SingScore to capture transition dynamics of transcriptional data in EMT and MET transitions

Singscore is a simple single-sample gene signature scoring method that uses rank-based statistics to analyze the sample's gene expression profile (7). It scores the expression activities of gene sets at a single-sample level. Genes are ranked by increasing m-RNA abundance in a sample transcriptome after correction of technical within-sample bias. For undirected gene signatures, SingScores are calculated by the following formula:

$$S_i = (\frac{\sum_g \hat{R_i^g}}{N_i}) \tag{1}$$

$$\bar{S_i} = \frac{(S_i - S_{min,i})}{S_{max,i} - S_{min,i}} \tag{2}$$

where S_i is the score for sample i against the undirected gene-set, $\hat{R_i^g}$ is the absolute median-centered rank of gene g in the undirected gene set; i.e.

$$\hat{R_i^g} = |R_i^g - ceil(\frac{N_{total}}{2})| \tag{3}$$

where ceil represents the ceiling function.
N_i is the number of genes in the undirected gene set that are observed within the data, $\bar{S_i}$ is the normalised score for sample i against the undirected gene-set, and $S_{min,i}$ and $S_{max,i}$ are the theoretical minimum and maximum mean ranks obtained from the arithmetic series expansion:

$$S_{min,i} = \frac{(ceil(\frac{N_{dir,i}}{2}) + 1)}{2} \tag{4}$$

$$S_{max,i} = \frac{(N_{total,i} - ceil(\frac{N_{dir,i}}{2}) + 1)}{2} \tag{5}$$

SingScores of an epithelial and a mesenchymal gene signature list were calculated for each sample at a particular time point undergoing transition. Resulting scores were plotted on an epithelial vs mesenchymal graph to observe the transition in scores on both axes across time points.

SingScore plots were made to describe EMT for TGF-β1 treated A549 Human Lung Cancer Cells. SingScores of an epithelial gene signature list and a mesenchymal gene signature list were calculated separately, for three replicates each at 9 time points - 0,0.5,1,2,4,8,16,24 and 72h. As expected, the SingScore for the mesenchymal genes increased while that of the epithelial genes decreased from 0 to 72h, indicating a clear transition from epithelial to mesenchymal character across time points.

The same procedure was applied to describe MET in prostate adenocarcinoma, LNCaP cells with inducible and reversible expression of SNAI1 and SNAI2. SingScores of an epithelial gene signature list and a mesenchymal gene signature list were calculated separately, for three replicates each at 4 time points - EMT.D5, MET.D3, MET.D5 and MET.D20 for all three groups - GFP (no EMT induced), SNAI1 (EMT induced by SNAI1) and SNAI2 (EMT induced by SNAI2).The GFP group shows mixed mesenchymal and epithelial character, and serves as a good control for the SNAI1 and SNAI2 group. There is a clear transition from mesenchymal to epithelial character in both these groups, across time points. The mesenchymal character of SNAI1 group is significantly greater than that of the SNAI2 group.

Surprisal Analysis

Surprisal analysis was employed to make a comparative study of the dynamics of epithelial to mesenchymal and mesenchymal to epithelial transitions. Surprisal analysis is an information-theoretical analysis technique that integrates principles of thermodynamics and maximal entropy. It is based on the premise that time-evolving biological systems are subject to the same thermodynamic considerations used to describe inanimate non-equilibrium physical systems. At any point in time, a dynamic biological system is in a state of maximal entropy subject to constraints. Surprisal analysis identifies these constraints by describing the entire transcriptional data in terms of very few, three or four patterns. A pattern(representing a constraint) is a set of genes that vary in concert (i.e. all transcription levels are time-dependent in the same way) and can be assigned a specific biological phenotypic role.

