 203570_at | LOXL1 | 2.30 |
| 204337_at | RGS4 | 2.69 | 213943_at | TWIST1 | 2.06 | 201850_at | CAPG | 2.25 |
| 204620_s_at | CSPG2 | 2.69 | 236028_at | IBSP | 2.05 | 222449_at | TMEPAI | 2.19 |
| 212354_at | SULF1 | 2.68 | 228481_at | POSTN | 2.00 | 227276_at | PLXDC2 | 2.16 |

^aBreast cancer list indicates genes overexpressed in invasive ductal carcinoma vs. ductal carcinoma in situ.

^bGastric cancer list indicates genes overexpressed in early gastric cancer vs. advanced gastric cancer.

^cPancreatic cancer list indicates genes overexpressed in pancreatic ductal adenocarcinoma vs. normal pancreatic tissue.

^dAffymetrix probe sets

Table 3: Genes associated with *COL11A1* from various cancers

| TCGA OV | CCR OV | CCR Colon | Moffitt Colon | Singapore Gastric | Yu Multicancer Tumor | expO data sets | | | | | | | | | | | | | | | | | | | | |
|-----------------|-----------------|-----------------|-----------------|-------------------|----------------------|-----------------|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | | | | | | batch 1 | batch 2 | batch 3 | batch 4 | batch 5 | batch 6 | batch 7 | batch 8 | batch 9 | batch 10 | batch 11 | batch 12 | batch 13 | batch 14 | batch 15 | batch 16 | | | | | |
| <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> | <i>COL11A1</i> |
| <i>THBS2</i> | <i>FAP</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> |
| <i>COL5A2</i> | <i>POSTN</i> | <i>THBS2</i> | <i>ADAM12</i> | <i>MMP11</i> | <i>LRR15</i> | <i>ADAM12</i> | <i>LRR15</i> | <i>COL1A1</i> | <i>THBS2</i> | <i>LRR15</i> | <i>CTHRC1</i> | <i>INHBA</i> | <i>LRR15</i> | <i>COL1A1</i> | <i>COL1A1</i> | <i>ADAMTS12</i> | <i>COL10A1</i> | <i>COL1A1</i> | <i>LRR15</i> | <i>LRR15</i> | <i>ADAMTS12</i> | <i>ADAMTS12</i> | <i>ADAMTS12</i> | <i>ADAMTS12</i> | <i>ADAMTS12</i> | <i>ADAMTS12</i> |
| <i>INHBA</i> | <i>THBS2</i> | <i>ADAM12</i> | <i>HNT</i> | <i>COL1A1</i> | <i>COL1A2</i> | <i>ADAMTS12</i> | <i>MXRAS</i> | <i>CTHRC1</i> | <i>COL1A2</i> | <i>ADAM12</i> | <i>COL5A2</i> | <i>COL1A1</i> | <i>COL1A1</i> | <i>LRR15</i> | <i>FNDC1</i> | <i>COL1A1</i> | <i>COL3A1</i> | <i>ADAMTS12</i> | <i>FNDC1</i> | <i>COL1A1</i> | <i>COL3A1</i> | <i>ADAMTS12</i> | <i>FNDC1</i> | <i>COL1A1</i> | <i>COL5A2</i> | <i>COL5A2</i> |
| <i>COL5A1</i> | <i>HNT</i> | <i>HNT</i> | <i>THBS2</i> | <i>ADAMTS12</i> | <i>MMP11</i> | <i>PPAPDC1A</i> | <i>ADAMTS12</i> | <i>COL5A2</i> | <i>FAP</i> | <i>COL1A1</i> | <i>LRR15</i> | <i>ADAMTS12</i> | <i>THBS2</i> | <i>CTHRC1</i> | <i>LRR15</i> | <i>COL5A2</i> | <i>COL12A1</i> | <i>COL1A1</i> | <i>ADAM12</i> | <i>COL3A1</i> | <i>CTHRC1</i> | <i>CTHRC1</i> | <i>CTHRC1</i> | <i>CTHRC1</i> | <i>CTHRC1</i> | <i>CTHRC1</i> |
