Title

Interpretable, scalable, and transferrable functional projection of large-scale transcriptome data using constrained matrix decomposition

Authors

Nicholas Panchy[1], Kazuhide Watanabe[2], and Tian Hong[1,3] *

[1] Department of Biochemistry and Cellular and Molecular Biology. The University of Tennessee, Knoxville. Knoxville, Tennessee, 37996, United States of America.

[2] RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan.

[3] National Institute for Mathematical and Biological Synthesis, Knoxville, Tennessee, 37996, United States of America.

To whom correspondence should be addressed to Tel: 01-865-974-3089. Fax: 01-865-974-6306. E-mail: hongtian@utk.edu.

Abstract

Large-scale transcriptome data, such as single-cell RNA-sequencing data, have provided unprecedented resources for studying biological processes at the systems level. Numerous dimensionality reduction methods have been developed to visualize and analyze these transcriptome data. In addition, several existing methods allow inference of functional variations among samples using gene sets with known biological functions. However, it remains challenging to analyze transcriptomes with reduced dimensions that are both interpretable in terms of dimensions' directionalities and transferrable to new data. In this study, we used gene set non-negative principal component analysis (gsPCA) and non-negative matrix factorization (gsNMF) to analyze large-scale transcriptome datasets. We found that these methods provide low-dimensional information about the progression of biological processes in a quantitative manner, and their performances are comparable to existing functional variation analysis methods in terms of distinguishing multiple cell states and samples from multiple conditions. Remarkably, upon training with a subset of data, these methods allow predictions of locations in the functional space using data from experimental conditions that are not exposed to the models. Specifically, our models predicted the extent of progression and reversion for cells in the epithelial-mesenchymal transition (EMT) continuum. These methods revealed conserved EMT program among multiple types of single cells and tumor samples. Finally, we provide several recommendations on the choice between the two linear methods and the optimal algorithmic parameters. Our methods show that simple constrained matrix decomposition can produce to low-dimensional information in functionally interpretable and transferrable space, and can be widely useful for analyzing large-scale transcriptome data.

Introduction

Recent developments in RNA-sequencing technology have enabled the collection of large-scale transcriptome data at high speed. For example, single-cell RNA-sequencing (scRNA-seq) data of many biological systems have been accumulating rapidly, and provide opportunities to gain insights into complex biological processes at both the systems level and the single-cell resolution. Together with the advances in experimental techniques, the recent development of computational methods, including those for dimensionality reduction, allow the visualization and analyses of high-dimensional transcriptome data in low-dimensional space. For example, t-distributed stochastic neighbor embedding (tSNE) and Uniform Manifold Approximation and Projection (UMAP) have been instrumental to tackling challenges in transcriptome data visualization and are widely used in biomedical research (Becht et al., 2019; Luecken and Theis, 2019; Stein-O'Brien et al., 2018; Van der Maaten and Hinton, 2008). However, dimensionality reduction methods usually do not provide low-dimensional space that is directly interpretable in terms of biological functions: while these approaches cluster related samples, the positioning of samples along the derived dimension does not correspond to the degree of any biological process even if a predefined gene set with similar functions is chosen before the reduction. The lack of interpretability of the dimensions makes it challenging to visualize and analyze the progression of the samples (cells) in known biologically functional space.

Existing methods for functional quantification, such as z-score and Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013), are useful for obtaining 'functional scores' with the expression levels of multiple genes involved in the same biological process. However, these methods do not have transferability in that the scoring systems obtained with one dataset cannot be used to analyze other datasets directly. This limits the utility of these methods in predicting the progress of new data points, and in studying the relationships between functional spaces in different experimental settings.

One example of cellular processes that contains crucial quantitative information is epithelial-mesenchymal transition (EMT). While extreme changes of cell fate and morphology occur in the classical

form of EMT, recent studies with cancer and fibrosis showed that partial EMT involving intermediate states are prevalent, and it may be responsible for pathogenesis (Pastushenko et al., 2018). To quantify the degree of EMT in EMT-induced cell lines and tumor samples, several previous studies analyzed transcriptomic data and their projections onto epithelial (E) and mesenchymal (M) dimensions (Cursons et al., 2018; Panchy et al., 2020; Tan et al., 2014). Recently, scRNA-seq analysis has shown that the progression of EMT is highly dependent on inducing signals and cell types (Cook and Vanderhyden, 2020). However, it remains challenging to analyze rapidly accumulating transcriptome information on EMT for obtaining biological insights across multiple conditions. Improvement of methods for reducing dimensions of expression data in a functionally meaningful manner is necessary.

In this study, we used gene set filtered variants of both non-negative principal component analysis (gsPCA) and non-negative matrix factorization (gsNMF) to analyze progression of EMT in single cells at multiple timepoints. We show that these methods describe large-scale transcriptome data of multiple EMT stages in low-dimensional and functionally interpretable space. Taking advantage of the methods' transferability, we constructed dimensionality reduction models that can predict the stages of EMT with data from timepoints that were not used for model construction. We show that these linear methods can be used to compare functional spaces across multiple experimental conditions. With a validation scheme for rigorous testing, we provide recommendations for the choice of the methods and the parametric settings. Overall, our work provides a new toolbox for analyzing large-scale transcriptome data with efficient visualization and functional quantification.

Results

gsPCA and gsNMF capture cell state progression in low dimensional functional space

To test the performance of gsPCA and gsNMF in capturing biological progression through functional space, we first used time-course datasets containing single cells treated with EMT-inducing signals for various periods of time (Cook and Vanderhyden, 2020). We used two signature gene sets whose high expressions represent the E and M states respectively (Panchy et al., 2020; Tan et al., 2014; Watanabe et al., 2019). With the expression levels of the 176 E genes and the 116 M genes covered by the data set, we first performed gsPCA and gsNMF on single-cell transcriptomes of TGF-β-treated A549 cells using two components per model for each gene set. We then projected the single-cell data from the first five time points, which represent continuous EMT progression, onto the leading dimension for each gene set. This produced two-dimensional plots with dimensions that can be viewed as the progression of cell states in the epithelial and the mesenchymal spectrums (Figures 1A and 1B). We then compared the performance to two widely used approaches: z-score and GSVA (Figure 1C and 1D) using two metrics. First, we assessed the overall correspondence with EMT progression using the variance in time explained by a linear model E and M dimensions (the coefficient of determination or R^2, see Methods). Second, we assessed the ability of scores to separate samples from individual time points using the common language effect size (*f*) (McGraw and Wong, 1992), defined here as the probability that a higher score on the dimensions corresponds properly with EMT progression when pairs of samples are taken at random from neighboring time points (i.e. higher E-score = earlier timepoint and higher M-score = later timepoint), This probability can be derived from the U-statistic of the Mann-Whitney U-test, which we used to assess the significance of separation (see Methods). We found that gsPCA and gsNMF both better explained the overall variance of time across the first five time points of EMT progression (Adjusted R^2 = 0.49 and 0.52, respectively) than z-score (Adjusted R^2 = 0.33) and GSVA (Adjusted R^2 = 0.24). Likewise, when considering neighboring time points, we found that E-scores tended to decrease and M-scores tended to increase with time of TGF-β treatment (Figure 1E), with both scores significantly separating all

neighboring time points for gsPCA and gsNMF and yielding higher *f* probabilities in all but one case (E-scores at 3 days vs 7 days, Figure 1F). This suggests that gsPCA and gsNMF not only serve as visualization methods of functional space with defined gene sets, but also describe heterogeneous cell populations containing transitional information in a rigorous fashion. More importantly, the reduced dimensions obtained with gsPCA and gsNMF can be directly used to analyze data that were not used during the derivation of these dimensions, a transferability that is not offered by previous functional projection methods. Between the two methods, we found the gsNMF performed better with regard to both overall variance (Adjusted R^2 0.52 vs 0.49) and separating time points (Figure 1F) than gsPCA. However, this requires selecting the leading dimension based directly on correlation with time of EMT progression, suggesting that gsPCA may be more reliable in a purely unsupervised setting (see Supplement Methods for a full discussion of the consistency of gsNMF and gsPCA). Nonetheless, both gsPCA and gsNMF gave consistent results when subjected to cross-validation analysis (see Supplemental Methods; Figure S1), with selecting a leading dimension from a two-component model providing the best and the most consistent explanation of variation in EMT progression.

In the next few sections, we show various utilities of these linear methods based on their transferability and high-performance features. Because gsNMF gives the best performance with the A549 EMT data set, our discussion will focus on results obtained with gsNMF. The results for similar performance obtained using gsPCA are included in supplementary materials. It should be also noted that we used the time information to evaluate the performance with the first data set, but both methods can be used to analyze and visualize data without time labels (see examples later).

Prediction of cell states with data from new conditions

The transferability of gsPCA and gsNMF methods allows the projection of new high-dimensional data points onto previously derived functional dimensions. Similarly, these methods can be used to derive

functional dimensions with partial information of the biological process in terms of its stages. To show the predictive power of gsNMF, we removed samples from the 0-day, 1-day and 7-day (including revertant) time points in the A549 EMT data (i.e. the start, middle, and the end of the continuous portion of TGF-β induction) and then performed the dimensionality reduction. We found that the low-dimensional functional space was robust with respect to the removal, regardless of whether the missing time point is in the middle of the progress or at the extremes (Figure 2, A through D), such that when we projected the removed data points onto the space derived from a partial dataset, their positions were highly correlated with their positions when they were included in the data set (Pearson correlation coefficient (PCC) > 0.95). However, while the inferred 1-day samples were similarly separable from samples in 8-hr (f = 0.83 for E, 0.87 for M) and 3-day (f = 0.81 for E, 0.82 for M) time points, we observed reduced separability between the both inferred 0-day vs. 8-hr (f = 0.52 for E, 0.56 for M) and 3-day vs inferred 7-day (f = 0.56 for E, 0.56 for M) time points, with E-scores not significantly separating 0-day and 8-hour samples (Mann-Whitney U-test, p = 0.061). We also applied the same inference procedure to samples which were exposed to a transient EMT-inducing signal and allowed to revert. However, because the 8-hr and 24-hr reversion samples largely overlap with 7-day (hence their removal for 7-day inference, Figure 2E), we focused on inferring 3-day reversion samples after performing dimensionality reduction on the data set without any reversion samples. We found that 3-day reversion samples were positioned in the middle of the EMT spectrum, consistent with when they were included in functional space construction (PCC = 0.99 for E and 0.98 for M, Figure 2F). Additionally, the inferred 3-day reversion samples were similarly separable from the 7-day samples (f = 0.84 for E, 0.88 for M) as when they were when included in functional space construction (f = 0.83 for M, 0.86 for M). We obtained similar results using gsPCA to infer the position of samples from missing time points (Figure S2), but neither E- nor M-scores significantly separated the end points (0-day vs 8-hr and 3-day vs 7day). These results suggest that gsPCA and gsNMF can predict cell states of new data without retraining the model, and that these methods can be used to predict new cell states that have not been observed directly, though doing so at the edge of the spectrum is more difficult where initial separability of time points is lower.

Using functional space across cell lines

The transferability of gsNMF can be extended to data from different cell lines. We performed gsNMF on single-cell transcriptomes of TGF-β-treated DU145 from Cook and Vanderhyden (Cook and Vanderhyden, 2020) using the same procedure as A549 and obtained a moderate explanation of variance in time of EMT progression using the E and M dimensions (Adjusted R^2 = 0.34). We then inferred the position of the five continuous time points in the A549 data set using the DU145 model and vice versa (Figure 3 A through D). Transferred models (DU145 on A549 and A549 on DU145) were able to separate the individual time points, but overall performance decreased as they can explain only about 70% of the variance seen in the original models (Adjusted R^2 = 0.37 for DU145 on A549 and 0.24 for A549 on DU145). Therefore, it was expected that the individual sample scores would be positively correlated between models along both the E (Figure 3 E and F) and M dimensions (Figure 3 G and H). However, while the correlations between all pairs of scores were significant (minimum p = 2.7e-73), the correlation between E-scores was limited and worse for models of DU145 (PCC = 0.39) than models of A549 (PCC = 0.50). Comparably, the M-scores for both models of A549 (PCC = 0.85) and models of DU145 (PCC = 0.83) were more highly correlated and consistent between models. However, none of the sample scores between A549 and DU145 models were as correlated as inferred sample scores from missing point and the complete A549 model (PCC > 0.95). This suggests a reduced transferability across cells lines compared to within cells lines. In addition, across the data sets we used, changes along the M dimension were more consistent than the E dimension. We observed similar results using gsPCA, including M-scores being more correlated (PCC, A549 = 0.92, DU145 = 0.93) than E-scores (PCC, A549 = 0.71, DU145 = 0.68) (Figure S3). This is consistent with the fact that the same inducing agent was used across all cell lines, and also implies that inducing EMT in different cell types may yield more consistent changes in M genes compared to E genes.