Surprisal analysis was applied as described previously (1-4). The measured expression level of a m-RNA transcript i at time t - $X_i(t)$ is expressed as a sum of terms:

$$\ln X_i(t) = \ln X_i^o - \sum_{\alpha=1} \lambda_\alpha(t) G_{\alpha i} \tag{6}$$

$$= - \sum_{\alpha=0} \lambda_\alpha(t) G_{\alpha i} \tag{7}$$

$$\ln X_i^o = -\lambda_0 G_{0i} \tag{8}$$

where X_i^o is the expression level of transcript i in the steady or balanced state . $\lambda_\alpha(t)$ is the Lagrange multiplier of constraint α at time t. It is the potential that equals the weight of the phenotype α at time t. $G_{\alpha i}$ represents the weight of the gene i in the transcription pattern α, or how much transcript i contributes to the constraint α. Further details about the implementation and the thermodynamic derivation of equation (6) can be found in the supplementary appendices and main texts of (1-4) and (8).

m-RNA expression levels during EMT and MET can be well-described by the first two patterns, $\alpha = 1, 2$. $\alpha = 0$ represents the zeroth or baseline phenotype which remains relatively unchanged across time points and is most correlated with cellular networks that regulate and maintain cellular homeostatis machinery in any transition.

Surprisal Analysis applied to describe EMT for TGF-β1 treated A549 Human Lung Cancer Cells

Surprisal analysis was applied to the gene microarray expression profiles, collected in triplicates, at time points 0,0.5,1,2,4,8,16,24 and 72h during a TGF-β1 induced EMT transition in A549 cells (5). Instead of using all the genes or the entire transcriptional data to identify the principal phenotypes or patterns, only a particular set of genes - an epithelial-mesenchymal gene signature list was used to describe the transition dynamics. There was significant reduction in error bars when this specific list of genes was used to perform surprisal analysis compared to when entire transcriptional data of all genes was used. The qualitative trends in the time-dependence of the phenotypes was also conserved, implying that an epithelial-mesenchymal gene signature list is sufficient to describe all major changes happening in EMT dynamics by surprisal analysis. The composition of the two major phenotypic patterns $\alpha = 1, 2$ was quantified in terms of probabilities of epithelial and mesenchymal genes getting downregulated and upregulated in them respectively.

The transcriptional data of the the epithelial-mesenchymal gene signature list can be reduced to three major patterns. Each pattern is associated with a specific biological phenotypic role in the underlying process. The patterns are ranked according to the extent of deviation from the balance state caused by them. The major phenotype $\alpha = 1$ has the largest value of Lagrange multiplier $\lambda_1(t)$ across time points and is responsible for most free-energy changes during the transition. The phenotype $\alpha = 1$ shows a sharp monotonic rise in weight ($\lambda_1(t)$) from 0 to 24h, after which this rise slows down at 72h. The $\lambda_1(t)$ sign changes at around 8 h which has been described previously in (8) as the point where the transition from the epithelial to mesenchymal state occurs and is correlated with the point of minimum free energy of the transcripts contributing most to the $\alpha = 1$ phenotype. The steady monotonically increasing trend in $\lambda_1(t)$ is consistent with the trends found in its top contributing genes, and correlates well with the fact that it is the major phenotypic pattern contributing to the epithelial to mesenchymal transition. It is calculated that mesenchymal genes have a greater probability of being upregulated in the $\alpha = 1$ phenotype, while downregulated genes have a greater probability of being epithelial. This trend is conserved across replicates, hence is robust. This trend falls when we move from the $\alpha = 1$ to the $\alpha = 2$ phenotype.

The second most important phenotype governing the epithelial to mesenchymal transition, the $\alpha = 2$ pattern shows a steady increase in weight($\lambda_2(t)$) from 0 to 8h, which then falls steadily from 16 to 72h. There are two sign changes in ($\lambda_2(t)$), one at 2h and the other at 24h. These sign changes correspond to a transition from the epithelial state to a maturation state (2h) and then from the maturation state to the mesenchymal state (24h) according to (8). The trends found in the top contributing genes in the $\alpha = 1$ pattern are not found in the $\alpha = 2$ pattern. Only downregulated genes still have a somewhat high probability of being epithelial, though this probability values are significantly smaller than those observed in the $\alpha = 1$ pattern.