| <i>POSTN</i> | <i>INHBA</i> | <i>CTHRC1</i> | <i>RAB31</i> | <i>THBS2</i> | <i>COL1A1</i> | <i>COL12A1</i> | <i>PPAPDC1A</i> | <i>COL5A1</i> | <i>COL1A1</i> | <i>COL1A2</i> | <i>FAP</i> | <i>THBS2</i> | <i>INHBA</i> | <i>INHBA</i> | <i>ADAMTS12</i> | <i>INHBA</i> | <i>COL5A1</i> | <i>ADAM12</i> | <i>INHBA</i> | <i>LRR15</i> | <i>LRR15</i> | <i>LRR15</i> | <i>LRR15</i> | <i>COL5A1</i> | <i>COL5A1</i> | <i>COL5A1</i> |
| <i>VCAN</i> | <i>EPYC</i> | <i>SPOCK1</i> | <i>COL5A2</i> | <i>INHBA</i> | <i>COL5A2</i> | <i>CTHRC1</i> | <i>ADAM12</i> | <i>THBS2</i> | <i>INHBA</i> | <i>FAP</i> | <i>POSTN</i> | <i>CTHRC1</i> | <i>COMP</i> | <i>COL5A2</i> | <i>COL5A2</i> | <i>LRR15</i> | <i>ADAMTS12</i> | <i>FN1</i> | <i>ADAMTS12</i> | <i>COL5A2</i> | <i>FNDC1</i> | <i>ADAMTS12</i> | <i>COL5A2</i> | <i>FNDC1</i> | <i>FNDC1</i> | <i>FNDC1</i> |
| <i>MMP11</i> | <i>COL5A2</i> | <i>BGN</i> | <i>CTHRC1</i> | <i>SULF1</i> | <i>INHBA</i> | <i>FNDC1</i> | <i>COL5A2</i> | <i>COL1A2</i> | <i>ADAMTS12</i> | <i>COL5A1</i> | <i>COL5A1</i> | <i>LRR15</i> | <i>COL1A2</i> | <i>COL1A2</i> | <i>COL5A1</i> | <i>CTHRC1</i> | <i>FNDC1</i> | <i>COL1A2</i> | <i>FAP</i> | <i>COL5A1</i> | <i>COL1A2</i> | <i>COL1A2</i> | <i>COL1A2</i> | <i>COL1A2</i> | <i>COL1A2</i> | <i>COL1A2</i> |
| <i>FAP</i> | <i>COL10A1</i> | <i>FAP</i> | <i>SPOCK1</i> | <i>CTHRC1</i> | <i>COL5A1</i> | <i>THBS2</i> | <i>FNDC1</i> | <i>MXRAS</i> | <i>COL5A2</i> | <i>ADAMTS12</i> | <i>ADAMTS12</i> | <i>FNDC1</i> | <i>COL5A2</i> | <i>COL5A1</i> | <i>FAP</i> | <i>FAP</i> | <i>COL5A2</i> | <i>FNDC1</i> | <i>CTHRC1</i> | <i>ITGA11</i> | <i>LRR15</i> | <i>LRR15</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL3A1</i> |
| <i>LRR15</i> | <i>LRR15</i> | <i>INHBA</i> | <i>SULF1</i> | <i>PPAPDC1A</i> | <i>VCAN</i> | <i>PLAU</i> | <i>COL1A2</i> | <i>LRR15</i> | <i>LRR15</i> | <i>COL12A1</i> | <i>COL1A1</i> | <i>COL5A2</i> | <i>CTHRC1</i> | <i>FN1</i> | <i>CTHRC1</i> | <i>POSTN</i> | <i>FAP</i> | <i>COL5A1</i> | <i>POSTN</i> | <i>FAP</i> | <i>COL5A1</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL3A1</i> |
| <i>COL3A1</i> | <i>SFRP2</i> | <i>PPAPDC1A</i> | <i>INHBA</i> | <i>WNT2</i> | <i>THBS2</i> | <i>ANTXR1</i> | <i>COL8A2</i> | <i>ADAMTS12</i> | <i>COL5A1</i> | <i>INHBA</i> | <i>THBS2</i> | <i>COL5A1</i> | <i>POSTN</i> | <i>FNDC1</i> | <i>POSTN</i> | <i>COL5A1</i> | <i>ADAM12</i> | <i>PLAU</i> | <i>COL1A1</i> | <i>FAP</i> | <i>INHBA</i> | <i>INHBA</i> | <i>INHBA</i> | <i>INHBA</i> | <i>INHBA</i> | <i>INHBA</i> |
