

# Cancer invasion associated gene expression signature is present in differentially expressed genes in the reprogramming of fibroblasts into stem cells

Wei-Yi Cheng<sup>1</sup>, Hoon Kim<sup>1</sup>, Jessica Kandel<sup>2</sup> and Dimitris Anastassiou<sup>1\*</sup>

<sup>1</sup> Center for Computational Biology and Bioinformatics and Department of Electrical Engineering, Columbia University, New York, NY, USA

<sup>2</sup> Department of Pediatrics, College of Physicians and Surgeons of Columbia University, New York, NY, USA

\* Corresponding author email: [anastas@ee.columbia.edu](mailto:anastas@ee.columbia.edu)

## Abstract

**Tumors become invasive by penetrating adjacent connective tissue, but the underlying biological mechanisms remain obscure. We recently identified a precise gene expression signature of fibroblastic origin associated with cancer invasion, the first step of the metastatic cascade. The signature contains many coordinately overexpressed genes, prominent among which are *COL11A1*, *THBS2* and *INHBA*. Here we show that there is a striking similarity between the set of expressed genes in this metastasis-associated fibroblastic (MAF) signature and the set of genes that are downregulated when fibroblasts are reprogrammed to induced pluripotent stem cells (iPSCs). Because it is known that fibroblast reprogramming involves a mesenchymal-epithelial transition (MET), the above facts suggest that, conversely, the metastasis-associated fibroblasts responsible for the signature may result from stem-like cells undergoing some type of epithelial-mesenchymal transition (EMT). Therefore, we speculate that cancer stem cells (CSCs) undergoing some type of EMT become fibroblastic to obtain motility and invasiveness, reactivating early embryonic developmental pathways, and that these fibroblast-like cells are the main source of the MAF signature that we previously identified.**

## Introduction

We recently identified and reported [1] a precise gene expression signature consisting of a set of genes that are coordinately overexpressed only in samples of cancer that have exceeded a particular stage, specific to each cancer type (a previous version also available in Nature Precedings at <http://hdl.handle.net/10101/npre.2010.4503.1>). The same signature appears in solid

cancer types, including pancreatic, ovarian, colon, prostate, breast, gastric, neuroblastoma and Ewing's sarcoma, the only exceptions that we found being blood and brain cancers.

Among the overexpressed genes are various collagens and proteinases, fibroblast activation protein,  $\alpha$ -SMA, fibronectin and proteoglycans, suggesting a fibroblastic source. The signature, however, is not of a general fibroblastic nature, but instead has its own precise special characteristics, one of which is that genes *COL11A1*, *THBS2* and *INHBA* have the most prominent presence. We identified collagen *COL11A1* as a reliable proxy for the signature. In fact, in each rich cancer dataset, but not in non-cancer datasets, finding the list of genes whose expression is most correlated with that of *COL11A1* consistently identifies the other genes of the signature as a result of the presence of those high-stage samples that contain it. The signature also contains several transcription factors associated with epithelial-mesenchymal transition (EMT), particularly slug (*SNAI2*). Notably, however, it does not contain snail (*SNAI1*), which we found, at least in ovarian cancer, to be methylated.

Therefore, we have hypothesized that the signature corresponds to a biological mechanism leading to the presence of a particular type of fibroblasts, to which we refer as the "Metastasis Associated Fibroblasts" (MAFs), because cancer invasiveness is the first step of the metastatic cascade. Table 1 shows a list of the 64 genes corresponding to the 100 most overexpressed probe sets, as we previously reported [1], of the signature, to which we refer as the "MAF signature."

To obtain clues about the origin and nature of the MAF signature, we compared it with other known signatures. Among all signatures that we searched outside cancer datasets, we found by far the largest enrichment to be present in the set of genes that are downregulated in induced pluripotent stem cells (iPSCs) as well as embryonic stem cells (ESCs) compared to the fibroblasts before reprogramming. The presence of the signature was most evident in several gene expression datasets comparing mouse embryonic fibroblasts (MEFs) reprogrammed into iPSCs.

Because it is known [2] that a mesenchymal-epithelial transition (MET) is part of the reprogramming of mouse fibroblasts into stem-like cells, we hypothesize that, conversely, the MAF signature is mainly produced by fibroblasts resulting, at least partly, from cancer stem cells (CSSs) undergoing some type of EMT and obtaining mesenchymal phenotype.

## Results

Since the three genes *COL11A1*, *THBS2*, *INHBA* are required and prominent in the MAF signature, we performed Gene Set Enrichment Analysis in the Molecular Signature Database of the Broad Institute for these three genes. The search only revealed five sets containing all genes (Table 2). Consistent with the discovery of the MAF signature [1], four of these sets correspond to genes upregulated in high-stage vs. low-stage cancer samples, specifically in lobular breast cancer, gastric cancer, ductal breast cancer, and papillary thyroid cancer. The fifth one included 514 genes downregulated in primary fibroblast cell culture after infection with HCMV (AD169

strain) at 48 h time point that were not downregulated at the previous time point, 24 h. However, only nine of these 514 genes were among the 64 MAF genes of Table 1.

