# Testing the Gene Expression Classification of the EMT Spectrum

Dongya Jia[1,2,*], Jason T. George[1,3,7,*], Satyendra C. Tripathi[8], Deepali L Kundnani[9], Mingyang Lu[10], Samir M. Hanash[8,9], José N. Onuchic[1,4,5,6], Mohit Kumar Jolly[1,11,#], Herbert Levine[1,3,5,6,#]

[1]Center for Theoretical Biological Physics, [2]Program in Systems, Synthetic and Physical Biology [3]Department of Bioengineering, [4]Department of Chemistry, [5]Department of Biosciences, [6]Department of Physics and Astronomy, Rice University, Houston, TX 77005;
[7]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX 77030;
[8]Department of Clinical Cancer Prevention, [9]Red and Charline McCombs Institute for the Early Detection and Treatment of Cancer, University of Texas MD Anderson Cancer Center, Houston, TX, USA
[10]The Jackson Laboratory, Bar Harbor, ME.
[11]Current Address: Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India.

[*]These authors contributed equally.

[#]To whom correspondence should be addressed:

Mohit Kumar Jolly (mkjolly@iisc.ac.in) and Herbert Levine (herbert.levine@rice.edu)

## Supplementary Information

### 1. Construction of the gene regulatory network of EMT

The gene regulatory network of EMT (**Fig. 1A**) is an extended version of the one used in our previous work [1]. The experimental evidence for the added genes and regulatory links are listed in **Table S1** and the experimental evidence for other genes and links can be found in [1].

### 2. E and M samples purification

RACIPE generates many gene expression samples (each of them being a steady state solution of a given set of parameters and initial conditions) and clusters them into four classes, which can be mapped on to E, hybrid E/M and M phenotypes. In order to isolate RACIPE-generated samples that are highly epithelial, E samples characterized by CDH1$^{high}$/CLDN7$^{high}$/OVOL2$^{high}$/GRHL2$^{high}$/ΔNp63α$^{high}$ are selected as purified E samples. These represent a purified subset selected from the distribution of all E samples and represent a signature for a maximally epithelial phenotype. Similarly, purified M samples are identified by selecting those with VIM$^{high}$/ZEB1$^{high}$/ZEB2$^{high}$/FOXC2$^{high}$/TGF-$\beta^{high}$.

### 3. Prediction of the mixture proportions for Mixtures of purified E and M samples

In order to assess the ability of the EMT scoring metric to resolve relative proportions of epithelial and mesenchymal signatures from a mixture, we analyzed signatures for mixtures of varying epithelial and mesenchymal proportions. The mean EMT metric-estimated proportions [E, M] in five mixtures are [56%,

44%], [46%, 54%], [65%, 35%], [25%, 75%], [83%, 17%], with respective mean error [5.6%, 6.0%, 5.1%, 4.8%, 3.2%] across replicates, which closely estimate their respective true proportions [50%, 50%], [40%, 60%], [60%, 40%], [20%, 80%], [80%, 20%].

## Supplementary Tables

**Table S1. Experimental evidence for the regulatory links of OVOL2, GRHL2, ΔNp63α and CLDN7.**

| Regulation | References |
|---|---|
| Mutual inhibition between OVOL2 and ZEB1 | [2] |
| Activation of OVOL2 by GRHL2 | [3,4] |
| Mutual Inhibition between ZEB1 and GRHL2 | [5] |
| Activation of CDH1 by OVOL2 | [6] |
| Self-inhibition of OVOL2 | [7] |
| Inhibition of VIM by OVOL2 | [7] |
| Inhibition of SNAI2 by OVOL2 | [7] |
| Inhibition of ΔNp63α by OVOL2 | [7] |
| Inhibition of TGF-$\beta$ by OVOL2 | [8] |
| Inhibition of CLDN7 by ZEB1 | [9] |
| Activation of CDH1 by CLDN7 | [10] |
| Inhibition of SNAI2 by CDH1 | [11] |
| Activation of SNAI2 by ΔNp63α | [12] |
| Activation of miR-205 by ΔNp63α | [12] |
| Mutual activation between ΔNp63α and GRHL2 | [13] |