Using functional space across experimental conditions

In addition to predicting the locations in the functional space across cell lines, gsPCA and gsNMF can be used across both experimental conditions and cell types. To test the cross-condition transferability, we first used our low-dimensional functional EMT space for A549 and DU145 cells to analyze tumor transcriptomes measured with bulk RNA-seq (The Cancer Genome Atlas, TCGA). To perform the most comparable transfer, we used lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD) data, which correspond to A549 and DU145 in terms of tissue type. We considered transfers between both similar (projecting LUAD data by a A549-trained model, and PRAD by a DU145-trained model) and dissimilar (LUAD by DU145 and PRAD by A549) cell types. We found that the low-dimensional functional space obtained with in vitro data captured tumor sample heterogeneity in the EMT spectrum when compared to our previous GSVA analysis of the same data (Figure 4). Overall, the original E- and M-scores were significantly correlated with the A549 models in all cases (smallest p = 2.1e-53). Correlations were consistent for LUAD across both the E (PCC = 0.76 for A549 and 0.73 for DU145) and M (PCC = 0.77 for A549 and 0.75 for DU145) dimensions. Comparably, PRAD scores were less consistent along the E dimension (PCC = 0.59 for A549 and 0.60 for DU145) than along the M dimension (PCC = 0.87 for A549 and 0.86 for DU145). Additionally, although the effect was more pronounced in PRAD data, much like transferring across cell lines, we once again observed greater consistency with M-scores suggesting a broader context for the consistency of M-gene expression compared to E-gene expression. Additionally, we obtained similar results with gsPCA (Figure S4), except that E-scores were more correlated in LUAD (A549 = 0.79, DU145 = 0.73) than M-scores (A549 = 0.73, DU145 = 0.70), although PRAD correlations remained similar for E (A549 = 0.54, DU145 = 0.57) and M (A549 = 0.88, DU145 = 0.87). Overall, this suggests that the robustness of derived functional spaces is broadly transferrable across data and cell types.

We next examined if gsNMF can produce transferrable models that reveal both spatial and temporal progression of EMT. Using single-cell RNA-seq, McFaline-Figueroa et al. previously found that

epithelial cells exhibit an E-to-M spectrum from the inner position of a colony to the outer position (McFaline-Figueroa et al., 2019). This dataset that contains binarized identities (inner and outer) obtained with macro-dissection (defined as spatial EMT data) from two experiments, one in which cells were allowed to migrate without external induction of EMT (Mock), and one in which EMT was induced with TGF-β (TGF-β). Since there are only two populations in this data set, the leading dimension for E- and M-scores was chosen to maximize the separation based on the *f* probability. Overall, three analyses were performed for each data set: spatial data with its own gsNMF model, spatial data with the model from the other spatial data set (TGF-β on Mock and Mock on TGF-β), and spatial data with A549 time series model (Figure 5). As with our previous results, the best separation of inner and outer data points was observed when Mock (f = 0.67 for E, 0.80 for M) and TGF-β (f = 0.86 for E, 0.85 for M) data sets had their own model applied to them. However, for Mock data, the TGF-β model (f = 0.67 for E, 0.78 for M) outperformed the A549 model (f = 0.48 for E, 0.64 for M) on both dimensions and, in fact, the E dimension of the A549 model did not effectively separate inner and outer points in the Mock data (p = 0.99). In comparison, the Mock model better separated TGF-β inner and out points in the E direction (f = 0.77 for E, 0.73 for M), while the A549 model better separated them in the M direction (f = 0.66 for E, 0.76 for M). gsPCA models gave similar results, including the A549 model yielding better performance along the M-dimension (f = 0.78) for TGF-β data than Mock data (f = 0.75) (Figure S5).

As with previous examples of transferrable models, we once again found that the performance with the new M-scores was better than that of the new E-scores obtained from the transferred models. However, the spatial context provides important additional observations in this regard. The fact that Mock and TGF-β spatial data sets were separated well by either spatial model along the E dimensions suggests that the variations in E expression were similar whether migration occurs with or without TGF-β induction. On the other hand, the fact the A549 model better separated TGF-β spatial points along the M dimension than the Mock model, but did not outperform TGF-β on the Mock model suggests that there is conserved TGF-β induced M-gene expression regardless of context. To explore the basis of this similarity/dissimilarity in

E/M-scores, we compared the loading matrices between Mock, TGF-β, and A549 models. We found little correlation between A549 and spatial E-gene loading along the lead dimensions (PCC = -0.08. p = 0.33 for Mock; PCC = -0.10, p = 0.20 for TGF-β), but positive correlation between A549 and spatial M-gene loading values (PCC = 0.56, p = 1.4e-7 for Mock; PCC = 0.64, p = 3.6e-12 for TGF-β). Additionally, M-genes with the highest loading values were common amongst spatial and A549 models: VIM, LGALS1, and TPM2 were in the 90th percentile of M-gene loading values of all models, while KRT8 was in the 90th percentile of both Mock and A549, and GLIPR1, SERPINE1 and TAGLN were in the 90th percentile of both TGF-β and A549. Conversely, while we observe E-genes in the 90th percentile of loading values for Mock, TGF-β, and A549 models (ELF3, S100P, PERP, SLPI), we also saw an equal (4, Mock vs A549) or greater (7, TGF-β vs A549) number of E-genes in the 90th percentile of one model but the 10th percentile of the other, consistent with the overall non-significant correlation of values. We observed similar results from gsPCA, with VIM, LGALS1, and TPM2 appearing in the 90th percentile of weights for Mock, TGF-β, and A549 M-gene models, but the overall correlation between M-gene loading values was weaker between Mock and A549 (PCC = 0.30, p = 0.008) as well as TGF-β and A549 (PCC = 0.55, p = 1.2e-8) as compared to the results from gsNMF. Together, these results suggest a coherence of the progression of EMT program in both the spatial and temporal context with regard to M-genes, while E-gene progression appears to be more sensitive to context. In addition, they highlight the usefulness of the transferability of gsPCA and gsNMF outside of a time series context, where performance may need to be evaluated on discrete groups or the analyses are conducted in an unsupervised fashion.

Characterizing relationships among multiple functional spectrums

To test the capacity of gsPCA and gsNMF in analyzing a wide range of gene sets, we returned to the A549 data set and examined the expression changes of gene sets relative to time. Taking advantage of the high-efficiency of those methods, we began with 5455 C2 curated gene sets from the Molecular Signature Database (see Methods) and applied a gsNMF model to A549 data for each, extracting the leading

dimension based on correlation with time of EMT progression. For simplicity, we used a two-component model, but we applied stricter convergence and selection criteria because of the diversity of gene set size and coverage by the data set (see Methods). This yielded 5120 (93.8% of the original) gene sets, of which 3887 had a leading dimension positively correlated with time and 1233 were negatively correlated with time (Figure 6 A and B). As expected, we found gene sets associated with cancer (SMID BREAST CANCER BASAL UP, SCHUETZ BREAST CANCER DUCTAL INVASIVE UP, DODD NASOPHARYNGEAL CARCINOMA UP) and stem cells (BOQUEST STEM CELL UP, LIM MAMMARY STEM CELL UP, WONG ADULT TISSUE STEM MODULE) produced dimensions that were highly correlated with EMT progression (PCC > 0.7), with cell migration falling just short of this threshold (WU CELL MIGRATION, PCC = 0.697). Likewise, gene sets associated with repression of metastasis (JAEGER METASTASIS DN, RICKMAN METASTASIS DN) produced some of the most anti-correlated dimensions (PCC < -0.6). We also observed separation between specific cancer types with gene sets favoring luminal breast cancers (CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL UP, PCC = -0.60; CHARAFE BREAST CANCER LUMINAL VS BASAL UP, PCC = -0.56) producing anti-correlated dimensions, while those which favored basal or mesenchymal breast cancers produced correlated dimensions (CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL DN, PCC = 0.73; CHARAFE BREAST CANCER LUMINAL VS BASAL DN, PCC = 0.72). As expected, the same gene sets up-regulated in luminal breast cancer tended to have more E-genes than M-genes (72 vs 7, 22 vs 3) while gene sets down-regulated in luminal breast cancer had more M-genes than E-genes (51 vs 5, 23 vs 13)

Overall, the number of EMT genes in these sets was correlated with the magnitude of correlation with time: M-gene count was correlated with the magnitude of correlation in positive gene sets (PCC = 0.55), while the E-gene count was correlated with the magnitude of correlation in negative gene sets (PCC = 0.45). However, this implies that only 20% (E-genes) to 30% (M-genes) of the variance in gene set performance can be explained by the presence of the expected type of EMT genes and, 81.3% of gene sets

were not significantly enriched in either E or M genes (Fisher's Exact Test, FDR < 0.05). Nevertheless, this group of non-enriched gene sets still contained gene sets which produced strongly correlated dimensions, including poorly differentiated tumors (RICKMAN TUMOR DIFFERENTIATED WELL VS POORLY DN, PCC = 0.71), adipogenesis (BURTON ADIPOGENESIS 7, PCC = 0.69), and graft vs. host disease (CHIBA GRAFT VERSUS HOST DISEASE 35D UP; PCC = 0.69), as well as strongly anti-correlated dimensions, including reduced drug resistance (MASSARWEH TAMOXIFEN RESISTANCE DN; PCC = -0.62), reduced RAS pathway expression (SWEET KRAS TARGETS DN; PCC = -0.62), and reduced expression in cancer (LINDGREN BLADDER CANCER CLUSTER 3 DN, PCC = -0.66; WANG ESOPHAGUS CANCER VS NORMAL DN, PCC = -0.6). A full list of selected gene set and their correlation with EMT progression can be found in Table S1. These results show that gsNMF can be used in exploratory or hypothesis-free analyses for extracting meaningful gene sets in a high-throughput and unbiased manner.

De-novo Gene-set Construction

We next sought to generate an A549 TGF-β induction specific gene set by exploring the performance of individual genes across C2 gene sets. This approach offers an advantage over traditional analysis by selecting genes which are not only correlated with EMT progression or which are part of a well correlated gene set, but also drive the correlation of models scores in multiple contexts. To accomplish this task, we first removed any gene which appears in less than 20 gene sets to avoid spuriously well-performing genes, then applied two selection criteria to select genes: its association with highly correlated gene sets (i.e. the average PCC of gene sets which the gene belongs to, mean_PCC) and its importance to defining leading dimensions (i.e. the percentile of its average loading value across gene sets, percentile). Additionally, we focused on genes sets positively correlated with the time of EMT progression (and thus presumably enriched with M-genes) as they were more frequent and had a higher magnitude of PCC with EMT on average compared to negatively correlated gene sets (0.28 vs 0.19).

We found that when using the most permissive mean_PCC (> 0.0), the correlation of the *de novo* gene set model with EMT progression improved with increasing percentile cutoffs up 98%, while using a stricter mean_PCC gave better performance at lower percentile cutoff, but declined as more stringent percentiles cutoffs were applied (Figure 6C). Additionally, when we considered the results relative to the size of the gene set (Figure 6D), we found that both extremes of our grid search, strong percentile selection alone (mean_PCC> 0.0 & percentile > 0.98) and strong PCC selection alone (mean_PCC > 0.5 & percentile > 0.80), yielded gene sets of similar size (93 for the former, 70 for the latter) and models of similar of performance (PCC = 0.70). While using a lower PCC threshold gave a slightly better model on average (mean_PCC> 0.4 & percentile > 0.80, PCC = 0.71), it also tripled the size of the gene set (248).