Additional independent search yielded some other results. For example, we found that part of the signature is present in keloid lesions [3], in which, however, *THBS2* is interestingly shown to be downregulated, contrary to its prominent upregulation in the MAF signature.

The most remarkable similarity by far, however, was identified in a study analyzing the reprogramming of somatic cells to induced pluripotent stem cells [4], where all three genes *COL11A1*, *THBS2*, *INHBA*, were mentioned in a list of genes found to be lower expressed in ES and iPS cells compared to mouse embryonic fibroblasts before reprogramming. Specifically, the section “Functional description of the genes lower expressed in ES and iPS cells compared to fibroblasts in mouse” in Supplementary text S1 mentions 22 among the top 64 MAF genes of Table 1, including, in addition to *COL11A1*, *THBS2*, *INHBA* an almost identical additional collagen composition (*COL5A1*, *COL5A2*, *COL1A1*, *COL1A2*, *COL3A1*, *COL6A1*, *COL6A3*) as well as genes *POSTN*, *ADAM12*, *LOX*, *FBN1*, *MMP2*, *TIMP3*, *DCN*, *ACTA2*, *PDGFRB*, *SNAI2*, *THY1*, *PRRX1*).

We then downloaded and analyzed the seven datasets from mouse fibroblast reprogramming mentioned in [4], as described in Materials and Methods, confirming the remarkable similarity of the differentially expressed genes with the MAF signature. Table 3 includes the 45 top-ranked genes that we identified, which includes 14 MAF genes from the 64 genes of Table 1 ( $P < 10^{-27}$ ). Four of these 14 MAF genes (*ASPN*, *LOXL2*, *CDH11*, *SERPINF1*) are in addition to the list of 22 MAF genes mentioned above. The resulting heat maps for the 64 MAF genes of Table 1 for the seven datasets are shown in Figures 1a-1g.

## Discussion

The parallels between embryonic development and tumor progression involving EMT have been well recognized [5]. On the other hand, recent research suggests that cell “stemness” and the ability to shift between epithelial and mesenchymal characteristics (probably in a continuous rather than abrupt manner) are qualities that appear to be closely related. For example, EMT generates cells with properties of stem cells [6]. Conversely, MET is involved in the reprogramming of fibroblasts into stem cells [2]. Therefore, our hypothesis, based purely on computational analysis, is consistent with the notion that cancer stem cells with mesenchymal characteristics, derived from cancer cells poised to undergo EMT, obtain metastatic potential [7].

We speculate that these fibroblastic cells (MAFs) originate from the primary tumor rather than the stroma and constitute the main source of the MAF signature. They could be triggered from contextual reactive stroma microenvironmental signals, and they would open passageways perhaps allowing for other cancer cells to also go through the adjacent connective tissue.

Furthermore, because the MAF signature is observed even in non-epithelial cancers, such as neuroblastoma [1], the above hypothesis would imply the employment of a mesenchymal transition mechanism more general than what EMT is assumed to be.

Many among the top MAF genes have previously been individually identified as associated with metastatic potential in cancer. If our hypothesis is correct, such associations can largely be explained by the fact that these genes are expressed by the cancer cells themselves after they have obtained mesenchymal phenotype and become invasive.

## Materials and Methods

### Gene Expression Data

The gene expression datasets were downloaded from Gene Expression Omnibus under accession ID GSE7815 [8], GSE7841 [9], GSE8024 [10], GSE13211 [11], GSE13770 [12], GSE14012 [13], GSE15267 [14]. The expression data were normalized as provided by the original author. We applied base-2 logarithm to the expression value if the dataset values were not log-transformed.

### Differential Expression Analysis

The differentially expressed genes were identified by the *limma* package in *R*. We applied linear model and empirical Bayesian methods on the probe-level data, then taking the minimum of the  $P$  values of all probes corresponding to the same genes as the gene's  $P$  value for differently expression. Since we are only interested in genes that are overexpressed in fibroblasts, we only computed the upper-tail  $P$  value in the fibroblast vs. ESC or iPSC comparisons.

After obtaining the  $P$  values of each gene in each datasets, we simply take the  $\log_{10}$ -average of the  $P$  values across datasets as the scores for overexpression for the gene across each datasets. We then rank the scores and Table 3 shows the top 45 entries.

### Data Visualization

We first collapsed the probe-level expression data into gene-level using the median of the probes corresponding to the same gene. Then we extracted the 64 MAF genes of Table 1 existing in the dataset to create heatmaps using GenePattern [15]. The column is clustered using pairwise average-linkage, with Euclidean distance as column distance measure.