| <i>COL1A1</i> | <i>VCAN</i> | <i>SULF1</i> | <i>GAS1</i> | <i>COL5A2</i> | <i>COL3A1</i> | <i>COL5A2</i> | <i>COL5A1</i> | <i>FNDC1</i> | <i>POSTN</i> | <i>POSTN</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL5A1</i> | <i>THBS2</i> | <i>INHBA</i> | <i>COL3A1</i> | <i>COL1A2</i> | <i>EPYC</i> | <i>POSTN</i> | <i>COL1A2</i> | <i>FAM49B</i> | <i>FAM49B</i> | <i>FAM49B</i> | <i>FAM49B</i> | <i>FAM49B</i> | <i>FAM49B</i> |
| <i>CTSK</i> | <i>FN1</i> | <i>COL1A1</i> | <i>AEBP1</i> | <i>C5orf46</i> | <i>ADAM12</i> | <i>ITGA11</i> | <i>CTHRC1</i> | <i>POSTN</i> | <i>FNDC1</i> | <i>THBS2</i> | <i>COL1A2</i> | <i>POSTN</i> | <i>SFRP2</i> | <i>ADAMTS12</i> | <i>THBS2</i> | <i>THBS2</i> | <i>CTHRC1</i> | <i>COL5A2</i> | <i>THBS2</i> | <i>FNDC1</i> | <i>FAP</i> | <i>FAP</i> | <i>FAP</i> | <i>FAP</i> | <i>FAP</i> | <i>FAP</i> |
| <i>LOX</i> | <i>ADAMTS12</i> | <i>VCAN</i> | <i>SPARC</i> | <i>ANTXR1</i> | <i>BGN</i> | <i>FAP</i> | <i>FAP</i> | <i>ADAM12</i> | <i>COL3A1</i> | <i>INHBA</i> | <i>COL1A2</i> | <i>COL3A1</i> | <i>FAP</i> | <i>COL3A1</i> | <i>CTSK</i> | <i>MXRAS</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> |
| <i>LUM</i> | <i>LOX</i> | <i>COL5A2</i> | <i>FBN1</i> | <i>HNT</i> | <i>FAP</i> | <i>COL1A2</i> | <i>COMP</i> | <i>COL3A1</i> | <i>COL3A1</i> | <i>COL5A2</i> | <i>FNDC1</i> | <i>ITGA11</i> | <i>AEBP1</i> | <i>COL3A1</i> | <i>COL1A2</i> | <i>FNDC1</i> | <i>PLAU</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> | <i>PPAPDC1A</i> |
| <i>CRISPLD2</i> | <i>ALPK2</i> | <i>ADAMTS12</i> | <i>COL1A2</i> | <i>FN1</i> | <i>AEBP1</i> | <i>NOX4</i> | <i>SFRP2</i> | <i>AEBP1</i> | <i>FN1</i> | <i>COL8A2</i> | <i>COL12A1</i> | <i>FAP</i> | <i>MXRAS</i> | <i>POSTN</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>LRR15</i> | <i>CTHRC1</i> | <i>COL5A1</i> | <i>POSTN</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> | <i>COL10A1</i> |
| <i>COL10A1</i> | <i>COL3A1</i> | <i>POSTN</i> | <i>COL5A1</i> | <i>FNDC1</i> | <i>MFAP5</i> | <i>COL5A1</i> | <i>COMP</i> | <i>CTHRC1</i> | <i>CTHRC1</i> | <i>COL3A1</i> | <i>SFRP2</i> | <i>ADAMTS12</i> | <i>COL12A1</i> | <i>COL12A1</i> | <i>COL1A2</i> | <i>POSTN</i> | <i>CTHRC1</i> | <i>TPX2</i> | <i>CTHRC1</i> | <i>PLAU</i> | <i>PLAU</i> | <i>PLAU</i> | <i>PLAU</i> | <i>PLAU</i> | <i>PLAU</i> | <i>PLAU</i> |