While the PCC and percentile selected gene sets shared more genes in common than expected by random chance alone (20 vs. 0.3), they were nonetheless distinct in regard to both M-gene and functional enrichment. While both the PCC and percentile gene sets were enriched for M-genes, even adjusting for the 5120 gene sets we analyzed (FDR = 1.2e-21 and 2.7e-5, respectively), the PCC gene set was significantly more enriched for M-genes than the percentile set when compared directly (16 vs 6; Fisher's Exact Test, p = 0.002). However, the percentile set does contain VIM, LGALS1, GLIPRI1, and TAGLN, which were in the 10th percentile of loading values for both TGF-β and A549 spatial models, while the PCC set only contains TAGLN (notably, TPM2 and SERPINE1 were in the 97.5th percentile, just short of the cutoff). Furthermore, we observed extensive GO functional enrichment in the percentile gene set, with over 200 enriched terms, including several which were related to ribosomes, actin filaments, actin polymerization, membrane projection, cell adhesion, cell motility, cell surface receptors, and cell differentiation (see Table S2 for a full list). Comparably, the PCC selected gene set was enriched for only 10 terms (see Table S3), though some EMT associated processes like extra-cellular matrix components and growth factor binding were included.

As such, selecting based on PCC yielded a gene set which was most notably enriched for M-genes, which we both expected and demonstrated would produce a model highly correlated with EMT progression, but

was relatively limited in terms of functional enrichment. Conversely, the percentile selected group exhibits enrichment for diverse, EMT related functions, and, while it was more M-gene poor, it includes the majority of the M-genes which have been shown to drive M-dimension correlation with EMT progression in both temporal and spatial contexts. This indicates that, not only can the loading values of the leading dimension be used to identify biological significant individual genes and functional gene sets, but can yield results that are distinct from gene sets derived from correlation with EMT progression alone.

Discussion

Previous methods that aimed to address the challenges of visualizing single-cell data in functional space were primarily based on weighted sum of expression values or Kolmogorov–Smirnov test with full datasets (DeTomaso and Yosef, 2016; Hänzelmann et al., 2013). These methods are useful to analyzing samples with functional gene sets, they do not provide transferability which is essential for predicting cell states with existing models and new data. We showed that constrained linear transformation enables good performance in depicting cell states with straightforward interpretation in functional space and satisfactory efficiency. While more sophisticated methods such as deep generative models have potentials to address similar problems, current methods primarily focus on the interpretability in terms of inter-sample distances in low dimensions rather than the dimensions themselves (Ding et al., 2018; Lopez et al., 2018), and we expect that the gsPCA and NMF methods are more efficient than models based on nonlinear connectivity.

Factorization approaches like PCA and NMF have previously been applied to the problem of gene expression, with NMF in particular having been used to deconvolute expression patterns scRNA-seq data sets (Fujita et al., 2018; Kotliar et al., 2019), but these approaches have primarily focused on the unsupervised clustering of samples and/or for *de-novo* module discoveries at relatively high dimensionality ($n > 10$). In contrast, our approach suggests there is a utility in applying these factorization approaches to interrogating the relationship between known gene modules and data with implicit structure and/or separable populations of samples, particularly when assessing a single biological process (EMT) across multiple contexts (e.g. cell line, time and space), such that the simplicity of low-dimension space ($n = 2$) can be leveraged for visualization and analysis.

In this work we have found that conserved EMT gene expression signatures can be used to describe stages of EMT in multiple cell lines (e.g. A549 and DU145), and these signatures not only capture the subpopulation heterogeneity resulting from differential times of treatment with EMT-inducing signals such as TGF-β, but also reflect the EMT program driven by spatial heterogeneity with cell populations

(McFaline-Figueroa et al., 2019). These results are consistent with the existence of conversed EMT program across cell lines (Cook and Vanderhyden, 2020), but do not contradict the idea of context specific expression as models trained and applied to the same data set always explained more variance in EMT progression. The coexistence of a common EMT signature and context specific expression is further supported by the observation that M-scores were more consistent and better separate data across different contexts of EMT than E-scores, and the related observation that M-gene loading values were correlated across spatial and temporal models, while E-genes were not. This suggests that M-gene induction by TGF-β is consistent across cellular contexts, while changes in E-gene expression are more variable, possibly due to greater sensitivity to cell line, environmental context, or other initial conditions effecting the cell prior to induction.

The transferability of models across EMT context indicates the synergy between spatial arrangement of cells and external signals (e.g. TGF-β) in determining the stages of EMT. In addition, we found that the functional dimensions obtained with TGF-β can serve as reasonable approximations for the positioning of tumor transcriptomes in the EMT spectrum. Similar to the EMT spectrum, many biological processes involve stepwise changes of gene expression programs. A possible mechanism underlying these non-binary programs is the feedback-driven formation of stable intermediate cell states (Ye et al., 2019; Yui and Rothenberg, 2014). With the rapid advances of the single-cell technology, transcriptome-wide gene expression data will become available for more biological systems. We expect that our functional projection methods can be widely useful for visualizing and analyzing these data. In particular, the transferability of the models can be a powerful feature for interrogating the relationships among different experimental conditions and cell types.

Methods

Gene expression data sources

scRNA-sequencing data and meta data for A549 and DU145 cell lines were obtained from Cook and Vanderhyden (Cook and Vanderhyden, 2020). In brief, we obtained pre-processed .rds data files and used the ScaleData function from Seurat to regress out mitochondrial gene expression, total unique reads in a sample, and cell cycle gene expression (Stuart et al., 2019). For McFaline-Figueroa et al. spatial scRNA-seq and TCGA bulk RNA-seq data, we obtained count data from GEO (GSE114687) and TCGAbiolinks, respectively (Colaprico et al., 2016; McFaline-Figueroa et al., 2019). Counts were transformed to $\log_2$TPM with a pseudo-count of 1 using gene models for the hg38 annotation of the human genome obtained from RefSeq (O'Leary et al., 2016). scRNA-seq data from McFaline-Figueroa et al. was further processed to remove genes expressed in < 1% of cells.

Dimension reduction approaches and evaluation

gsPCA uses the non-negative approach to PCA pioneered by Sigg and Buhnam (2008). In brief, the vectors of weights, w, used to define the first principal component of PCA is defined such that it maximizes the variance of the first component, i.e:

$$\arg\max_{\boldsymbol{w}} w^{\mathrm{T}} C w$$

Where C is the covariance matrix of the original data set X and w is unit vector ($||w||^2 = 1$)). In our case, X is an m by n matrix of expression values where m is the number of samples and n is the number of genes in the selected gene set. This method for determining w can be treated as an expectation maximization problem where the original data is projected using the current estimate of w ($y = Xw_t$) and this projection is used to re-estimate w using the following minimization step:

$$w_{t+1} = \arg\min_w \sum_{n=1}^{N} ||x_n - y_n w||_2^2$$

Where x_n are the rows of the original data and y_n are the rows of the projected data (Sigg and Buhmann, 2008). This expectation-maximization formulation allows additional constraints on w, including forcing the component values to be non-negative. Note that the non-negativity constraint applies only to the weight components such that negative scores can still exist if there are negative values in underlying data, such as those produced by centering expression data to zero which we did for all gsPCA inputs. Subsequent components are calculated in the same way, under the constraint that they are orthogonal to the preceding ones. We implemented non-negative PCA in R using the *nsprcomp* function (with the option nneg=TRUE) from the package of the same name (Sigg and Buhmann, 2008).

NMF involves factorizing the original data matrix of non-negative values into two matrices whose product estimates the original data, i,e.:

$$X \cong WH$$

Where X is the original matrix (m by n)**,** W is the transform matrix (m by p), and H is the component matrix (p by n), such that m is the number of rows in the original matrix (samples in our case), n is the columns (genes in our case), and p is the number of components used in the factorization. Because the original matrix is constrained to being non-negative, we subtracted the minimum of value of the scaled expression matrix from all values to create a non-negative input matrix. As a consequence, the values of the W and H matrices must likewise be non-negative such that product is non-negative. Multiple approaches and implementations for choose W and H matrices exist, but for ease use and accessibility, we used the Scikit-learn implementation of NMF (Pedregosa et al., 2011).

To focus specifically on EMT progression, input data was filtered using signature E and M gene sets constructed from EMT genes identified in Tan et al. (Tan et al., 2014), but using annotations from our previous study of TGF-β induced EMT (Watanabe et al., 2019). For both gsPCA and gsNMF, we used the

standard convergence parameters and two components for analysis. Other parameters were considered, but a cross-validation analysis was performed, in which a 2-component model produced the most consistent results (the full results of these tests are discussed in the Supplemental Methods). For other data sets, we tested ten random seeds for consistency (average PCC of dimensions > 0.99) and tuned the convergence parameters as necessary (this was only necessary for gsNMF on the TGF-β spatial data set, for which we used 2500 iterations and a tolerance of 1e-9). GSVA and z-score methods were implemented using the GSVA package in R (Hänzelmann et al., 2013).

gsPCA and gsNMF both produce two outputs of interest. For gsPCA, the principal component scores were used as sample level E/M-scores while the weight vectors (w) values represent the contribution of individual genes in each set to the respective scores. For gsNMF, the transform matrix (W) was used as sample level scores while the component matrix (H) loading values represent the gene level contributions. Z-scores and GSVA both directly generate scores, but determining the contribution of individual genes is non-trivial. Sample scores were evaluated using several criteria. First, for gsPCA and gsNMF, the leading dimensions for E and M gene sets were selected based on the correlation between sample score and EMT progression in both A549 and DU145 data, picking the greatest negative PCC for E and the greatest positive PCC for M. PCC was used in this case to distinguish direction as we expect E to anti-correlated and M to be correlated with EMT progression (for spatial data, where there was only two populations, *f* probability was used instead, see below). Then, the overall ability for a pair of E/M-scores to explain EMT progression over time was assessed by creating a linear model (time ~ E + M) and determining the coefficient of determination, which is the percent of overall variance in the dependent variable explained by the independent variable (Adjsuted R^2). Finally, the ability of E/M-scores to separate sample from different time/spatial populations was assessed using Mann Whitney U-statistic, which has two uses. First, the U-statistic can be used to determine the significance of score when used to separate two populations of samples. Second, by dividing the U-statistic by the total number of possible pairs of samples for two populations, we can calculate the common language effect size (*f*) which is the probability that if a pair of

samples are chosen at random from two populations, the score of the sample from the 'first' population is higher. Because this probability is reciprocal, the choice of the 'first' population is arbitrary, so we chose to calculate the f probability such that the 'first' population is the more progressed for M and less progressed for E. Therefore, a higher probability always indicates better correspondence with EMT progression. As such, a probability of 1.0 would indicate perfect correspondence of the E/M-scores with progression, while a probability of 0.5 indicates random association and values of < 0.5 suggests an inverse relationship compared to what we would expect. Note that this probability is equivalent to the area-under the receiver operating curve for a classifier built on the respective score.

Inference and model transfer

To infer the position of new data in functional space for gsPCA, we multiplied the new data directly by weight vectors (w) of the E- and M-scores. For gsNMF, we used the Scikit-learn 'transform' method which transforms the input data according to the fitted model (i.e., returns a new W transform matrix while holding H constant). In both cases, we used the same leading dimension for inference as in the original model. For inferring missing data points, no further steps were required as the new data always had the same coverage of the E and M gene sets as the original. However, for transferring models across cell-line, TCGA, and spatial data, we first had to determine the common set of genes between the two data sets. Common genes were then used to filter the weight vectors for gsPCA and to refit the model to be transferred on the common subset of genes for gsNMF. Input data was then filtered by the same subset of data and inference was done as described previously. Transferred models were assessed against the new data set using the same approaches as the original models, but relationship between E/M-scores and gene rotation/loading values between models built on the original data and transferred models were assessed by PCC for both magnitude and direction.