## References

1. Kim H, Watkinson J, Varadan V, Anastassiou D: **Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1.** *BMC Med Genomics* 2010, **3**:51.
2. Polo JM, Hochedlinger K: **When fibroblasts MET iPSCs.** *Cell Stem Cell*, **7**:5-6.
3. Seifert O, Bayat A, Geffers R, Dienus K, Buer J, Lofgren S, Matussek A: **Identification of unique gene expression patterns within different lesional sites of keloids.** *Wound Repair Regen* 2008, **16**:254-265.
4. Boue S, Paramonov I, Barrero MJ, Izpisua Belmonte JC: **Analysis of human and mouse reprogramming of somatic cells to induced pluripotent stem cells. What is in the plate?** *PLoS One* 2010, **5**.
5. Micalizzi DS, Farabaugh SM, Ford HL: **Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression.** *J Mammary Gland Biol Neoplasia* 2010, **15**:117-134.
6. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M, et al: **The epithelial-mesenchymal transition generates cells with properties of stem cells.** *Cell* 2008, **133**:704-715.
7. Chaffer CL, Weinberg RA: **A perspective on cancer cell metastasis.** *Science*, **331**:1559-1564.
8. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch R, et al: **Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution.** *Cell Stem Cell* 2007, **1**:55-70.
9. Okita K, Ichisaka T, Yamanaka S: **Generation of germline-competent induced pluripotent stem cells.** *Nature* 2007, **448**:313-317.
10. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553-560.
11. Feng B, Jiang J, Kraus P, Ng JH, Heng JC, Chan YS, Yaw LP, Zhang W, Loh YH, Han J, et al: **Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb.** *Nat Cell Biol* 2009, **11**:197-203.
12. Cho HJ, Lee CS, Kwon YW, Paek JS, Lee SH, Hur J, Lee EJ, Roh TY, Chu IS, Leem SH, et al: **Induction of pluripotent stem cells from adult somatic cells by protein-based reprogramming without genetic manipulation.** *Blood*, **116**:386-395.
13. Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, Plath K: **Role of the murine reprogramming factors in the induction of pluripotency.** *Cell* 2009, **136**:364-377.
14. Chen J, Liu J, Han Q, Qin D, Xu J, Chen Y, Yang J, Song H, Yang D, Peng M, et al: **Towards an optimized culture medium for the generation of mouse induced pluripotent stem cells.** *J Biol Chem*, **285**:31066-31072.
15. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0.** *Nat Genet* 2006, **38**:500-501.

**Table 1.** Top genes overexpressed in the MAF signature.

<b>Rank</b>	<b>Gene</b>	<b>Rank</b>	<b>Gene</b>
1	COL11A1	33	ENTPD4 /// LOXL2
2	THBS2	34	COL6A3
3	COL10A1	35	MXRA5
4	COL5A2	36	MFAP5
5	INHBA	37	NUAK1
6	LRRC15	38	RAB31
7	COL5A1	39	TIMP3
8	VCAN	40	CRISPLD2
9	FAP	41	ITGBL1
10	COL1A1	42	CDH11
11	MMP11	43	TMEM158
12	POSTN	44	SPOCK1
13	COL1A2	45	SFRP4
14	ADAM12	46	SERPINF1
15	COL3A1	47	DCN
16	LOX	48	C7orf10
17	FN1	49	COPZ2
18	AEBP1	50	NOX4
19	SULF1	51	EDNRA
20	FBN1	52	ACTA2
21	ASPN	53	PDGFRB
22	SPARC	54	RCN3
23	CTSK	55	SNAI2
24	TNFAIP6	56	AMACR ///C1QTNF3
25	HNT	57	COMP
26	EPYC	58	LGALS1
27	MMP2	59	THY1
28	PLAU	60	PCOLCE
29	GREM1	61	COL6A2
30	BGN	62	GLT8D2
31	OLFML2B	63	NID2
32	LUM	64	PRRX1

**Table 2.** MSigDB results that contains all three *COL11A1*, *INHBA*, *THBS2* genes.

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p value 
<a href="#">TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_VS_DUCTAL_NORMAL_UP [67]</a>	Genes up-regulated in lobular carcinoma vs normal ductal breast cells.	3		2.52 e <sup>-8</sup>
<a href="#">VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_UP [167]</a>	Up-regulated genes distinguishing between two subtypes of gastric cancer: advanced (AGC) and early (EGC).	3		4.01 e <sup>-7</sup>
<a href="#">SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_VE_UP [355]</a>	Genes up-regulated in invasive ductal carcinoma (IDC) relative to ductal carcinoma in situ (DCIS, non-invasive).	3		3.89 e <sup>-6</sup>
<a href="#">DELYS_THYROID_CANCER_UP [400]</a>	Genes up-regulated in papillary thyroid carcinoma (PTC) compared to normal tissue.	3		5.57 e <sup>-6</sup>
<a href="#">BROWNE_HCMV_INFECTION_48HR_DN [514]</a>	Genes down-regulated in primary fibroblast cell culture after infection with HCMV (AD169 strain) at 48 h time point that were not down-regulated at the previous time point, 24 h.	3		1.18 e <sup>-5</sup>

**Table 3.** Top 45 differentially expressed genes ranked by the log-average of *P* values of ESC vs. MEF in seven mouse datasets