| <i>FBN1</i> | <i>GREM1</i> | <i>RAB31</i> | <i>PRRX1</i> | <i>SPHK1</i> | <i>CDH11</i> | <i>SULF1</i> | <i>COL3A1</i> | <i>FBN1</i> | <i>ITGA11</i> | <i>ITGA11</i> | <i>SFRP2</i> | <i>FN1</i> | <i>FAP</i> | <i>ITGBL1</i> | <i>COL6A3</i> | <i>TNFSF4</i> | <i>THBS2</i> | <i>FAM49B</i> | <i>FN1</i> | <i>MFAP2</i> | <i>POPN</i> | <i>POPN</i> | <i>POPN</i> | <i>POPN</i> | <i>POPN</i> | <i>POPN</i> |
| <i>FN1</i> | <i>COL5A1</i> | <i>AEBP1</i> | <i>COL1A1</i> | <i>KAL1</i> | <i>SULF1</i> | <i>FN1</i> | <i>EPYC</i> | <i>COL6A3</i> | <i>PPAPDC1A</i> | <i>MXRAS</i> | <i>SULF1</i> | <i>PPAPDC1A</i> | <i>COL8A2</i> | <i>COMP</i> | <i>COMP</i> | <i>EPYC</i> | <i>EPYC</i> | <i>SULF1</i> | <i>SFRP2</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> |
| <i>ADAM12</i> | <i>ADAM12</i> | <i>COL1A2</i> | <i>MXRAS</i> | <i>COL1A2</i> | <i>COL6A3</i> | <i>SRPX2</i> | <i>MFAP5</i> | <i>FN1</i> | <i>AEBP1</i> | <i>CDH11</i> | <i>COMP</i> | <i>AEBP1</i> | <i>LOXL1</i> | <i>SFRP2</i> | <i>COL8A2</i> | <i>CTSK</i> | <i>COL8A2</i> | <i>CTSK</i> | <i>ITGBL1</i> | <i>CTSK</i> | <i>ITGBL1</i> | <i>ITGBL1</i> | <i>ITGBL1</i> | <i>ITGBL1</i> | <i>ITGBL1</i> | <i>ITGBL1</i> |
| <i>MMP2</i> | <i>COL1A1</i> | <i>PRRX1</i> | <i>FAP</i> | <i>NOX4</i> | <i>POSTN</i> | <i>COL1A1</i> | <i>COL8A2</i> | <i>SULF1</i> | <i>COL6A3</i> | <i>MXRAS</i> | <i>VCAN</i> | <i>ADAM12</i> | <i>COL8A2</i> | <i>CTSK</i> | <i>CTSK</i> | <i>COL8A2</i> | <i>CTSK</i> | <i>ITGBL1</i> | <i>CTSK</i> | <i>IGFL2</i> | <i>HGD</i> | <i>HTRA3</i> | <i>KIAA1199</i> | <i>KIAA1199</i> | <i>KIAA1199</i> | |
| <i>AEBP1</i> | <i>CRISPLD2</i> | <i>COL5A1</i> | <i>BGN</i> | <i>CST1</i> | <i>MMP13</i> | <i>MXRAS</i> | <i>LOC401097</i> | <i>SPARC</i> | <i>HNT</i> | <i>FN1</i> | <i>ADAM12</i> | <i>SULF1</i> | <i>FNDC1</i> | <i>ITGA11</i> | <i>GLT8D2</i> | <i>MXRAS</i> | <i>POPN</i> | <i>POSTN</i> | <i>UBE2C</i> | <i>ALDH1L2</i> | <i>MMP14</i> | <i>MMP14</i> | <i>MMP14</i> | <i>MMP14</i> | <i>MMP14</i> | <i>MMP14</i> |
| <i>PLAU</i> | <i>MMP11</i> | <i>KAL1</i> | <i>POSTN</i> | <i>BGN</i> | <i>CTSK</i> | <i>EPYC</i> | <i>COL12A1</i> | <i>SFRP2</i> | <i>COL6A3</i> | <i>FNDC1</i> | <i>PRRX1</i> | <i>MXRAS</i> | <i>COL6A3</i> | <i>SULF1</i> | <i>ITGBL1</i> | <i>COL12A1</i> | <i>SFRP2</i> | <i>SLC12A8</i> | <i>COL1A2</i> | <i>RCN3</i> | <i>SRPX2</i> | <i>SRPX2</i> | <i>SRPX2</i> | <i>SRPX2</i> | <i>SRPX2</i> | <i>SRPX2</i> |