Multi-gene set evaluation and de-novo gene set construction

C2 gene sets were obtained from the Molecular Signature Database (http://www.gsea-msigdb.org/gsea/msigdb/index.jsp) (Liberzon et al., 2015; Liberzon et al., 2011; Subramanian et al., 2005). gsNMF was performed as described for EMT gene sets expect that we increased the iteration (2500) and convergence threshold (10e-9) of the NMF algorithm to ensure consistent results across the gene sets which varied widely in size (2 to 1581 genes present in the data set) and coverage by the A549 data set due to the sparsity of scRNA-seq data. Additionally, we dropped any gene set where the variance of the PCC of scores with time across EMT progression was > 5% of the mean value across ten random initializations, leaving 5120 get sets. As such, the results for C2 models were highly correlated across the ten random seeds we analyzed (average PCC between seeds = 0.998). We also calculated the mean PCC with time of EMT progression and mean per gene loading value across all ten seeds and used those values to represent each gene set in subsequent analyses. EMT gene enrichment was evaluated using Fisher's Exact Test and multiple test correction was done using the Benjamini-Hocberg method.

De-novo gene sets were constructed using a grid search of two parameters: the mean PCC with EMT progression across all C2 greater than or equal to the following values (0.0, 0.1, 0.2, 0.3, 0.4, 0.5) and the mean component loading value across in the following percentile of all mean loading values (0.8, 0.85, 0.9, 0.95, 0.98, 0.99, 0.995). In both cases, the mean PCC/percentile is calculated across all C2 gene sets and gsNMF models which the gene is a part of. gsNMF was run for the A549 data with de-novo gene sets using the same parameters as C2 gene sets. GO Terms were identified using GO Enrichment Analysis tool available at http://geneontology.org/ (Ashburner et al., 2000; Carbon et al., 2021). Functional and M-gene enrichment was evaluated using Fisher's Exact Test and multiple test correction was done using the Benjamini-Hocberg method.

Data and Code availability

Code and data for generating the primarily results of this study can be found at https://github.com/panchyni/gsNMF

Author contributions

Designed research: TH. Performed research: NP and TH. Analyzed data: NP, KW, and TH. Wrote manuscript: N.P. and T.H.

Conflict of Interest

The authors declare no conflict of interest.

Figure Legends

Figure 1. Visualization of EMT progression in TGF-β induced A549 cells. (**A-D**) Contour plots of gene set scores of E (x-axis) and M (y-axis) genes from four different scoring methods, z-score (A), GSVA (B), gsPCA (C), and gsNMF (D). Color indicates the time of TGF-β induction from 0 days (dark green) to 7 days (dark purple). Circles indicate the mean E- and M-score of samples from each time point and the associated error bars show the standard deviation. (**E**) A box-plot showing the distribution of E (red) and M (blue) scores across all five time points of TGF-β induction from the gsNMF model. Whiskers indicate the 1.5 inter-quartile range of each distribution while the red points indicate outliers beyond this range. (**F**) Bar chart of the *f* probability values for E (top) and M (bottom) scores between all consecutive pairs of time points. Color indicates the method used to produce the score: red is z-score, orange is GSVA, blue is gsPCA, and purple is gsNMF. Bars marked by an '**x**' indicates that the score did not significantly separate the samples from those time points (Mann-Whitney U-test, $p < 0.05$).

Figure 2. Predicting A549 sample from specific time points using gsNMF. (**A**) Scatter plot of E (x-axis) and M (y-axis) scores for all TGF-β induction samples using gsNMF. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). (**B-D**) Scatter plot of 0-day (green, B), 1-day (yellow, C), and 7-day samples (purple, D) inferred using a gsNMF model built with all other time points (gray). (**E)** A scatter plot of TGF-β induction samples with TGF-β reversion samples (i.e. 7 days induction followed by removal from TGF-β). Induction samples are labeled as in (A), while reversion samples are colored blue, with darker shade indicating longer time since removal. (**F**) Scatter plot of 3-day reversion samples (dark blue) inferred using a gsNMF model built with all non-reversion time points (gray).

Figure 3. Transferring gsNMF models between A549 and DU145 TGF-β induced samples. (**A-D**) Scatter plot of E (x-axis) and M (y-axis) scores for different combination of data and gsNMF model: (A)

A549 model on A549 data, (B) DU145 model on A549 data, (C) A549 model on DU145 data, and (D) DU145 model on DU145 data. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). (**E-F**) Comparison of E-scores of samples from A549 (E) and DU145 (F) data. The x-axis is the E-score from using the model from the same data set (A549 on A549, DU145 by DU145), while the y-axis is the E-score from the opposite model (DU145 on A549, A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). (**G-H**) Comparison of M-scores of samples from A549 (G) and DU145 (H) data. The x-axis is the M-score from using the model from the same data set (A549 on A549, DU145 by DU145), while the y-axis is the M-score from the opposite model (DU145 on A549, A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple).

Figure 4. Transferring gsNMF models to TCGA data. (**A-B**) Scatter plots of E-scores for PRAD (A) and LUAD (B) from transferring gsNMF models built on A549 (x-axis) and DU145 (y-axis) data. The color of individual points indicates the original GSVA based E-score of the TCGA data set. (**C-D**) Scatter plots of M-scores for PRAD (C) and LUAD (D) from transferring gsNMF models built on A549 (x-axis) and DU145 (y-axis) data. The color of individual points indicates the original GSVA based M-score of the TCGA data set.

Figure 5. Transferring gsNMF models between temporal and spatial data sets. (**A-C**) Scatter plots of E (x-axis) and M (y-axis) scores for Mock spatial data from gsNMF models built on different data sets: Mock spatial data (A), TGF-β induced spatial data (B), and TGF-β induced A549 temporal data (C). The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue). (**D-F**) Scatter plots of E (x-axis) and M (y-axis) scores for TGF-β spatial data from gsNMF models built on different data sets: TGF-β induced spatial data (D), Mock spatial data (E), and TGF-β induced A549 temporal data (F). The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue).

Figure 6. Assessing EMT progression with curated gene sets. (**A-B**) Scatter plot showing the relationship between gene set size (x-axis, $\log_{10}$-scale) and the correlation (PCC) of the leading dimension of the gsNMF model of the gene set with progression of EMT in TGF-β induced A549 data. Gene sets that are positively correlated with the time of EMT progression are shown in (A), while negatively correlated gene sets are shown in (B). The color samples indicate the number of M (A) or E (B) genes in the gene set. (**C**) A line plot showing the relationship between the loading value percentile used to select genes for *de-novo* gene sets (x-axis) and the correlation of the resulting gsNMF model of the gene set with the progression of EMT in in TGF-β induced A549 data. The color of each line indicates the minimum mean_PCC used to select gene (black = 0.0, red = 0.1, green = 0.2, blue = 0.3, purple = 0.4, brown = 0.5). (**D**) A line plot showing the relationship between the size of *de-novo* gene sets (x-axis) and its correlation with the progression of EMT in in TGF-β induced A549 data. The color of each line indicates the minimum mean_PCC used to select gene (black = 0.0, red = 0.1, green = 0.2, blue = 0.3, purple = 0.4, brown = 0.5).

Supplemental Figure S1. Performance of gsPCA and gsNMF models across cross-validation data sets. (A) Boxplots showing the Adjusted R^2 of the linear model of gsPCA leading E and M dimensions across different number of model components (x-axis) using basic convergence criteria (algorithm standard). Adjusted R^2 is measure two ways: the Adjusted R^2 of each independent validation fold (left) and the Adjusted R^2 averaged across all five validation folds for each of 60 training data sets (see Supplemental Methods). (**B-D**) Boxplots showing Adjusted R^2 of the linear model of gsNMF leading E and M dimensions across different number of model components (x-axis) and different convergence criteria: basic (algorithm standard, **B**), strong (1e-6 tolerance, 500 iterations, **C**), and nndsvd (initialization with non-negative singular value decomposition, 1e-6 tolerance, 500 iterations, **D**). Left and right panels are as in (**A**). The yellow line indicates the average of each distribution while the whiskers show 1.5 times the interquartile range. Red dots show the individual Adjusted R^2 values in each distribution.

Supplemental Figure S2. Predicting A549 sample from specific time points using gsPCA. (**A**) Scatter plot of E (x-axis) and M (y-axis) scores for all TGF-β induction samples using gsPCA. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). (**B-D**) Scatter plot of 0-day (green, B), 1-day (yellow, C), and 7-day samples (purple, D) inferred using a gsPCA model built using all other time points (gray). (**E)** A scatter plot of TGF-β induction samples with TGF-β reversion samples (i.e. 7 days induction followed by removal from TGF-β). Induction samples are labeled as in (A), while reversion samples are colored blue, with darker shade indicating longer time since removal. (**F**) Scatter plot of 3-day reversion samples (dark blue) inferred using a gsPCA model built using all non-reversion time points (gray).

Supplemental Figure S3. Transferring gsPCA models between A549 and DU145 TGF-β induced samples. (**A-D**) Scatter plot of E (x-axis) and M (y-axis) scores for different combination of data and gsPCA model: (A) A549 model on A549 data, (B) DU145 model on A549 data, (C) A549 model on DU145 data, and (D) DU145 model on DU145 data. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). (**E-F**) Comparison of E-scores of samples from A549 (E) and DU145 (F) data. The x-axis is the E-score from using the model from the same data set (A549 on A549, DU145 by DU145), while the y-axis is the E-score from the opposite model (DU145 on A549, A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). (**G-H**) Comparison of M-scores of samples from A549 (G) and DU145 (H) data. The x-axis is the M-score from using the model from the same data set (A549 on A549, DU145 by DU145), while the y-axis is the M-score from the opposite model (DU145 on A549, A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple).

Supplemental Figure S4. Transferring gsPCA models to TCGA data. (**A-B**) Scatter plots of E-scores for PRAD (A) and LUAD (B) from transferring gsPCA models built on A549 (x-axis) and DU145 (y-axis) data. The color of individual points indicates the original GSVA based E-score of the TCGA data set. (**C-**

D) Scatter plots of M-scores for PRAD (C) and LUAD (D) from transferring gsPCA models built on A549 (x-axis) and DU145 (y-axis) data. The color of individual points indicates the original GSVA based M-score of the TCGA data set.

Supplemental Figure S5. Transferring gsPCA models between temporal and spatial data sets. (**A-C**) Scatter plots of E (x-axis) and M (y-axis) scores for Mock spatial data from gsPCA models built on different data sets: Mock spatial data (A), TGF-β induced spatial data (B), and TGF-β induced A549 temporal data (C). The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue). (**D-F**) Scatter plots of E (x-axis) and M (y-axis) scores for TGF-β spatial data from gsPCA models built on different data sets: TGF-β induced spatial data (D), Mock spatial data (E), and TGF-β induced A549 temporal data (F). The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue).

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol *37*, 38-44.

Carbon, S., Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., and Hartline, E. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res *49*, D325-D334.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., *et al.* (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res *44*, e71.

Cook, D.P., and Vanderhyden, B.C. (2020). Context specificity of the EMT transcriptional response. Nat Commun *11*, 1-9.

Cursons, J., Pillman, K.A., Scheer, K.G., Gregory, P.A., Foroutan, M., Hediyeh-Zadeh, S., Toubia, J., Crampin, E.J., Goodall, G.J., Bracken, C.P., *et al.* (2018). Combinatorial Targeting by MicroRNAs Co-ordinates Post-transcriptional Control of EMT. Cell Systems *7*, 77-91.e77.

DeTomaso, D., and Yosef, N. (2016). FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics *17*, 1-12.

Ding, J., Condon, A., and Shah, S.P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun *9*, 1-13.

Fujita, N., Mizuarai, S., Murakami, K., and Nakai, K. (2018). Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Sci Rep *8*, 1-10.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7.

Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., and Sabeti, P.C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife *8*, e43803.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell systems *1*, 417-425.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739-1740.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat Methods *15*, 1053-1058.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single‐cell RNA‐seq analysis: a tutorial. Mol Syst Biol *15*, e8746.

McFaline-Figueroa, J.L., Hill, A.J., Qiu, X., Jackson, D., Shendure, J., and Trapnell, C. (2019). A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. Nat Genet *51*, 1389-1398.

McGraw, K.O., and Wong, S.P. (1992). A common language effect size statistic. Psychol Bull *111*, 361.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., and Ako-Adjei, D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-D745.

Panchy, N., Azeredo-Tseng, C., Luo, M., Randall, N., and Hong, T. (2020). Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic Cells. Front Oncol *9*, 1479.

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi, S., Van Keymeulen, A., Brown, D., Moers, V., and Lemaire, S. (2018). Identification of the tumour transition states occurring during EMT. Nature *556*, 463-468.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research *12*, 2825-2830.

Sigg, C.D., and Buhmann, J.M. (2008). Expectation-maximization for sparse and non-negative PCA. Paper presented at: Proceedings of the 25th international conference on Machine learning.

Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., and Ochs, M.F. (2018). Enter the matrix: factorization uncovers knowledge from omics. Trends Genet *34*, 790-805.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck Iii, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888-1902.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., and Lander, E.S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA *102*, 15545-15550.

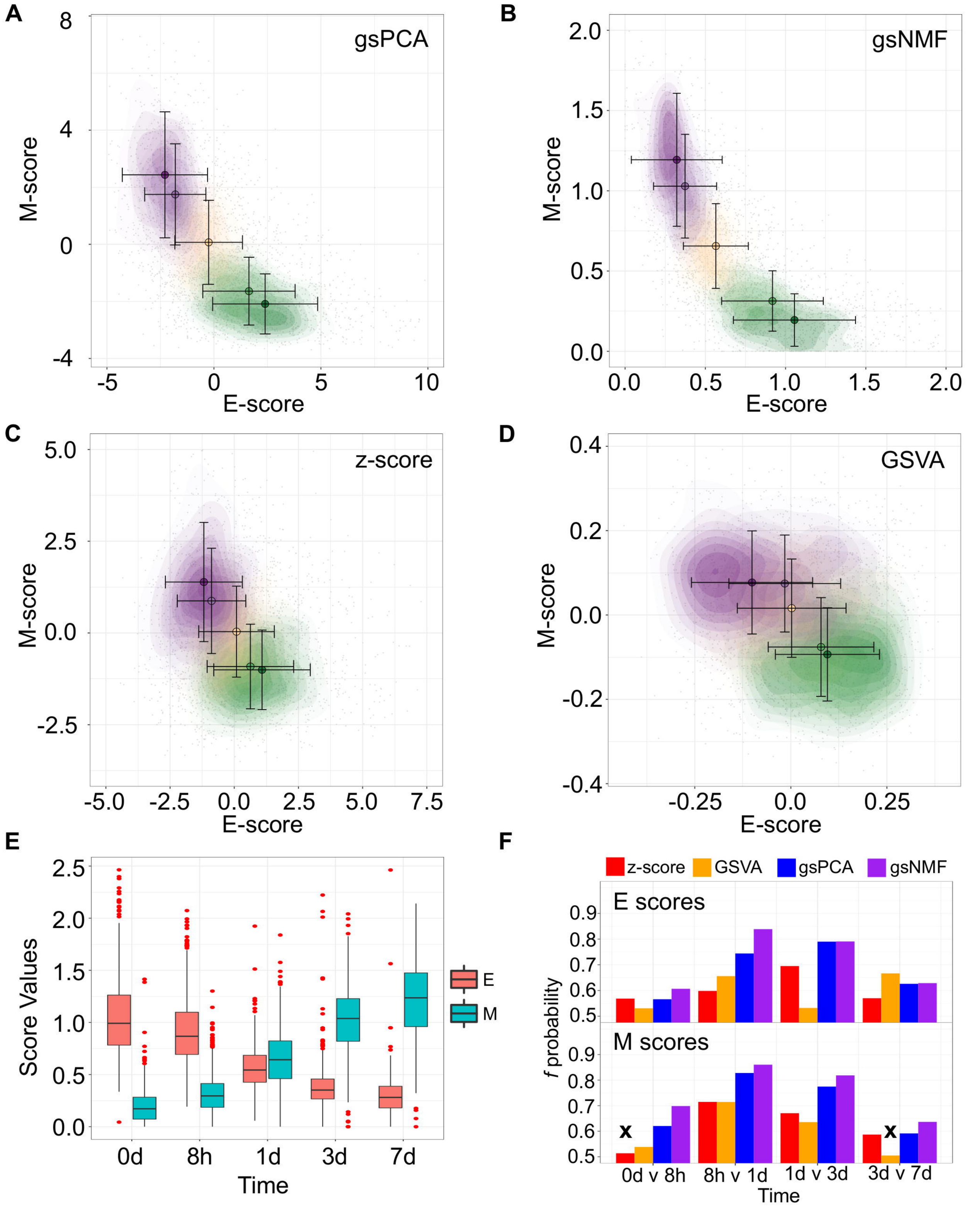
Tan, T.Z., Miow, Q.H., Miki, Y., Noda, T., Mori, S., Huang, R.Y.J., and Thiery, J.P. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med *6*, 1279-1293.

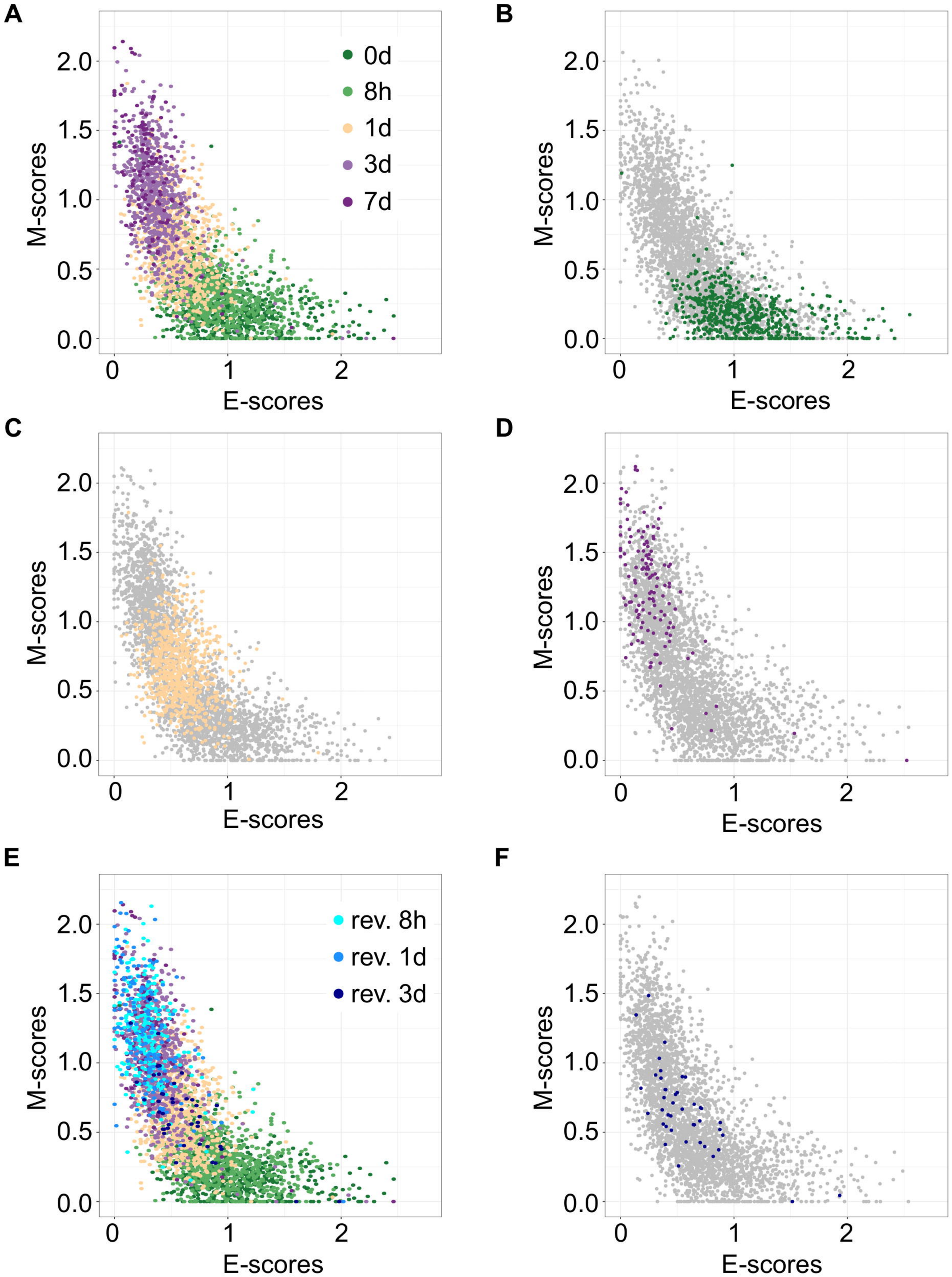
Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research *9*.

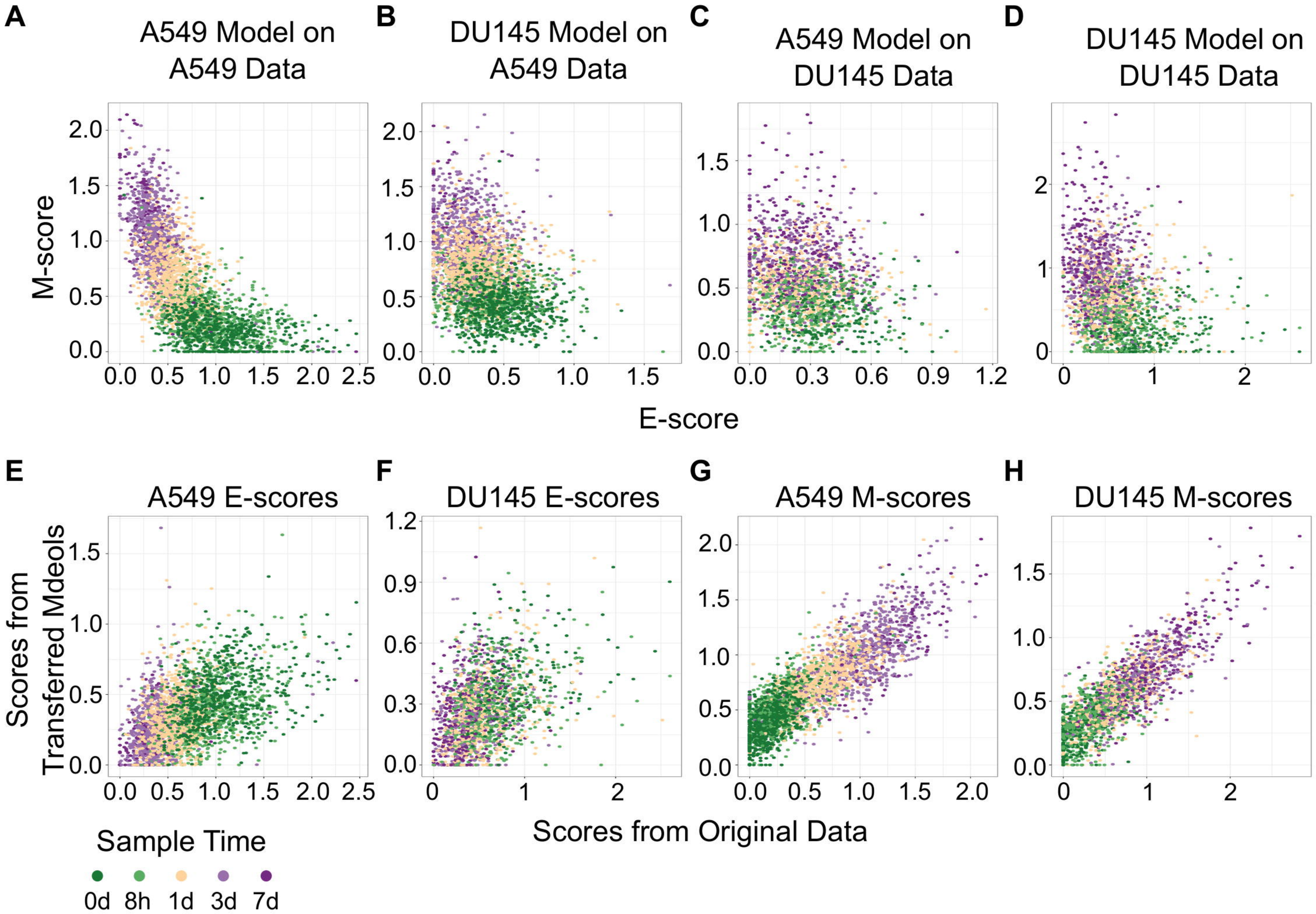
Watanabe, K., Panchy, N., Noguchi, S., Suzuki, H., and Hong, T. (2019). Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition. npj Syst Biol Appl *5*, 21.