Rank	Gene	log.avg	MAF?	Rank	Gene	log.avg	MAF?
1	KIF26B	7.605591		24	THBS2	5.886092	<b>MAF</b>
2	ZEB2	7.466951		25	COL5A1	5.831077	<b>MAF</b>
3	POSTN	7.141856	<b>MAF</b>	26	CXCL12	5.806576	
4	MXRA7	6.537514		27	RHOJ	5.77863	
5	COL11A1	6.446498	<b>MAF</b>	28	TGFB3	5.756418	
6	RBMS3	6.398836		29	FARP1	5.744399	
7	FBN1	6.288468	<b>MAF</b>	30	SLC24A3	5.708312	
8	A730054J21RIK	6.249052		31	COL1A2	5.706657	<b>MAF</b>
9	STX2	6.229643		32	GHR	5.675805	
10	COL1A1	6.208823	<b>MAF</b>	33	TMEM176A	5.650911	
11	TMEM176B	6.178267		34	MSRB3	5.644383	
12	WISP1	6.153605		35	TWIST2	5.624833	
13	LOXL2	6.100156	<b>MAF</b>	36	LSP1	5.617865	
14	PRRX1	6.08831	<b>MAF</b>	37	OGN	5.616873	
15	TMEM119	6.078214		38	PTX3	5.595969	
16	CMTM3	6.07064		39	COL3A1	5.590278	<b>MAF</b>
17	P4HA3	6.068265		40	NCAM1	5.57965	
18	MMP14	6.054013		41	FKBP10	5.571953	
19	SRPX	6.045175		42	ASPN	5.560051	<b>MAF</b>
20	ALDH1L2	6.007391		43	CDH11	5.525121	<b>MAF</b>
21	PDGFRB	5.980855	<b>MAF</b>	44	SERPINF1	5.522322	<b>MAF</b>
22	SH3PXD2B	5.960571		45	IGF1	5.520588	
23	SGCB	5.954762					