| <i>ASPN</i> | <i>PRRX1</i> | <i>SFRP4</i> | <i>SFRP2</i> | <i>SOX11</i> | <i>LUM</i> | <i>COL8A2</i> | <i>ANTXR1</i> | <i>INHBA</i> | <i>SPARC</i> | <i>COMP</i> | <i>AEBP1</i> | <i>CDH11</i> | <i>RCN3</i> | <i>SPARC</i> | <i>CDH11</i> | <i>COL6A3</i> | <i>MMP2</i> | <i>C15orf48</i> | <i>ASPM</i> | <i>ASPN</i> | <i>ASPN</i> | <i>ASPN</i> | <i>ASPN</i> | <i>ASPN</i> | <i>ASPN</i> | <i>ASPN</i> |
| <i>TMEM158</i> | <i>TMEM158</i> | <i>FBN1</i> | <i>PPAPDC1A</i> | <i>ITGA11</i> | <i>PPAP2B</i> | <i>WISP1</i> | <i>ITGA11</i> | <i>SULF1</i> | <i>FBN1</i> | <i>ASPN</i> | <i>COL16A1</i> | <i>ADAM12</i> | <i>SULF1</i> | <i>TIMP3</i> | <i>FN1</i> | <i>GREM1</i> | <i>SRPX2</i> | <i>THBS2</i> | <i>KIF4A</i> | <i>COL6A3</i> | <i>ADA</i> | <i>ADA</i> | <i>ADA</i> | <i>ADA</i> | <i>ADA</i> | <i>ADA</i> |
| <i>ECM1</i> | <i>CTSK</i> | <i>SFRP2</i> | <i>NOX4</i> | <i>CST4</i> | <i>FN1</i> | <i>THY1</i> | <i>RUNX2</i> | <i>VCAN</i> | <i>COL12A1</i> | <i>MFAP2</i> | <i>PPAPDC1A</i> | <i>COL12A1</i> | <i>MFAP5</i> | <i>GLT8D2</i> | <i>COL8A2</i> | <i>GLT8D2</i> | <i>SULF1</i> | <i>MXRAS</i> | <i>NUF2</i> | <i>ITGBL1</i> | <i>ADAM19</i> | <i>ADAM19</i> | <i>ADAM19</i> | <i>ADAM19</i> | <i>ADAM19</i> | <i>ADAM19</i> |
| <i>COL1A2</i> | <i>SPARC</i> | <i>ANTXR1</i> | <i>VCAN</i> | <i>COL5A1</i> | <i>RCN3</i> | <i>LRR15</i> | <i>TNFSF4</i> | <i>PPAPDC1A</i> | <i>MXRAS</i> | <i>PPAPDC1A</i> | <i>MFAP2</i> | <i>COL6A3</i> | <i>VCAN</i> | <i>ADAM12</i> | <i>EPYC</i> | <i>POPN</i> | <i>RCN3</i> | <i>EPH2</i> | <i>CENPF</i> | <i>SRPX2</i> | <i>BMP1</i> | <i>BMP1</i> | <i>BMP1</i> | <i>BMP1</i> | <i>BMP1</i> | <i>BMP1</i> |
| <i>SPARC</i> | <i>FBN1</i> | <i>GAS1</i> | <i>ANTXR1</i> | <i>FAP</i> | <i>TPO</i> | <i>IGFL2</i> | <i>COL6A3</i> | <i>COL12A1</i> | <i>VCAN</i> | <i>AEBP1</i> | <i>CTSK</i> | <i>COL8A2</i> | <i>LUM</i> | <i>CDH11</i> | <i>VCAN</i> | <i>SULF1</i> | <i>COL6A3</i> | <i>DLGAP5</i> | <i>ACADL</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> | <i>ADAM12</i> |
| <i>EPYC</i> | <i>CTHRC1</i> | <i>COL3A1</i> | <i>FRMD6</i> | <i>MMP14</i> | <i>OLFM12B</i> | <i>CDH11</i> | <i>PDGFRL</i> | <i>DCN</i> | <i>COMP</i> | <i>MMP13</i> | <i>COL8A2</i> | <i>BGN</i> | <i>PCOLCE</i> | <i>MXRAS</i> | <i>SULF1</i> | <i>FN1</i> | <i>MMP1</i> | <i>KIF20A</i> | <i>PLAU</i> | <i>MXRAS</i> | <i>C11orf52</i> | <i>C11orf52</i> | <i>C11orf52</i> | <i>C11orf52</i> | <i>C11orf52</i> | <i>C11orf52</i> |
| <i>COL6A3</i> | <i>GJB2</i> | <i>FNDC1</i> | <i>COL3A1</i> | <i>LRR15</i> | <i>KCNJ16</i> | <i>POSTN</i> | <i>ALDH1L2</i> | <i>CTSK</i> | <i>COL8A2</i> | <i>SFRP2</i> | <i>SERPINF1</i> | <i>HNT</i> | <i>COL12A1</i> | <i>RCN3</i> | <i>SPARC</i> | <i>COMP</i> | <i>PLAU</i> | <i>MFAP5</i> | <i>DTL</i> | <i>FBN1</i> | <i>HPN</i> | <i>HPN</i> | <i>HPN</i> | <i>HPN</i> | <i>HPN</i> | <i>HPN</i> |