Surprisal Analysis applied to describe MET in prostate adenocarcinoma, LNCaP cells with inducible and reversible expression of SNAI1 and SNAI2

Cells were treated with doxycycline to induce a SNAI1- or SNAI2-mediated EMT for 5 days and then subsequently removed for 3, 5, and 20 days to allow for MET. LNCaP-iGFP cells were also generated for control purposes and were subjected to the same treatment regime (6). Surprisal analysis was applied to the gene microarray expression profile, collected in triplicates, at time points EMT.D5 (day 5 of EMT induction), MET.D3 (day 3 of EMT withdrawal), MET.D5 (day 5 of EMT withdrawal) and MET.D20 (day 20 of EMT withdrawal) for all three groups - GFP (no EMT induced), SNAI1 (EMT induced by SNAI1) and SNAI2 (EMT induced by SNAI2). The GFP group serves as a good control to compare the results of SNAI1 and SNAI2 against. As before, only the epithelial-mesenchymal gene signature list was used to perform surprisal analysis, and it could efficiently capture the dynamics of the transition. The transition process could be well-described by three major transcriptional patterns in all three groups.

In the SNAI1 group (EMT induced by SNAI1), the major transcription pattern governing the mesenchymal to epithelial transition - the $\alpha = 1$ pattern shows a monotonically decreasing trend from time points EMT.D5 to MET.D3. There is a very slight increase in weight ($\lambda_1(t)$) towards the end of the transition from MET.D5 to MET.D20. There is a sign change in $\lambda_1(t)$ just before MET.D3, which may be related to the point of transition from the mesenchymal to the epithelial state, like in the A549 EMT case described above. It is found that mesenchymal genes have a really high probability of being upregulated in the $\alpha = 1$ pattern, while downregulated genes have a really high probability of being epithelial. But due to the monotonic decreasing trend of the $\alpha = 1$ pattern, this implies that mesenchymal genes have a really high probability of being downregulated in the biological process, while upregulated genes have a really high probability of being epithelial, which is consistent with a mesenchymal to epithelial transition and is corroborated by SingScore plots on mesenchymal and epithelial axes across time points. The trends in the top contributing genes to the $\alpha = 1$ pattern are consistent across replicates, hence robust.

The second most important transcription pattern $\alpha = 2$ in the SNAI1 group shows a decrease in weight ($\lambda_2(t)$) from EMT.D5 to MET.D3, after which it increases steadily till MET.D20. There are two sign changes in $\lambda_2(t)$, the first between EMT.D5 and MET.D3 and the second between MET.D5 and MET.D20. The trends found in the top contributing genes in the $\alpha = 1$ pattern fall in significance in the $\alpha = 2$ pattern. Downregulated genes still have a greater probability of being epithelial, but these probability values are significantly smaller than those observed in the $\alpha = 1$ pattern.

The trends observed from surprisal analysis in the SNAI2 group (EMT induced by SNAI2) are quite similar to those found in the SNAI1 group. The major transcription pattern $\alpha = 1$ shows a monotonic decrease in weight ($\lambda_1(t)$) from EMT.D5 to MET.D5, followed by a slight increase from MET.D5 to MET.D20. This slight increase from MET.D5 to MET.D20 is consistent with a SingScore plot, which shows an unexpected increase in mesenchymal scores from MET.D5 to MET.D20. There is a sign change in $\lambda_1(t)$ just before MET.D3, similar to the SNAI1 group. Just like in the SNAI1 group, mesenchymal genes have a really high probability of being upregulated in the $\alpha = 1$ pattern, while downregulated genes have a really high probability of being epithelial. But due to the monotonic decreasing trend of the $\alpha = 1$ pattern, this implies that mesenchymal genes have a really high probability of being downregulated in the biological process, while upregulated genes have a really high probability of being epithelial, which is consistent with a mesenchymal to epithelial transition and is corroborated by SingScore plots. The trends are conserved across replicates.