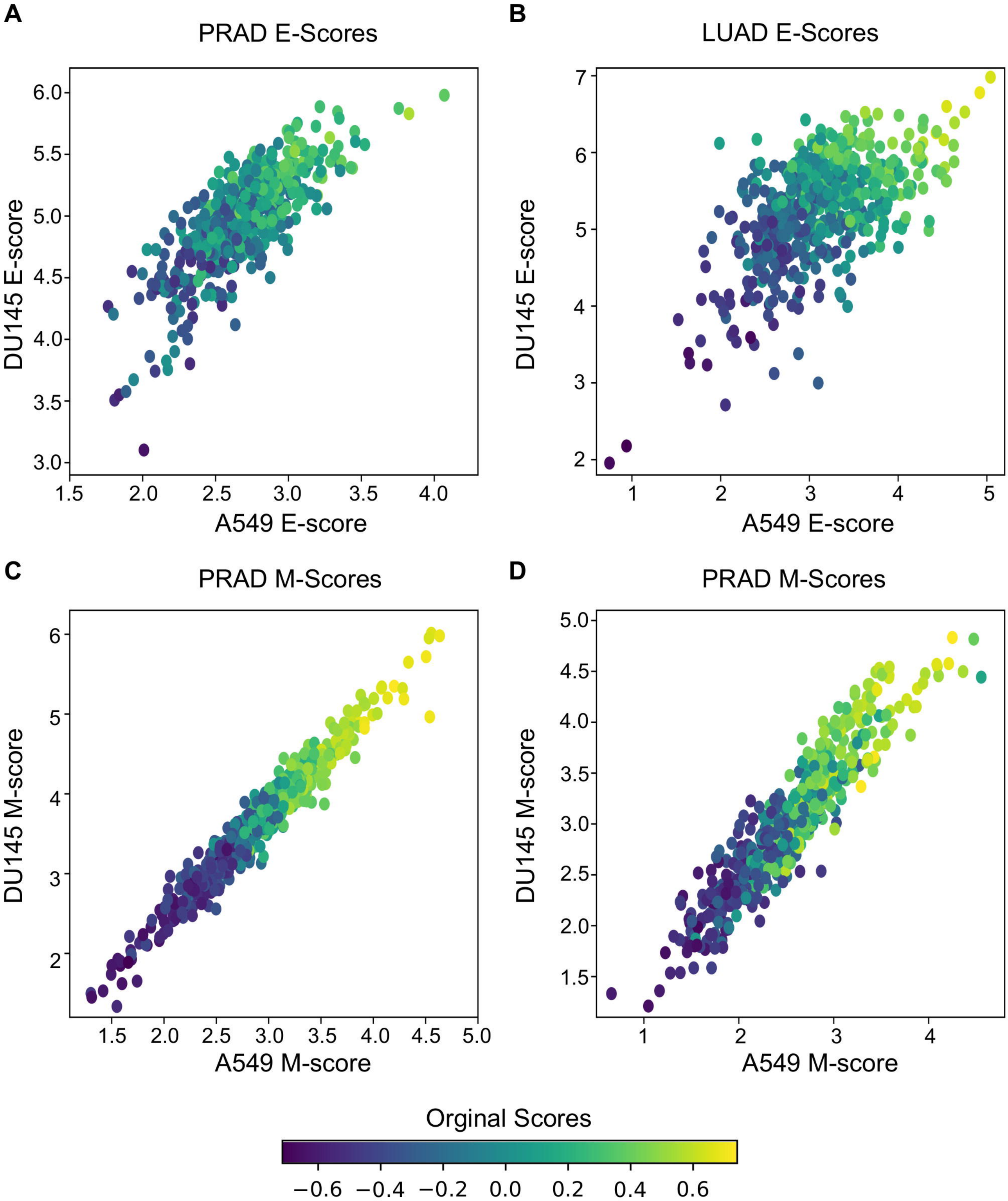
Ye, Y., Kang, X., Bailey, J., Li, C., and Hong, T. (2019). An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. PLoS Comput Biol *15*, e1006855.

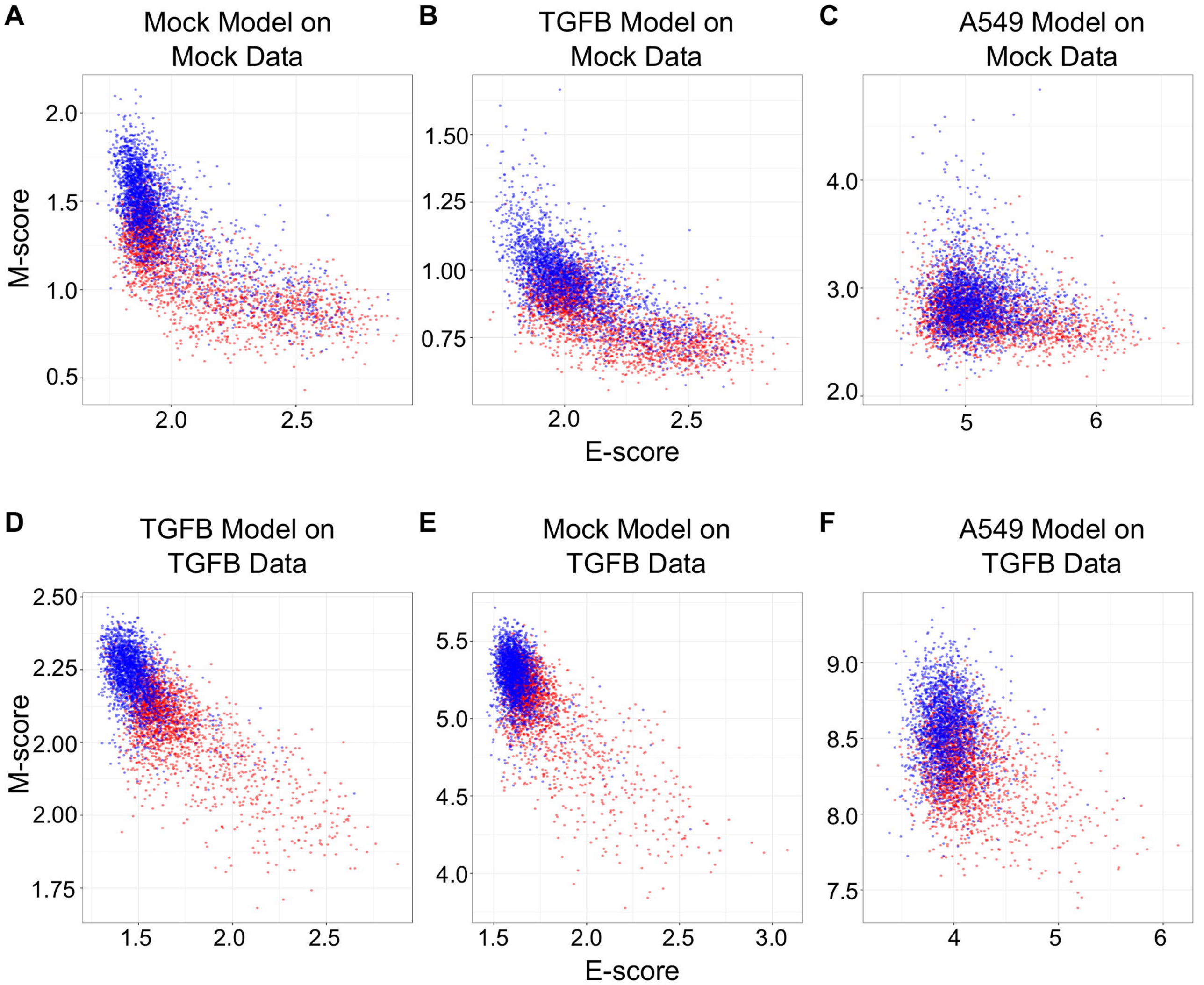
Yui, M.A., and Rothenberg, E.V. (2014). Developmental gene networks: a triathlon on the course to T cell identity. Nature Reviews Immunology *14*, 529-545.

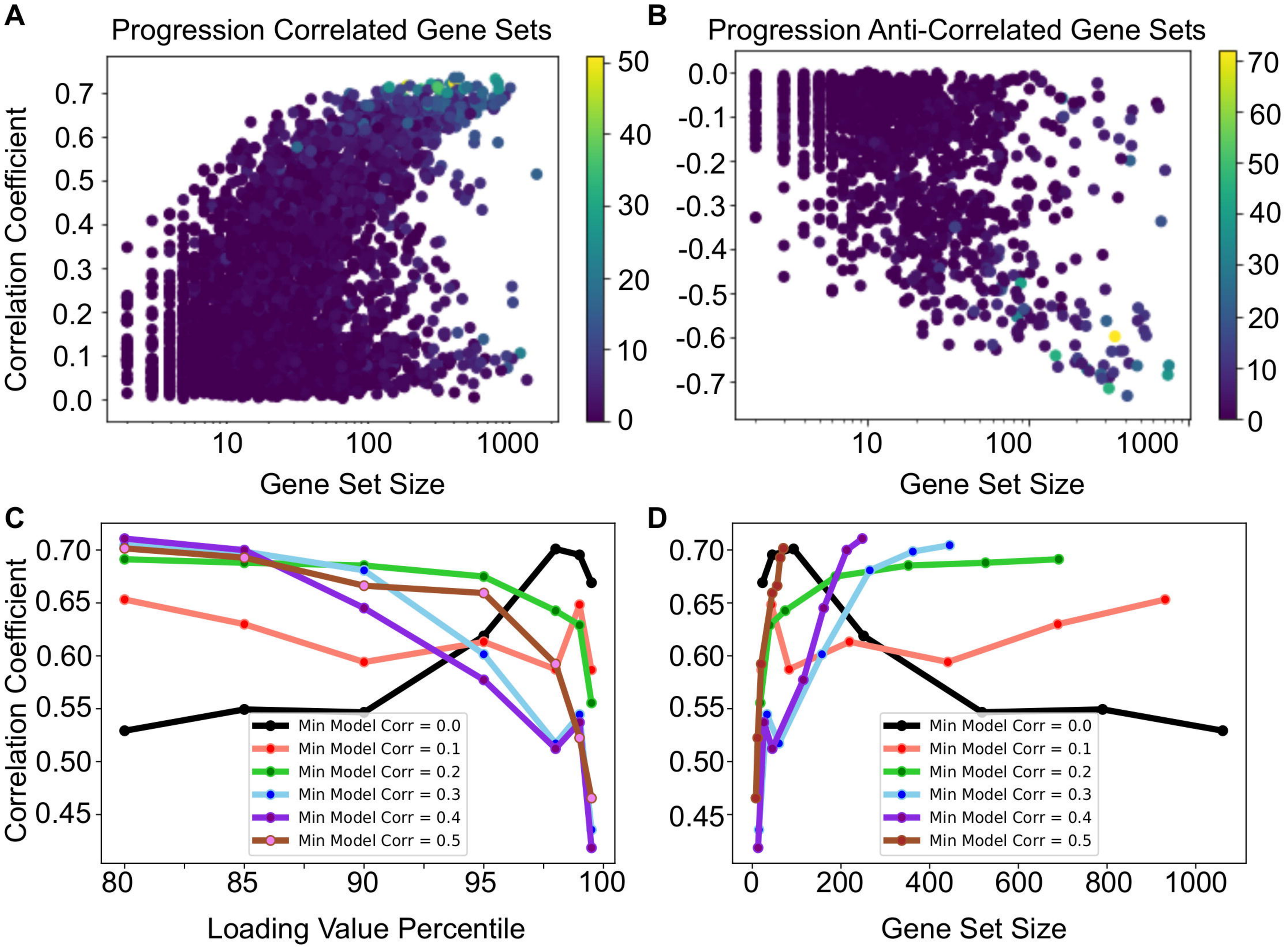

A
gsPCA
M-score
E-score
B
gsNMF
M-score
E-score
C
z-score
M-score
E-score
D
GSVA
M-score
E-score
E
Score Values
Time
0d
8h
1d
3d
7d
E
M
F
z-score
GSVA
gsPCA
gsNMF
E scores
M scores
f probability
0d v 8h
8h v 1d
1d v 3d
3d v 7d
Time