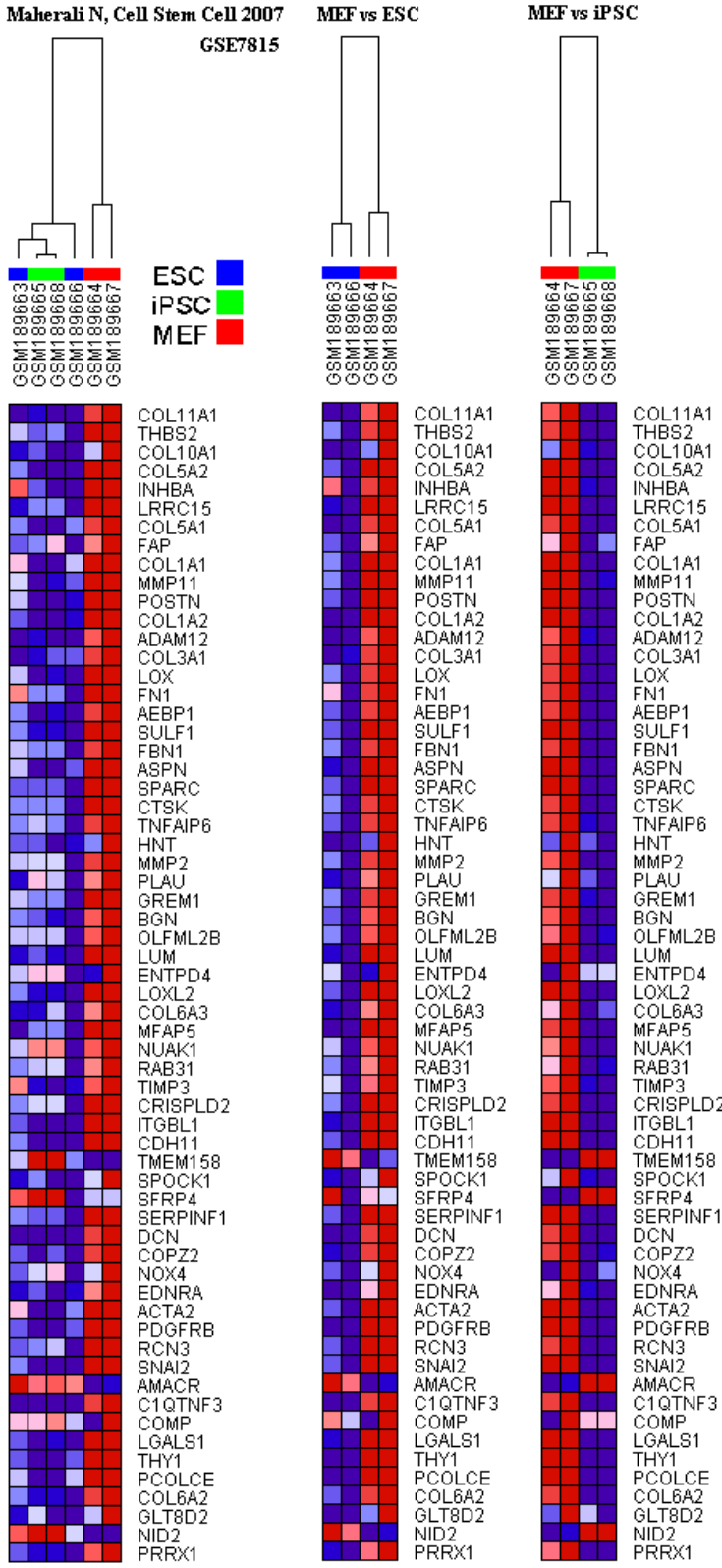


Figure 1a

Okita K, Nature 2007  
GSE7841

MEF vs ESC

MEF vs iPSC