**Table S2. Prediction error for RACIPE-generated gene expression data**

**A**

EMT Metric Predictions: RACIPE Data

| Category | Binned as E | Binned as E/M | Binned as M |
|---|---|---|---|
| **E** | 45.06% | 54.94% | 0.00% |
| **E/M I** | 25.69% | 74.25% | 0.05% |
| **E/M II** | 0.21% | 95.12% | 4.67% |
| **M** | 0.02% | 28.72% | 71.25% |
| **Diagnostic Accuracy: 60.83%** | | | |

**B**

EMT Metric Predictions: Purified RACIPE Data

| Category | Binned as E | Binned as E/M | Binned as M |
|---|---|---|---|
| **Pure E** | 98.95% | 1.05% | 0.00% |
| **Pure M** | 0.00% | 6.67% | 93.33% |
| **Diagnostic Accuracy: 96.57%** | | | |

**Table S3. Prediction error for mixtures**

**A**

Absolute Mixture Error: RACIPE Data

| Category | Mix 1 | Mix 2 | Mix 3 | Mix 4 | Mix 5 |
|---|---|---|---|---|---|
| **Replicate 1** | 0.002 | 0.018 | 0.008 | 0.006 | 0.000 |
| **Replicate 2** | 0.023 | 0.009 | 0.001 | 0.003 | 0.002 |
| **Replicate 3** | 0.016 | 0.006 | 0.013 | 0.006 | 0.000 |
| **Average** | **0.013** | **0.011** | **0.007** | **0.005** | **0.001** |

**B**

Absolute Mixture Error: Purified RACIPE Data

| Category | Mix 1 | Mix 2 | Mix 3 | Mix 4 | Mix 5 |
|---|---|---|---|---|---|
| **Replicate 1** | 0.002 | 0.000 | 0.003 | 0.002 | 0.001 |
| **Replicate 2** | 0.001 | 0.001 | 0.002 | 0.003 | 0.002 |
| **Replicate 3** | 0.001 | 0.001 | 0.006 | 0.001 | 0.001 |
| **Average** | **0.001** | **0.001** | **0.004** | **0.002** | **0.001** |