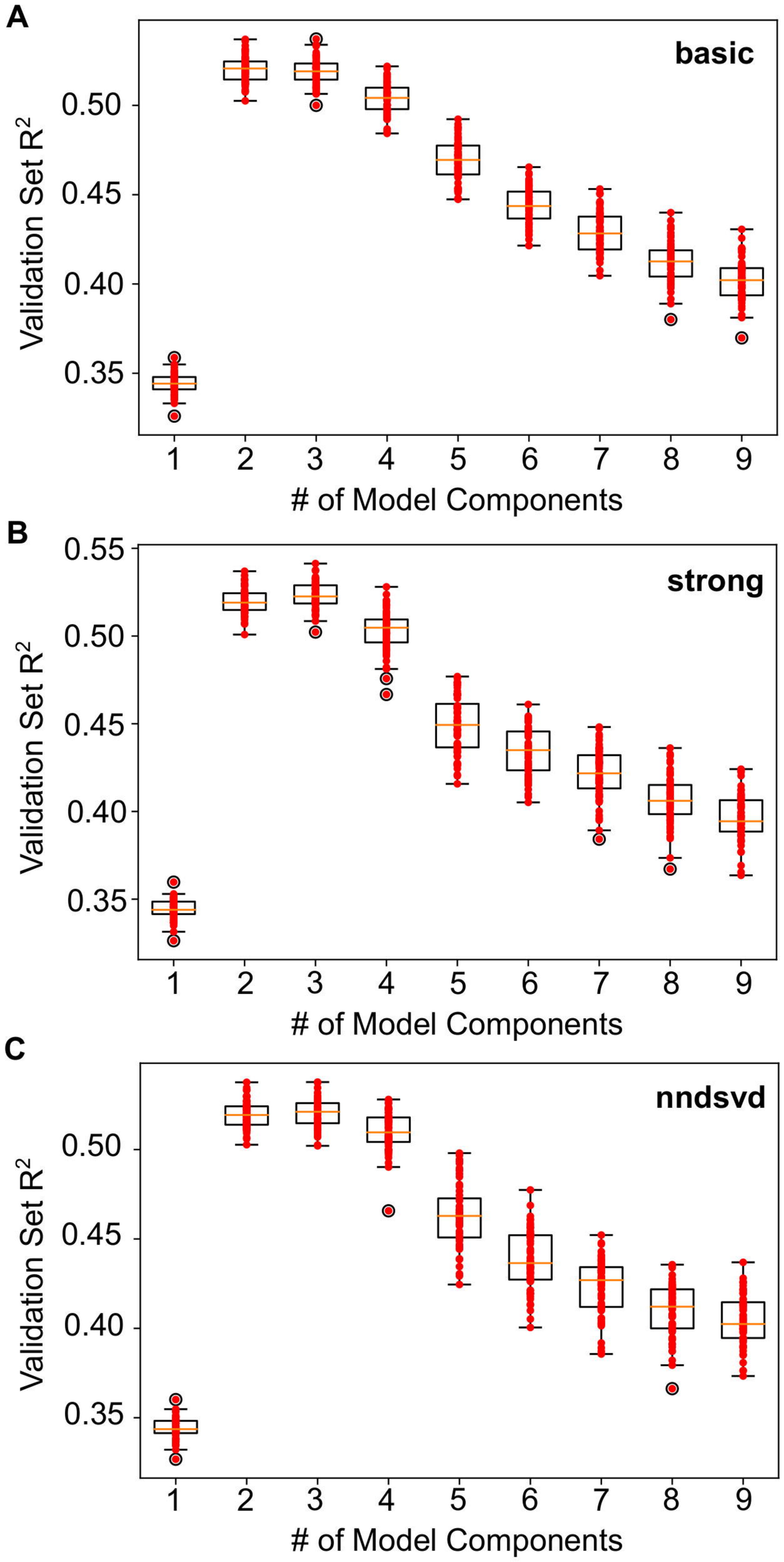

A
B
C
D
E
F
0d
8h
1d
3d
7d
rev. 8h
rev. 1d
rev. 3d
M-scores
E-scores
0.0
0.5
1.0
1.5
2.0
0
1
2

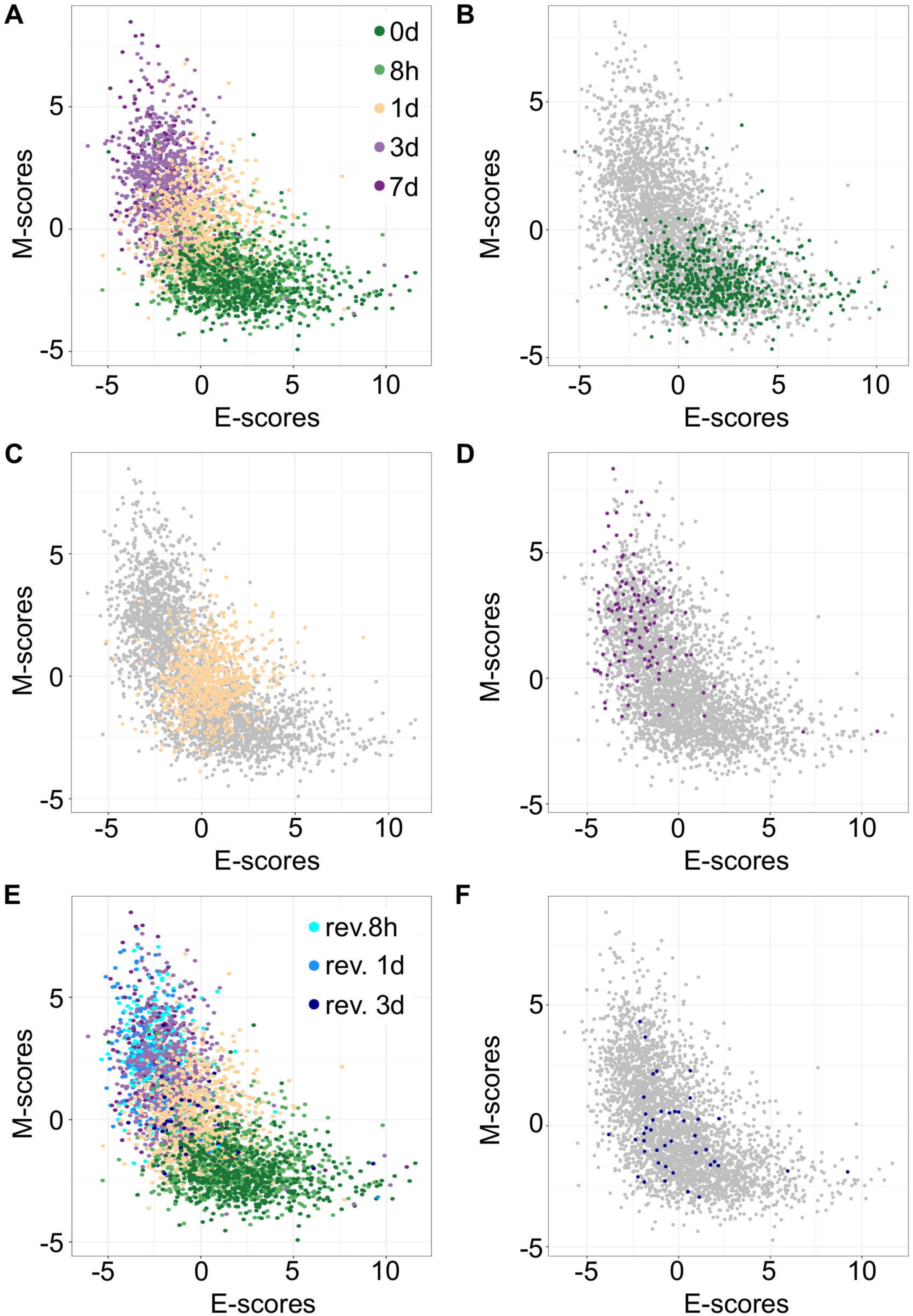

A
A549 Model on A549 Data
B
DU145 Model on A549 Data
C
A549 Model on DU145 Data
D
DU145 Model on DU145 Data
M-score
E-score
E
A549 E-scores
F
DU145 E-scores
G
A549 M-scores
H
DU145 M-scores
Scores from Transferred Mdeols
Scores from Original Data
Sample Time
0d
8h
1d
3d
7d

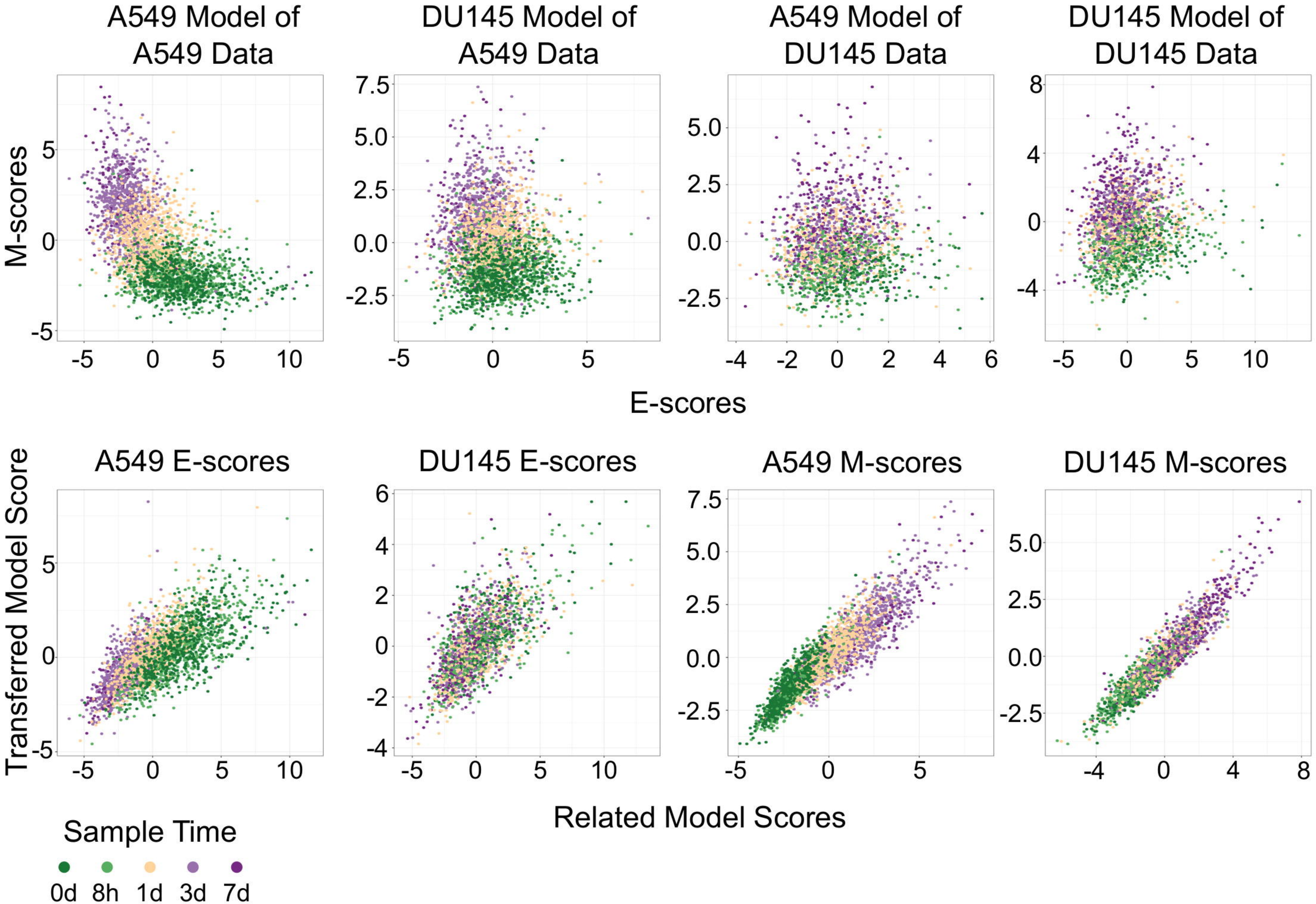

A
PRAD E-Scores
DU145 E-score
A549 E-score
B
LUAD E-Scores
DU145 E-score
A549 E-score
C
PRAD M-Scores
DU145 M-score
A549 M-score
D
PRAD M-Scores
DU145 M-score
A549 M-score
Orginal Scores
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6