| <i>SERPINF1</i> | <i>TNFAIP6</i> | <i>HOPX</i> | <i>ADAMTS12</i> | <i>SERPINF1</i> | <i>MXRAS</i> | <i>COL3A1</i> | <i>GLT8D2</i> | <i>ITGBL1</i> | <i>WISP1</i> | <i>COL6A2</i> | <i>GLT8D2</i> | <i>SPARC</i> | <i>FN1</i> | <i>FBN1</i> | <i>DCN</i> | <i>CDH11</i> | <i>GREM1</i> | <i>CDC2</i> | <i>COL12A1</i> | <i>BMP1</i> | <i>TRAM2</i> | <i>TRAM2</i> | <i>TRAM2</i> | <i>TRAM2</i> | <i>TRAM2</i> | <i>TRAM2</i> |
| <i>OLFM12B</i> | <i>NUAK1</i> | <i>NOX4</i> | <i>TIMP2</i> | <i>POSTN</i> | <i>NWMT</i> | <i>HNT</i> | <i>CTSK</i> | <i>GLT8D2</i> | <i>GLT8D2</i> | <i>VCAN</i> | <i>CDH11</i> | <i>COMP</i> | <i>FBN1</i> | <i>AEBP1</i> | <i>LUM</i> | <i>WISP1</i> | <i>ALDH1L2</i> | <i>LIMK2</i> | <i>C2orf67</i> | <i>COL12A1</i> | <i>GJB2</i> | <i>GJB2</i> | <i>GJB2</i> | <i>GJB2</i> | <i>GJB2</i> | <i>GJB2</i> |
| <i>DCN</i> | <i>ASPN</i> | <i>LOX</i> | <i>COL12A1</i> | <i>SPP1</i> | <i>PAX8</i> | <i>INHBA</i> | <i>FBN1</i> | <i>OLFM12B</i> | <i>MRC2</i> | <i>SULF1</i> | <i>PCOLCE</i> | <i>TNFSF4</i> | <i>MFAP2</i> | <i>VCAN</i> | <i>SFRP2</i> | <i>HNT</i> | <i>FBN1</i> | <i>MFAP2</i> | <i>RNF180</i> | <i>MRC2</i> | <i>ECHDC2</i> | <i>ECHDC2</i> | <i>ECHDC2</i> | <i>ECHDC2</i> | <i>ECHDC2</i> | <i>ECHDC2</i> |
| <i>ACTA2</i> | <i>AEBP1</i> | <i>CERCAM</i> | <i>LOX</i> | <i>ADAMTS12</i> | <i>DUOX2</i> | <i>TNFSF4</i> | <i>TWIST1</i> | <i>ANTXR1</i> | <i>RCN3</i> | <i>SULF1</i> | <i>ANTXR1</i> | <i>FN1</i> | <i>RCN3</i> | <i>PRRX1</i> | <i>HNT</i> | <i>PLAU</i> | <i>FBN1</i> | <i>MMP3</i> | <i>LOC401097</i> | <i>BIRC5</i> | <i>COMP</i> | <i>ELF4</i> | <i>ELF4</i> | <i>ELF4</i> | <i>ELF4</i> | <i>ELF4</i> |
| <i>NTM</i> | <i>COL8A1</i> | <i>COL12A1</i> | <i>HTRA1</i> | <i>CLECSA</i> | <i>DCN</i> | <i>COL6A3</i> | <i>INHBA</i> | <i>TNFAIP6</i> | <i>ANTXR1</i> | <i>CTSK</i> | <i>HNT</i> | <i>GREM1</i> | <i>PLAU</i> | <i>COL8A1</i> | <i>HNT</i> | <i>ALDH1L2</i> | <i>CEACAM6</i> | <i>MMP11</i> | <i>CDC20</i> | <i>VCAN</i> | <i>CTSK</i> | <i>CTSK</i> | <i>CTSK</i> | <i>CTSK</i> | <i>CTSK</i> | <i>CTSK</i> |
| <i>GREM1</i> | <i>COL1A2</i> | <i>SPARC</i> | <i>TWIST1</i> | <i>BMP1</i> | <i>TG</i> | <i>PDGFRL</i> | <i>FN1</i> | <i>MMP2</i> | <i>CTSK</i> | <i>PCOLCE</i> | <i>FBN1</i> | <i>EPYC</i> | <i>RAB31</i> | <i>LOX</i> | <i>IGHG1</i> | <i>RCN3</i> | <i>C11orf24</i> | <i>SLAMF8</i> | <i>CENPA</i> | <i>CERCAM</i> | <i>C3orf15</i> | <i>C3</i> | | | | |