The $\alpha = 2$ pattern in the SNAI2 group shows a decrease in weight ($\lambda_2(t)$) from EMT.D5 to MET.D3, followed by a monotonic increase from MET.D3 to MET.D20. $\lambda_2(t)$ undergoes sign changes at the same points as in the SNAI1 group, and there are similar trends in the top contributing genes to the pattern.

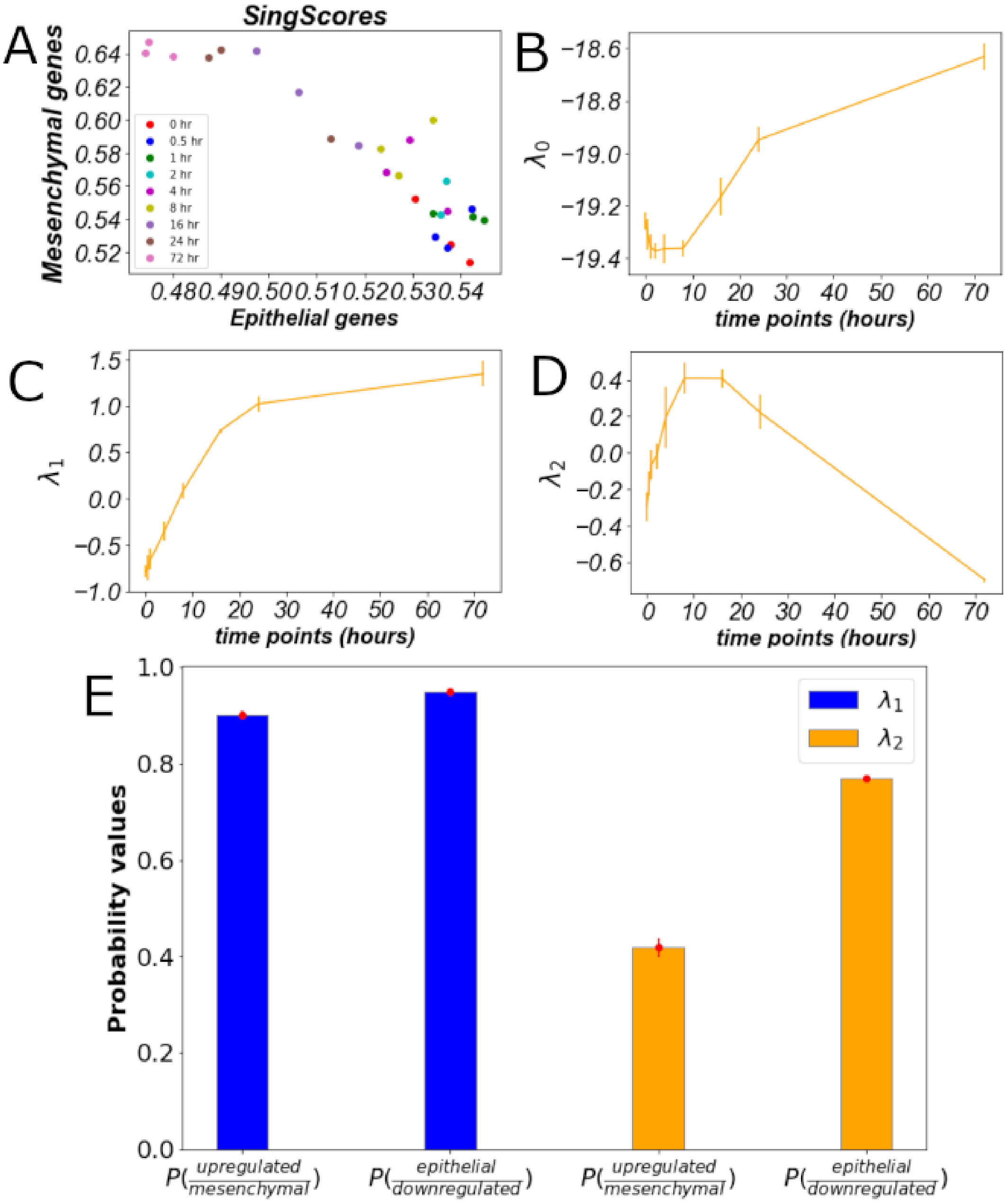

Figure 1. Transition dynamics of EMT in A549 Lung Cancer Cells.(A) SingScores of an epithelial and a mesenchymal gene signature list plotted across time points reveals a clear transition from epithelial to mesenchymal character. (B) Surprisal analysis performed on an EMT-specific gene signature list reveals three major phenotypic patterns. Weight of λ_0 pattern plotted across time points. (C) λ_1 phenotype shows a monotonic increase in weight from t=0 to t=72 h. (D) λ_2 phenotype shows an initial increase followed by a decrease in weight across time points. (E) Conditional probability values of genes getting upregulated given they are mesenchymal and genes being epithelial given they are downregulated in λ_1 and λ_2.