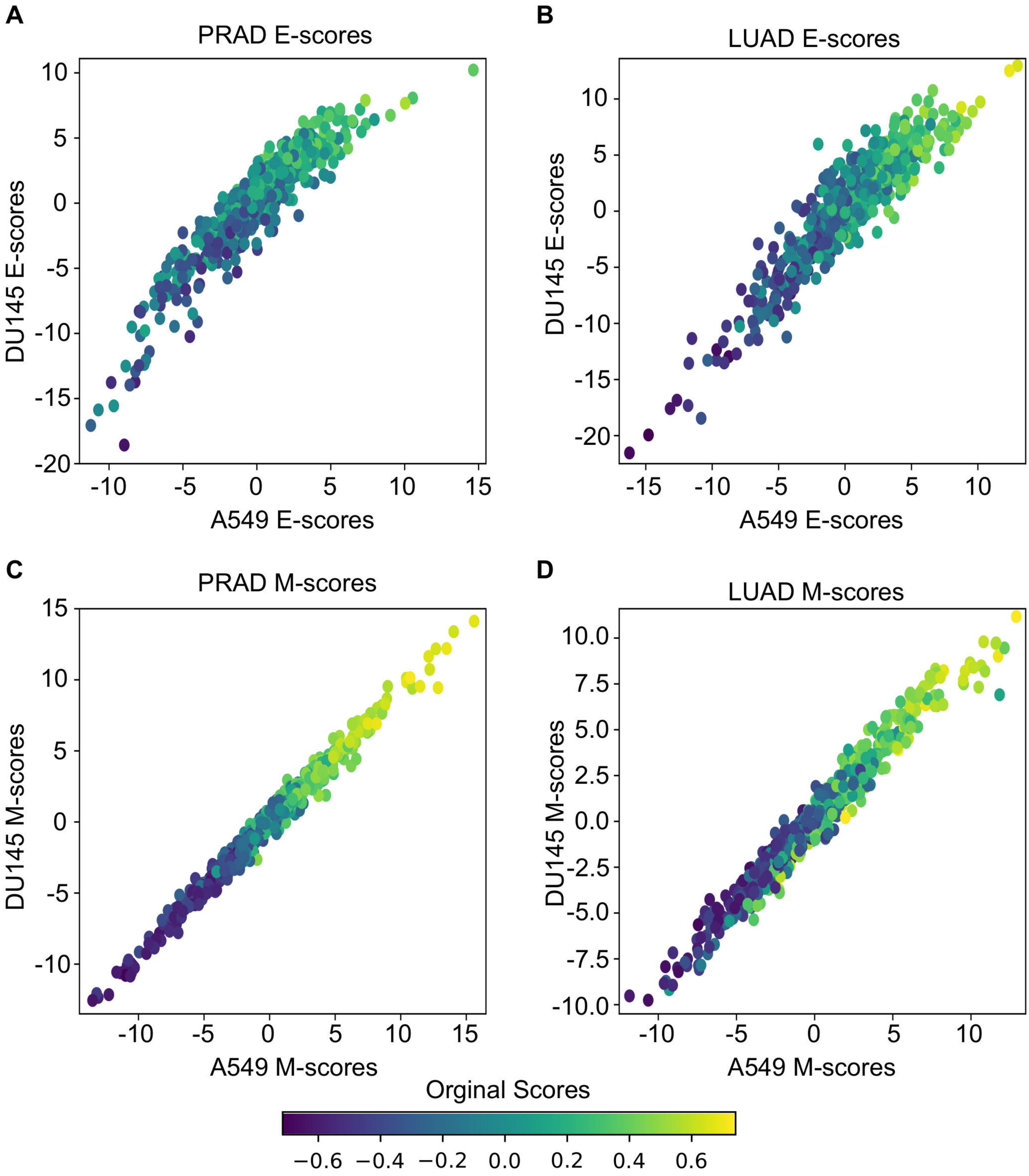

A
Mock Model on Mock Data
B
TGFB Model on Mock Data
C
A549 Model on Mock Data
D
TGFB Model on TGFB Data
E
Mock Model on TGFB Data
F
A549 Model on TGFB Data
M-score
E-score

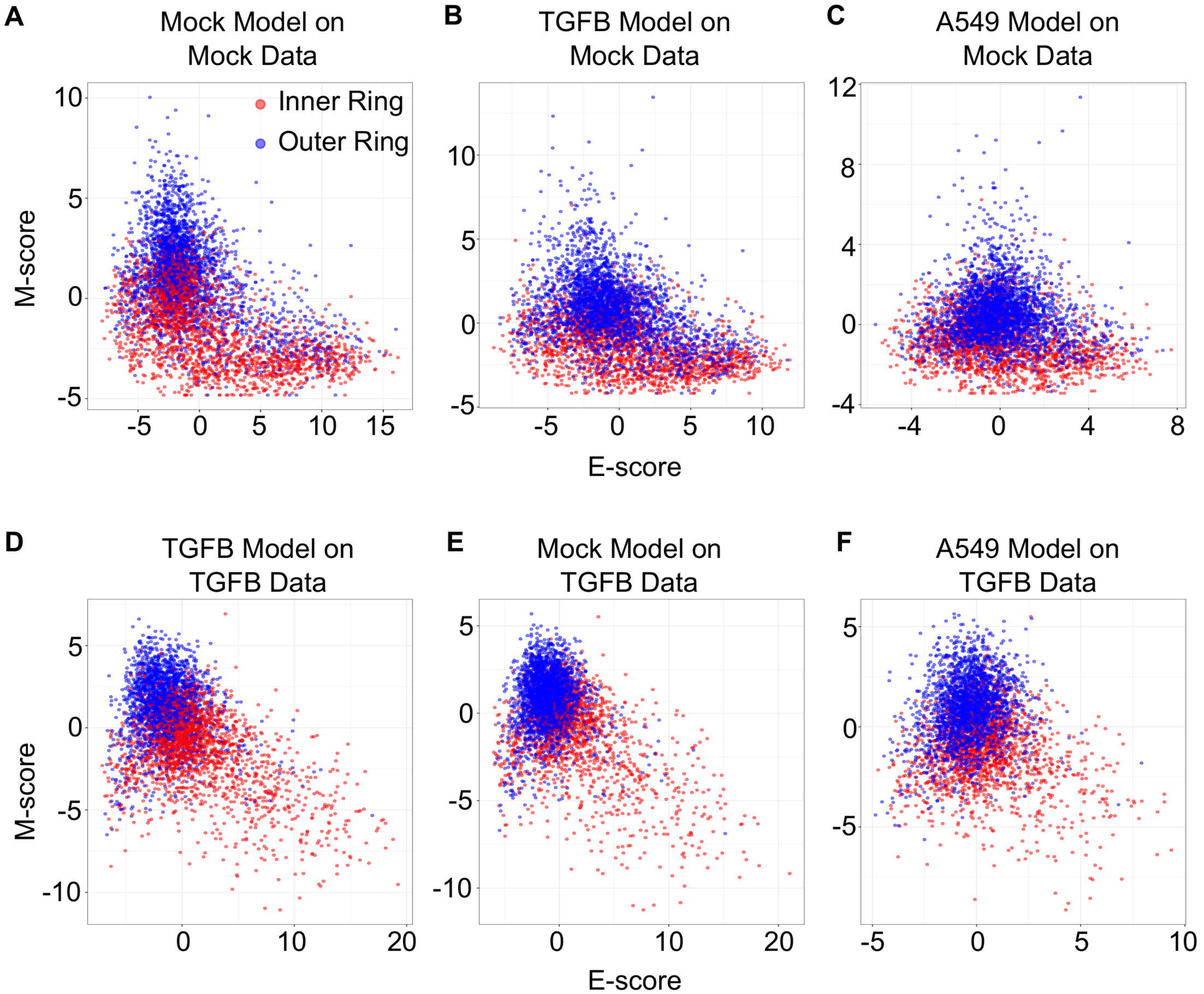

A
Progression Correlated Gene Sets
Correlation Coefficient
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
10
100
1000
Gene Set Size
50
40
30
20
10
0
B
Progression Anti-Correlated Gene Sets
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
10
100
1000
Gene Set Size
70
60
50
40
30
20
10
0
C
Correlation Coefficient
0.70
0.65
0.60
0.55
0.50
0.45
80
85
90
95
100
Loading Value Percentile
Min Model Corr = 0.0
Min Model Corr = 0.1
Min Model Corr = 0.2
Min Model Corr = 0.3
Min Model Corr = 0.4
Min Model Corr = 0.5
D
0.70
0.65
0.60
0.55
0.50
0.45
0
200
400
600
800
1000
Gene Set Size
Min Model Corr = 0.0
Min Model Corr = 0.1
Min Model Corr = 0.2
Min Model Corr = 0.3
Min Model Corr = 0.4
Min Model Corr = 0.5


A
basic
Validation Set R2
0.50
0.45
0.40
0.35
1
2
3
4
5
6
7
8
9
# of Model Components
B
strong
Validation Set R2
0.55
0.50
0.45
0.40
0.35
1
2
3
4
5
6
7
8
9
# of Model Components
C
nndsvd
Validation Set R2
0.50
0.45
0.40
0.35
1
2
3
4
5
6
7
8
9
# of Model Components


A
0d
8h
1d
3d
7d
M-scores
E-scores
B
M-scores
E-scores
C
M-scores
E-scores
D
M-scores
E-scores
E
rev.8h
rev. 1d
rev. 3d
M-scores
E-scores
F
M-scores
E-scores


A549 Model of A549 Data
DU145 Model of A549 Data
A549 Model of DU145 Data
DU145 Model of DU145 Data
M-scores
E-scores
A549 E-scores
DU145 E-scores
A549 M-scores
DU145 M-scores
Transferred Model Scores
Related Model Scores
Sample Time
0d
8h
1d
3d
7d


A
PRAD E-scores
DU145 E-scores
A549 E-scores
B
LUAD E-scores
DU145 E-scores
A549 E-scores
C
PRAD M-scores
DU145 M-scores
A549 M-scores
D
LUAD M-scores
DU145 M-scores
A549 M-scores
Orginal Scores
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6


A
Mock Model on Mock Data
Inner Ring
Outer Ring
M-score
B
TGFB Model on Mock Data
C
A549 Model on Mock Data
E-score
D
TGFB Model on TGFB Data
E
Mock Model on TGFB Data
F
A549 Model on TGFB Data
M-score
E-score