Table 4: Top ranked (multiple-test corrected $P < 10^{-16}$) differentially expressed miRNAs in MAF signature in the TCGA ovarian cancer data set in terms of their association with *COL11A1*.

| miRNA | MI | Up/Down Regulated |
|---|-----------|--------------------------|
| hsa-miR-22 | 0.204 | <i>Up</i> |
| hsa-miR-514-1 hsa-miR-514-2 hsa-miR-514-3 | 0.193 | <i>Down</i> |
| hsa-miR-152 | 0.187 | <i>Up</i> |
| hsa-miR-508 | 0.168 | <i>Down</i> |
| hsa-miR-509-1 hsa-miR-509-2 hsa-miR-509-3 | 0.164 | <i>Down</i> |
| hsa-miR-507 | 0.152 | <i>Down</i> |
| hsa-miR-509-1 hsa-miR-509-2 | 0.147 | <i>Down</i> |
| hsa-miR-506 | 0.146 | <i>Down</i> |
| hsa-miR-509-3 | 0.144 | <i>Down</i> |
| hsa-miR-214 | 0.128 | <i>Up</i> |
| hsa-miR-510 | 0.116 | <i>Down</i> |
| hsa-miR-199a-1 hsa-miR-199a-2 | 0.115 | <i>Up</i> |
| hsa-miR-21 | 0.112 | <i>Up</i> |
| hsa-miR-513c | 0.108 | <i>Down</i> |
| hsa-miR-199b | 0.103 | <i>Up</i> |