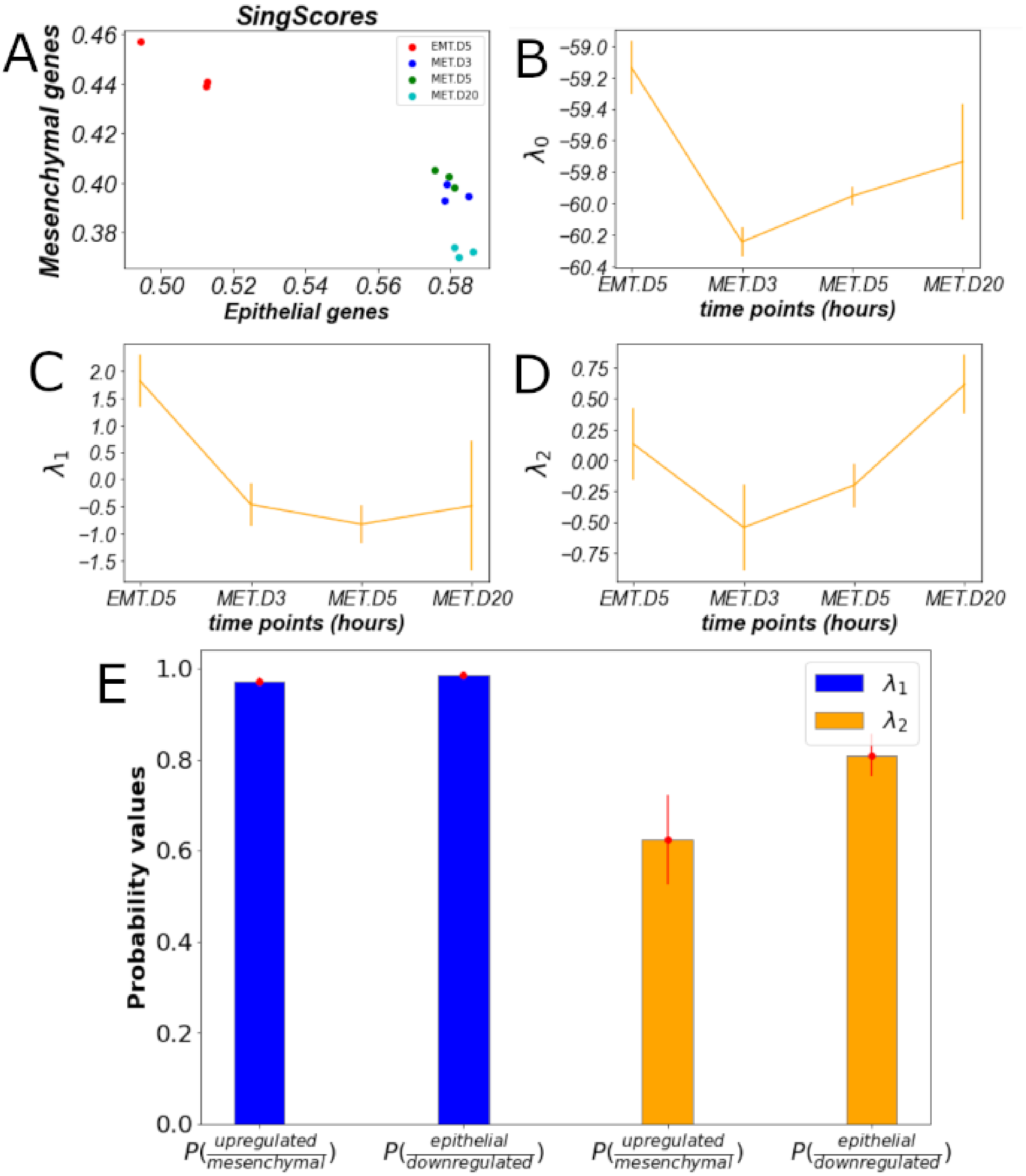

Figure 2. MET studied in LNCaP cells with inducible and reversible expression of SNAI1. (A) SingScores of an epithelial and a mesenchymal gene signature list plotted across time points reveals a clear transition from mesenchymal to epithelial character. (B) Surprisal analysis performed on an EMT-specific gene signature list reveals three major phenotypic patterns. Weight of λ_0 pattern plotted across time points. (C) λ_1 phenotype shows an almost monotonic decrease in weight throughout. (D) λ_2 phenotype shows an initial decrease followed by an increase in weight across time points. (E) Conditional probability values of genes getting upregulated given they are mesenchymal and genes being epithelial given they are downregulated in λ_1 and λ_2.