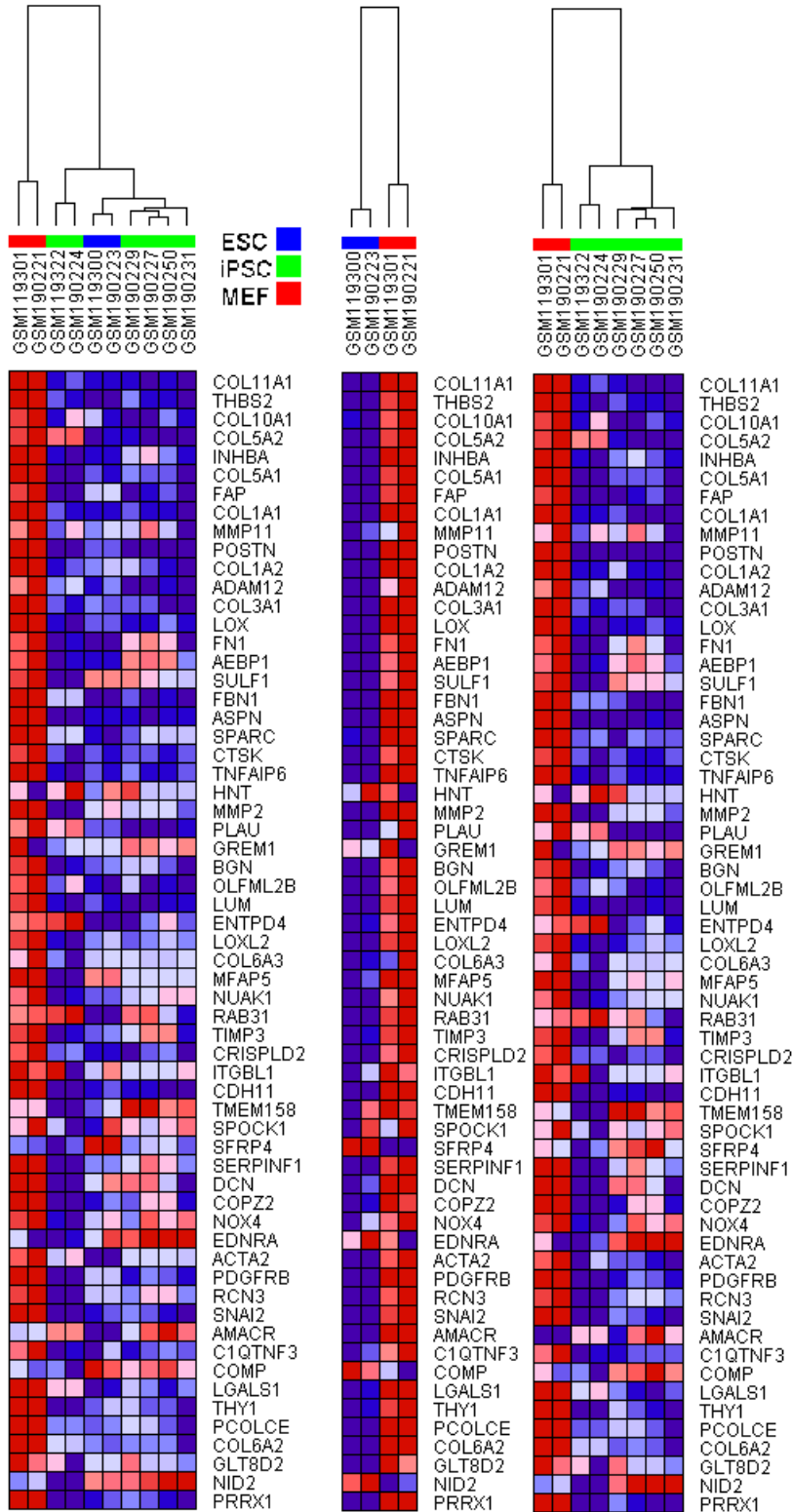


Figure 1b

Mikkelsen TS, Nature 2007  
GSE8024

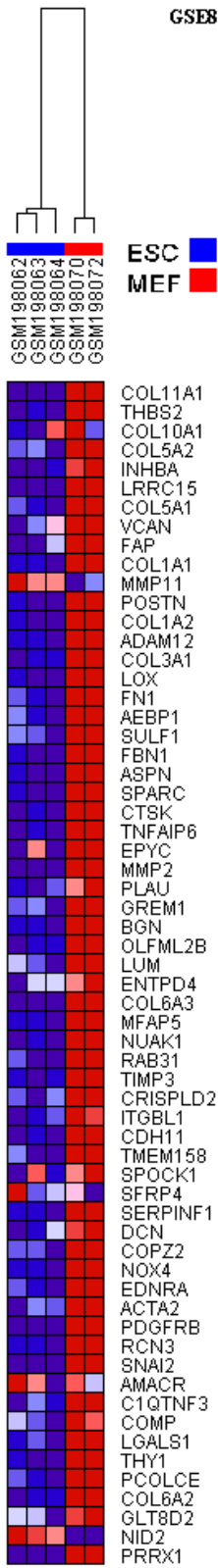