**Supplementary Figures:**

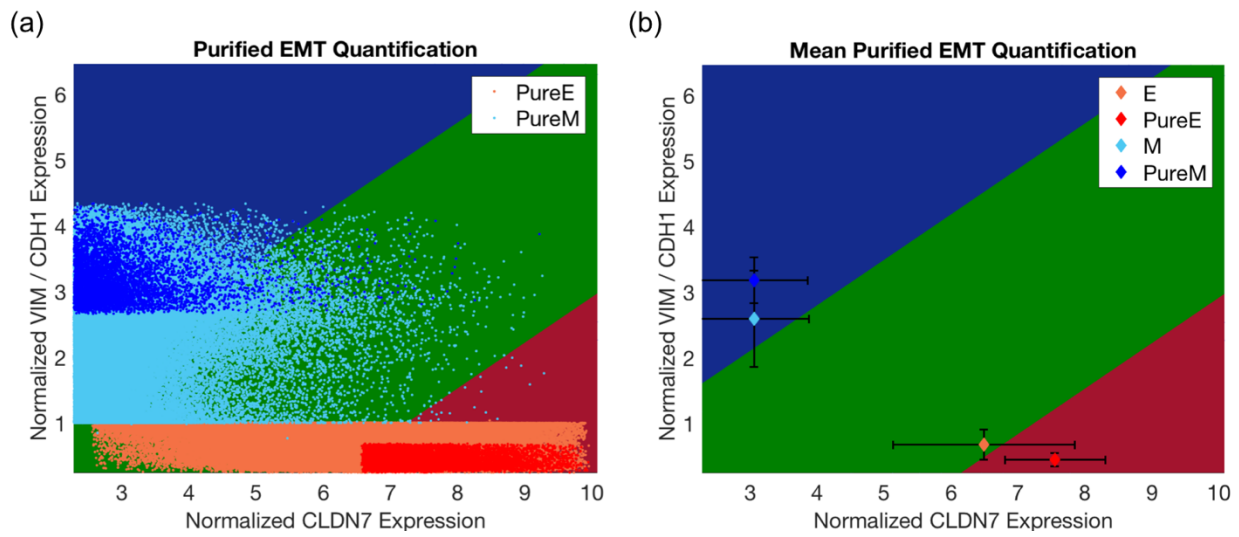


**Figure S1. Characterization of the purified E and M samples.** (a) Population level projection of RACIPE generated data onto EMT metric predictor space. VIM, CDH1, and CLDN7 expression values for each sample generated from RACIPE are plotted for various populations as categorized by RACIPE (red corresponding to E, green corresponding to hybrid E/M I and hybrid E/M II, and blue corresponding to Mesenchymal) and compared to NCI-60 gene expression training data (black). (b) Individual population members are plotted for each RACIPE-generated category as in Fig. 2A (salmon corresponding to E, light green to hybrid E/M I, magenta to hybrid E/M II, and sky blue to M). Purified E (resp. M) samples are represented by dark red (resp. blue) subsets of their respective populations. Black error bars represent standard deviations along each dimension for all samples in each average.

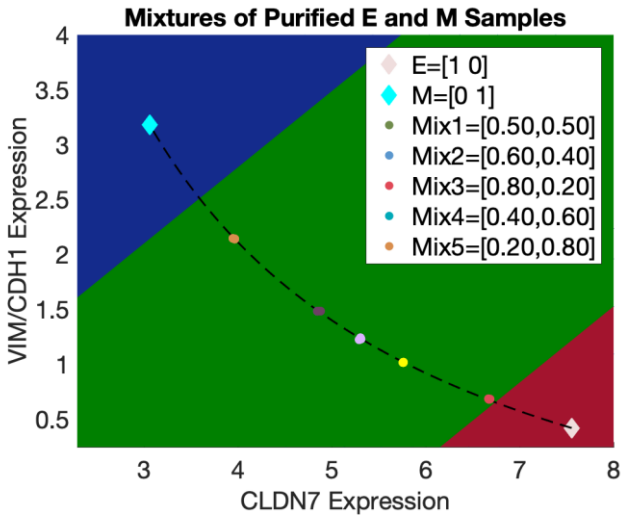


**Figure S2. Epithelial and mesenchymal subpopulation estimation by the EMT metric.** Predicted mixture proportions contained in each mixture of purified E and M samples are reported in the figure legends and predicted based on their projection to the curve of convex combinations for each non-mixed sample pair. In all cases, all three replicates for one mixture are tightly distributed and therefore have significant overlap.

## Reference


[1] Huang B, Lu M, Jia D, Ben-Jacob E, Levine H and Onuchic J N 2017 Interrogating the topological robustness of gene regulatory circuits by randomization *PLOS Comput. Biol.* **13** e1005456

[2] Roca H, Hernandez J, Weidner S, McEachin R C, Fuller D, Sud S, Schumann T, Wilkinson J E, Zaslavsky A, Li H, Maher C A, Daignault-Newton S, Healy P N and Pienta K J 2013 Transcription Factors OVOL1 and OVOL2 Induce the Mesenchymal to Epithelial Transition in Human Cancer *PLOS ONE* **8** e76773

[3] Aue A, Hinze C, Walentin K, Ruffert J, Yurtdas Y, Werth M, Chen W, Rabien A, Kilic E, Schulzke J-D, Schumann M and Schmidt-Ott K M 2015 A Grainyhead-Like 2/Ovo-Like 2 Pathway Regulates Renal Epithelial Barrier Function and Lumen Expansion *J. Am. Soc. Nephrol. JASN* **26** 2704–15