Table 5: Top ranked (multiple-test corrected $P < 10^{-16}$) differentially methylated genes in MAF signature in the TCGA ovarian cancer data set in terms of their association with *COL11A1*.

| Methylation site | MI | Hyper-/Hypomethylated |
|-------------------------|-----------|------------------------------|
| PRAME | 0.223 | <i>Hyper</i> |
| SNAI1 | 0.183 | <i>Hyper</i> |
| KRT7 | 0.158 | <i>Hyper</i> |
| RASSF5 | 0.157 | <i>Hyper</i> |
| FLJ14816 | 0.155 | <i>Hyper</i> |
| PPL | 0.155 | <i>Hyper</i> |
| CXCR6 | 0.153 | <i>Hypo</i> |
| SLC12A8 | 0.148 | <i>Hyper</i> |
| NFATC2 | 0.148 | <i>Hyper</i> |
| HOM-TES-103 | 0.147 | <i>Hypo</i> |
| ZNF556 | 0.147 | <i>Hyper</i> |
| OCIAD2 | 0.146 | <i>Hyper</i> |
| APS | 0.142 | <i>Hyper</i> |
| MGC9712 | 0.139 | <i>Hyper</i> |
| SLC1A2 | 0.136 | <i>Hyper</i> |
| HAK | 0.131 | <i>Hypo</i> |
| C3orf18 | 0.130 | <i>Hyper</i> |
| GMPR | 0.130 | <i>Hyper</i> |
| CORO6 | 0.128 | <i>Hyper</i> |

Table 6: Data sets that were used in the paper

| Data set name | Source Site | GEO Accession | Affymetrix platform | Sample size |
|--------------------------|-------------------------|----------------------|-----------------------------|------------------------|
| TCGA ovarian cancer | The Cancer Genome Atlas | | HT_HG-U133A ^a | 377 |
| CCR ovarian cancer | Gene Expression Omnibus | GSE9891 | HG-U133_Plus_2 ^b | 285 |
| CCR colon cancer | Gene Expression Omnibus | GSE14333 | HG-U133_Plus_2 | 290 |
| Moffitt colon cancer | Gene Expression Omnibus | GSE17536 | HG-U133_Plus_2 | 177 |
| Singapore gastric cancer | Gene Expression Omnibus | GSE15459 | HG-U133_Plus_2 | 200 |
| Yu multicancer tumor | Gene Expression Omnibus | GSE5364 | HG-U133A ^c | 270 [341] ^d |
| expO batch 1 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 60 |
| expO batch 2 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 163 |
| expO batch 3 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 164 |
| expO batch 4 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 163 |
| expO batch 5 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 160 |
| expO batch 6 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 160 |
| expO batch 7 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 156 |
| expO batch 8 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 160 |
| expO batch 9 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 240 |
| expO batch 10 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 180 |
| expO batch 11 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 180 |
| expO batch 12 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 125 |
| expO batch 13 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 62 |
| expO batch 14 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 62 |
| expO batch 15 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 60 |
| expO batch 16 | Gene Expression Omnibus | GSE2109 | HG-U133_Plus_2 | 63 |

^aAffymetrix HT Human Genome U133A Array

^bAffymetrix Human Genome U133 Plus 2.0

^cAffymetrix Human Genome U133A

^d270 out of 341 samples are tumor samples.