Figure 1c

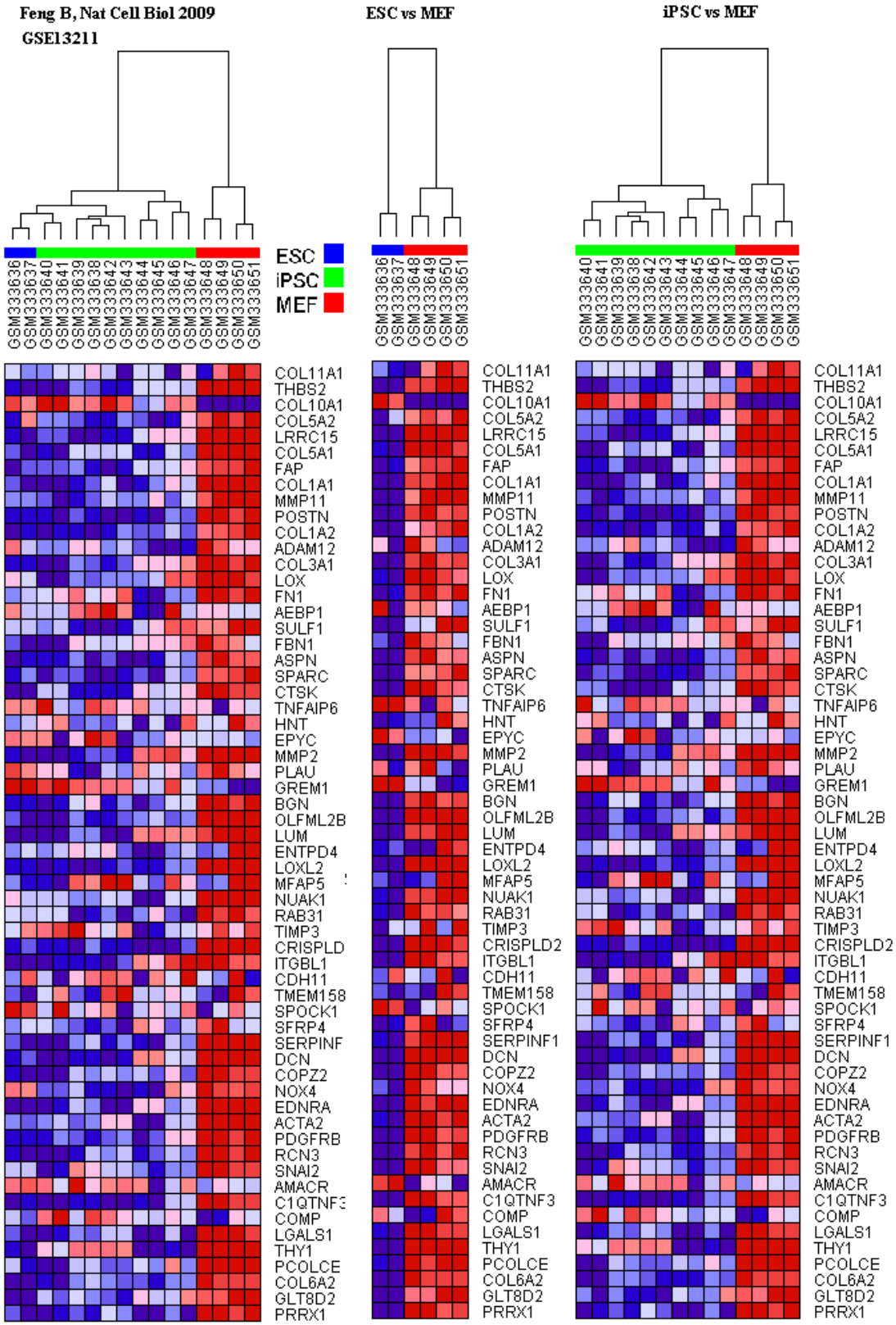


Figure 1d

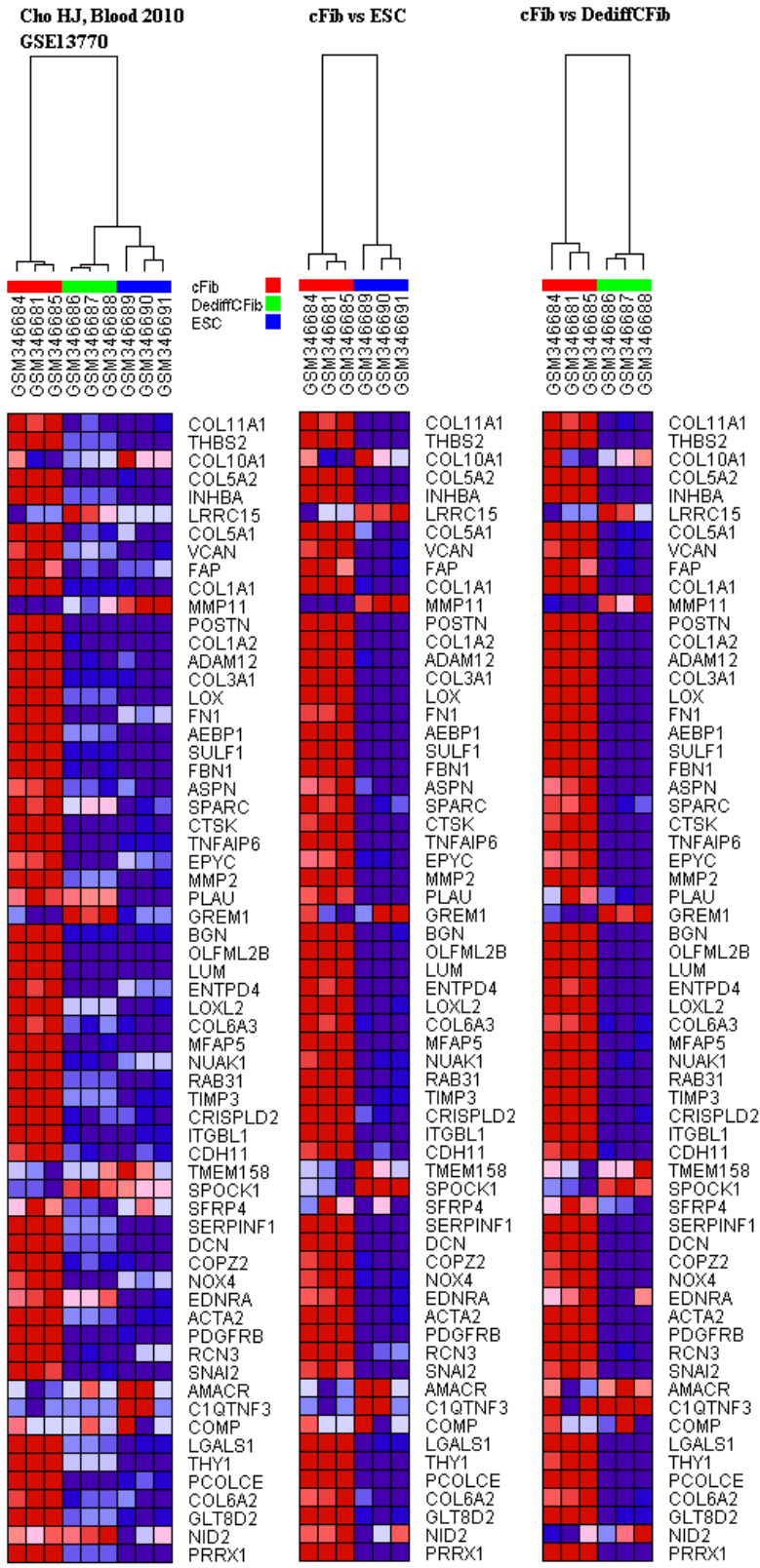