[4] Walentin K, Hinze C, Werth M, Haase N, Varma S, Morell R, Aue A, Pötschke E, Warburton D, Qiu A, Barasch J, Purfürst B, Dieterich C, Popova E, Bader M, Dechend R, Staff A C, Yurtdas Z Y, Kilic E and Schmidt-Ott K M 2015 A Grhl2-dependent gene network controls trophoblast branching morphogenesis *Dev. Camb. Engl.* **142** 1125–36

[5] Werner S, Frey S, Riethdorf S, Schulze C, Alawi M, Kling L, Vafaizadeh V, Sauter G, Terracciano L, Schumacher U, Pantel K and Assmann V 2013 Dual Roles of the Transcription Factor Grainyhead-like 2 (GRHL2) in Breast Cancer *J. Biol. Chem.* **288** 22993–3008

[6] Kitazawa K, Hikichi T, Nakamura T, Mitsunaga K, Tanaka A, Nakamura M, Yamakawa T, Furukawa S, Takasaka M, Goshima N, Watanabe A, Okita K, Kawasaki S, Ueno M, Kinoshita S and Masui S 2016 OVOL2 Maintains the Transcriptional Program of Human Corneal Epithelium by Suppressing Epithelial-to-Mesenchymal Transition *Cell Rep.* **15** 1359–68

[7] Lee B, Villarreal-Ponce A, Fallahi M, Ovadia J, Sun P, Yu Q-C, Ito S, Sinha S, Nie Q and Dai X 2014 Transcriptional mechanisms link epithelial plasticity to adhesion and differentiation of epidermal progenitor cells *Dev. Cell* **29** 47–58

[8] Wu R-S, Hong J-J, Wu J-F, Yan S, Wu D, Liu N, Liu Q-F, Wu Q-W, Xie Y-Y, Liu Y-J, Zheng Z-Z, Chan E-C, Zhang Z-M and Li B-A 2017 OVOL2 antagonizes TGF-β signaling to regulate epithelial to mesenchymal transition during mammary tumor metastasis *Oncotarget* **8** 39401–16

[9] Barbáchano A, Fernández-Barral A, Pereira F, Segura M F, Ordóñez-Morán P, Carrillo-de Santa Pau E, González-Sancho J M, Hanniford D, Martínez N, Costales-Carrera A, Real F X, Pálmer H G, Rojas J M, Hernando E and Muñoz A 2016 SPROUTY-2 represses the epithelial phenotype of colon carcinoma cells via upregulation of ZEB1 mediated by ETS1 and miR-200/miR-150 *Oncogene* **35** 2991–3003

[10] Lioni M, Brafford P, Andl C, Rustgi A, El-Deiry W, Herlyn M and Smalley K S M 2007 Dysregulation of claudin-7 leads to loss of E-cadherin expression and the increased invasion of esophageal squamous cell carcinoma cells *Am. J. Pathol.* **170** 709–21

[11] Schmalhofer O, Brabletz S and Brabletz T 2009 E-cadherin, beta-catenin, and ZEB1 in malignant progression of cancer *Cancer Metastasis Rev.* **28** 151–66

[12] Dang T T, Esparza M A, Maine E A, Westcott J M and Pearson G W 2015 ΔNp63α Promotes Breast Cancer Cell Motility through the Selective Activation of Components of the Epithelial-to-Mesenchymal Transition Program *Cancer Res.* **75** 3925–35

[13] Mehrazarin S, Chen W, Oh J-E, Liu Z X, Kang K L, Yi J K, Kim R H, Shin K-H, Park N-H and Kang M K 2015 The p63 Gene Is Regulated by Grainyhead-like 2 (GRHL2) through Reciprocal Feedback and Determines the Epithelial Phenotype in Human Keratinocytes *J. Biol. Chem.* **290** 19999–20008