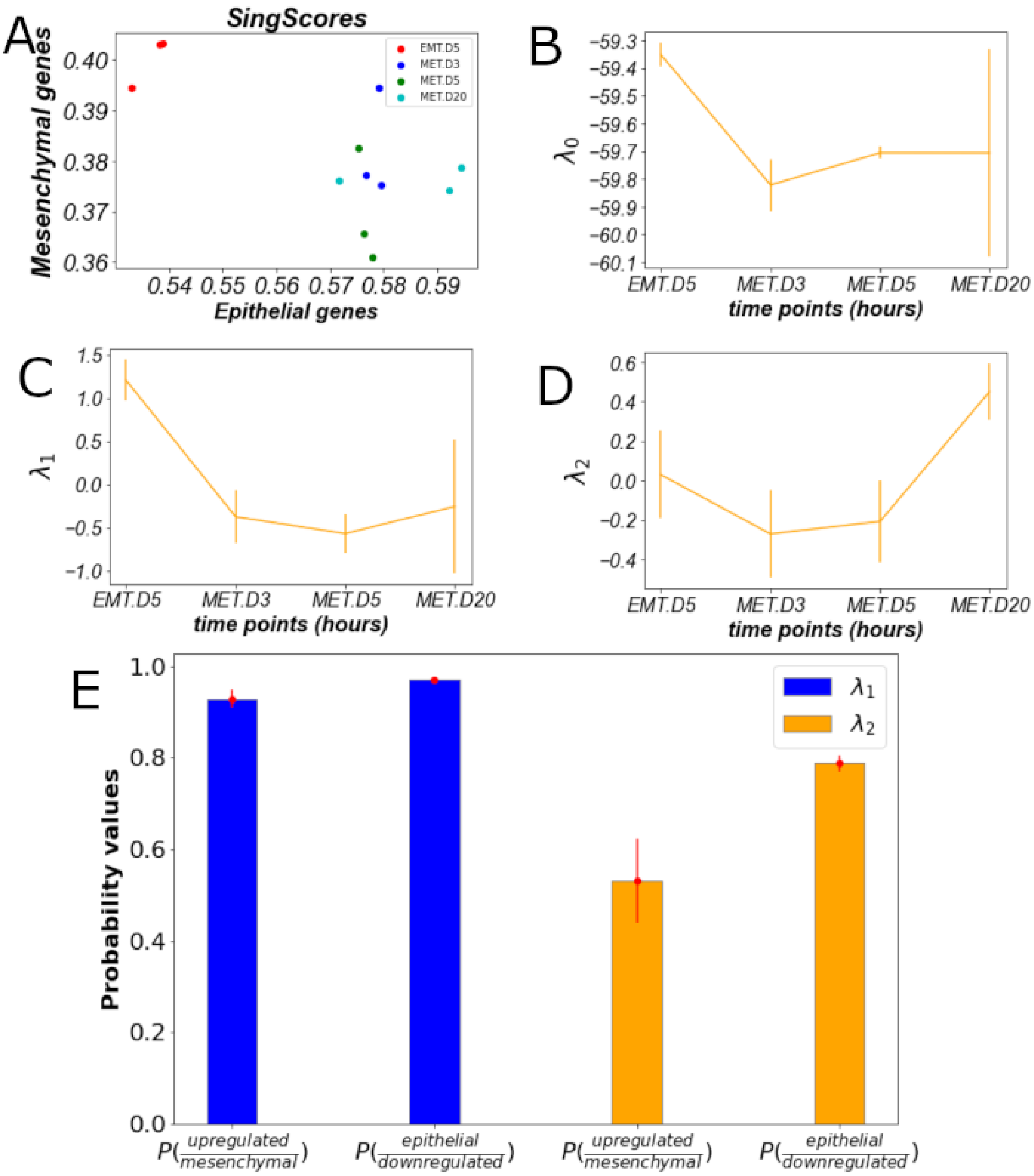

Figure 3. MET studied in LNCaP cells with inducible and reversible expression of SNAI2. (A) SingScore analysis reveals that overall mesenchymal character in SNAI2 group is lower than that of SNAI1, i.e. SNAI1 induces stronger EMT. (B) Surprisal analysis performed on an EMT-specific gene signature list reveals three major phenotypic patterns. Weight of λ_0 pattern plotted across time points. (C) λ_1 phenotype shows a nearly similar trend to that in SNAI1 group. (D) λ_2 phenotype shows an initial decrease followed by an increase in weight across time points. (E) Conditional probability values of genes getting upregulated given they are mesenchymal and genes being epithelial given they are downregulated in λ_1 and λ_2.