Figure 1e

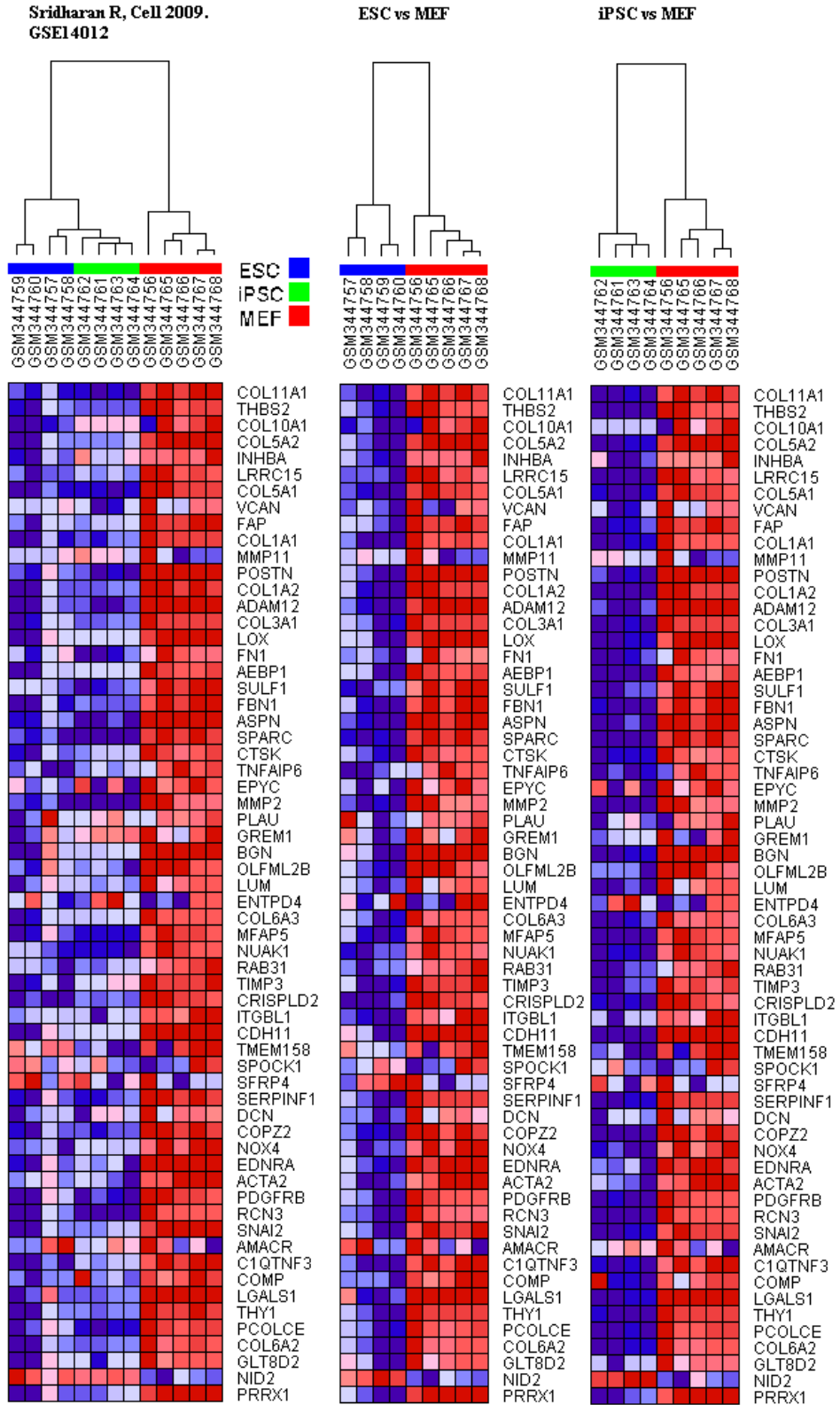


Figure 1f

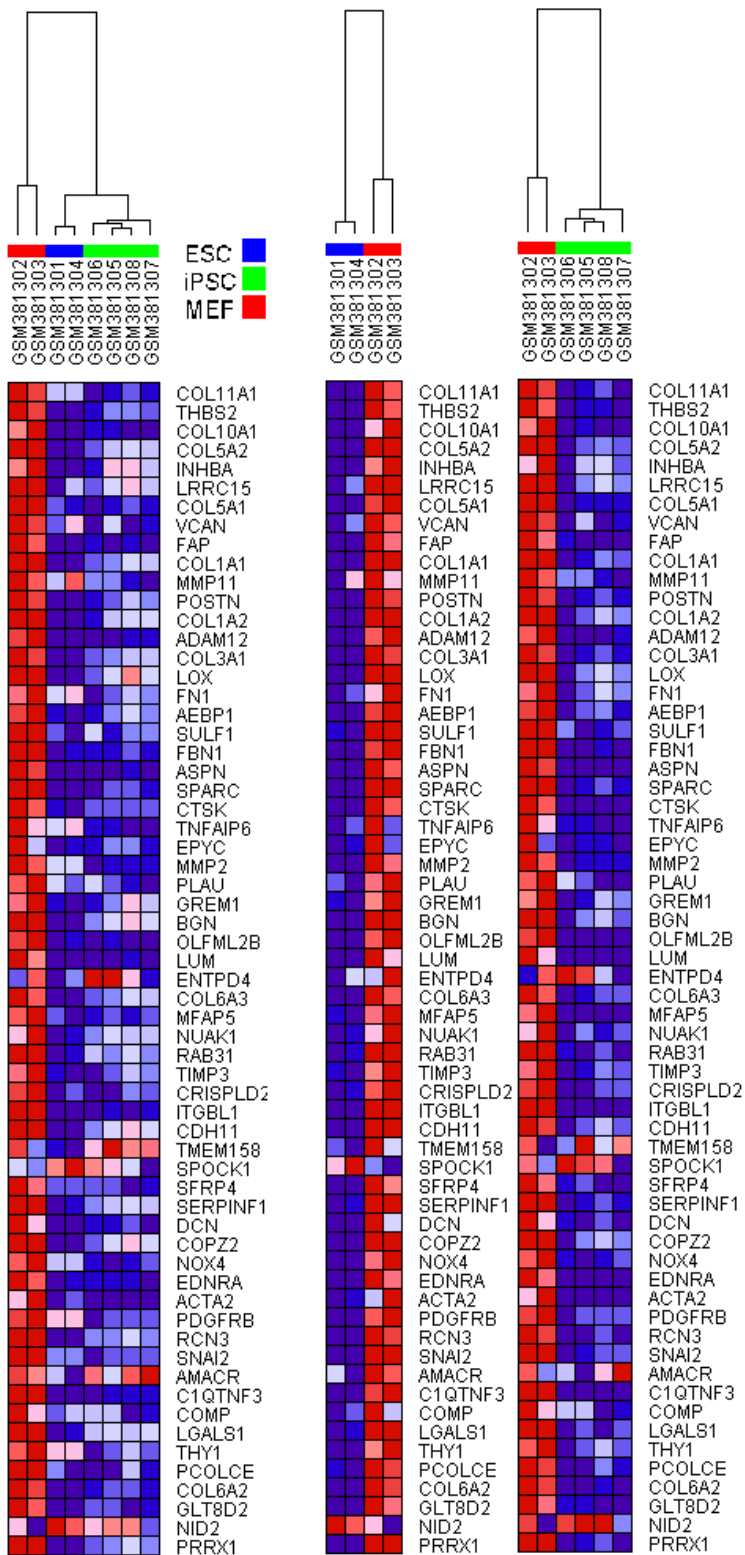


Figure 1g