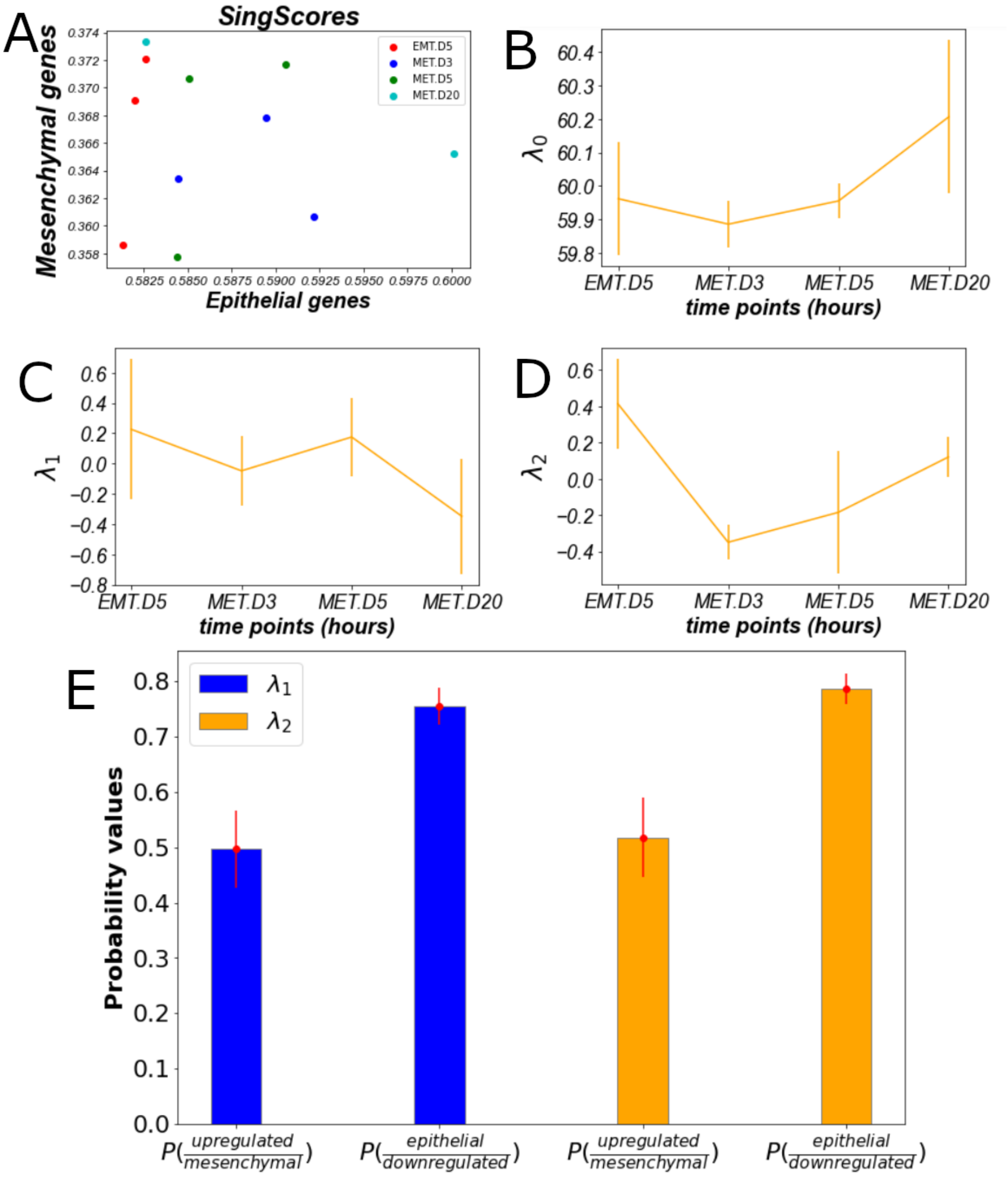

Figure 4. Surprisal Analysis performed on a control group for MET, GFP group in which no EMT was induced. (A) SingScore analysis reveals no clear transitions, as expected. (B) Surprisal analysis performed on an EMT-specific gene signature list reveals three major phenotypic patterns. Weight of λ_0 pattern plotted across time points. (C) λ_1 phenotype shows a highly non-monotonic trend. (D) λ_2 phenotype shows an initial decrease followed by an increase in weight across time points. (E) Conditional probability values of genes getting upregulated given they are mesenchymal and genes being epithelial given they are downregulated in λ_1 and λ_2.

References

1. Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA. 2010;107(22):10324–10329.

2. Kravchenko-Balasha N, et al. On a fundamental structure of gene networks in living cells. Proc Natl Acad Sci USA. 2012;109(12):4702–4707.

3. Kravchenko-Balasha N, et al. Convergence of logic of cellular regulation in different premalignant cells by an information theoretic approach. BMC Syst Biol. 2011;5:42.

4. Gross A, Levine RD. Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. PLoS One. 2013;8(4):e61554.

5. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J et al. ConceptGen: a gene set enrichment and gene set relation mapping tool. Bioinformatics 2010 Feb 15;26(4):456-63. PMID: 20007254

6. Stylianou N, Lehman ML, Wang C, Fard AT et al. A molecular portrait of epithelial-mesenchymal plasticity in prostate cancer associated with clinical outcome. Oncogene 2019 Feb;38(7):913-934. PMID: 30194451

7. Foroutan, M., Bhuva, D.D., Lyu, R. et al. Single sample scoring of molecular phenotypes. BMC Bioinformatics 19, 404 (2018).

8. Zadran S, Arumugam R, Herschman H, Phelps ME, Levine RD. Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. Proc Natl Acad Sci U S A. 2014 Sep 9;111(36):13235-40.