Supplementary Information for "Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories"

Pablo Cordero

Joshua M. Stuart

UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA

Supplementary Methods

Simulated data with correlated, trajectory-dependent noise

To simulate noise structures that correlate with the trajectory and that change throughout the progression (as would occur when gene networks rewire from one state to the next), we simulated two gene-gene networks, Net_{start} and Net_{end} using a Watts-Strogatz model with 3 connected neighbors and a 0.2 probability of edge reassignment. For a constant factor α and for each network, we simulated a covariance matrix Σ_{net} as follows: for each edge in the network and for each element in the diagonal, we set the corresponding entry in Σ_{net} to a random value drawn from a Gaussian distribution centered at zero and with standard deviation α. To ensure positive-definitiveness, we squared all elements in the diagonal and added to each a factor $n \times \alpha$ with n being the minimum non-negative integer such that Σ_{net} is numerically positive definite. We then simulated a covariance matrix function with initial values set to the the associated covariance matrix of Net_{start} and with a final matrix values corresponding to the associated covariance matrix of Net_{end}. To obtain this simulated matrix function while maintaining positive-definitiveness across the trajectory, we started by decomposing the covariance matrices of Net_{start} and Net_{end} to their respective Cholesky decompositions, C_{start} and C_{end}, and generated 100 quadratic polynomials in the $[0, 1]$ interval, forcing their starting and ending values to be entries from C_{start} and C_{end}. These polynomials then gave a lower triangular matrix function C_{sim} from which a covariance function could be computed: $\Sigma_{sim}(t) = C_{sim}(t)^T C_{sim}(t), t \in [0, 1]$. To obtain the mean function $\mu(t)$ of this morphing distribution, we generated 10 random quadratic polynomials, one for each entry of μ, and multiplied their values with a constant factor β.

Finally, we obtained 100 samples from this distribution by first sampling 100 timepoints from the $[0, 1]$ interval and then sampling from the Gaussian distributions given by the simulated covariance and mean functions at each timepoint. We performed 6 of these simulations with increasing noise-to-signal ratio $\frac{\alpha}{\beta}$ but maintaining the same noise structure (see Fig1B of main text). See locally-linear embeddings of these simulations in Supplemental Fig1.

Simulated data and scripts can be found in the supplemental data package supplementary_data.zip.

Supplementary Results

In silico benchmarks with non-heteroscedastic covariance estimation

To assess whether heteroscedastic covariance estimation is necessary for robust psuedotime estimation in our model, we performed SCIMITAR inference in our *in silico* benchmark by fixing the covariance function to a constant matrix function in two settings. In the first setting, we set all covariances to the identity matrix, essentially reducing our model to a principal curve smoothed differently by each functional class (see Supplemental Fig2A). The results were equal or inferior to the heteroscedastic SCIMITAR model (compare to Fig 2 in the main text). In particular, the Gaussian Process functional class tended to under-smooth and produce significantly inferior results, especially in the highly non-linear setting of our correlated noise benchmark (Fig2A left). In the second setting, we set all covariances to the average covariance matrix at all pseudo time-points at each iteration of coordinate ascent, summarizing local inferences into one global value. This resulted in a homoscedastic model that was informed by heteroscedastic inferences. These models were again equal or inferior to their heteroscedastic counterparts and especially under-performed in high uncorrelated noise and low smoothing (with Gaussian Processes) settings (see Supplemental Fig2B). Furthermore, homoscedastic models by definition are not able to track changing co-expression associations between genes, limiting their power for characterizing the functional changes between cell states across the trajectory.

Neurodifferentiation-associated genes

Exhaustive list and plots for neurodifferentiation progression associated genes and co-regulatory state modules are given in the supplementary_data.zip package.

Supplemental Figures

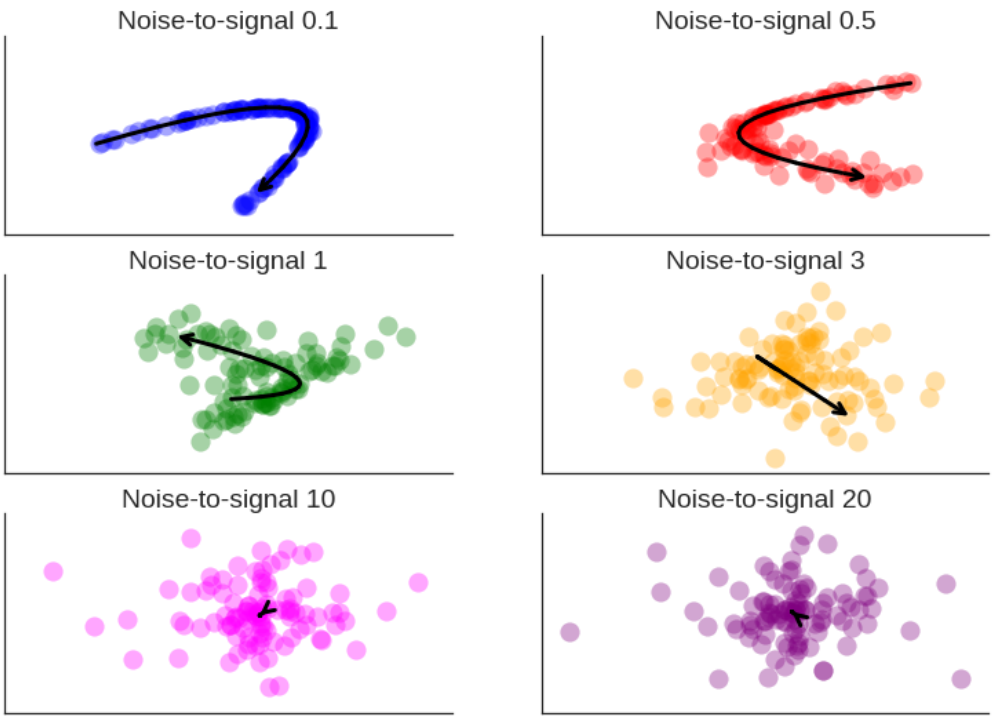

Figure 1. Locally linear embedding visualizations of the correlated noise simulations, as noise-to-signal $\frac{\alpha}{\beta}$ (see Suppl. Methods) increases

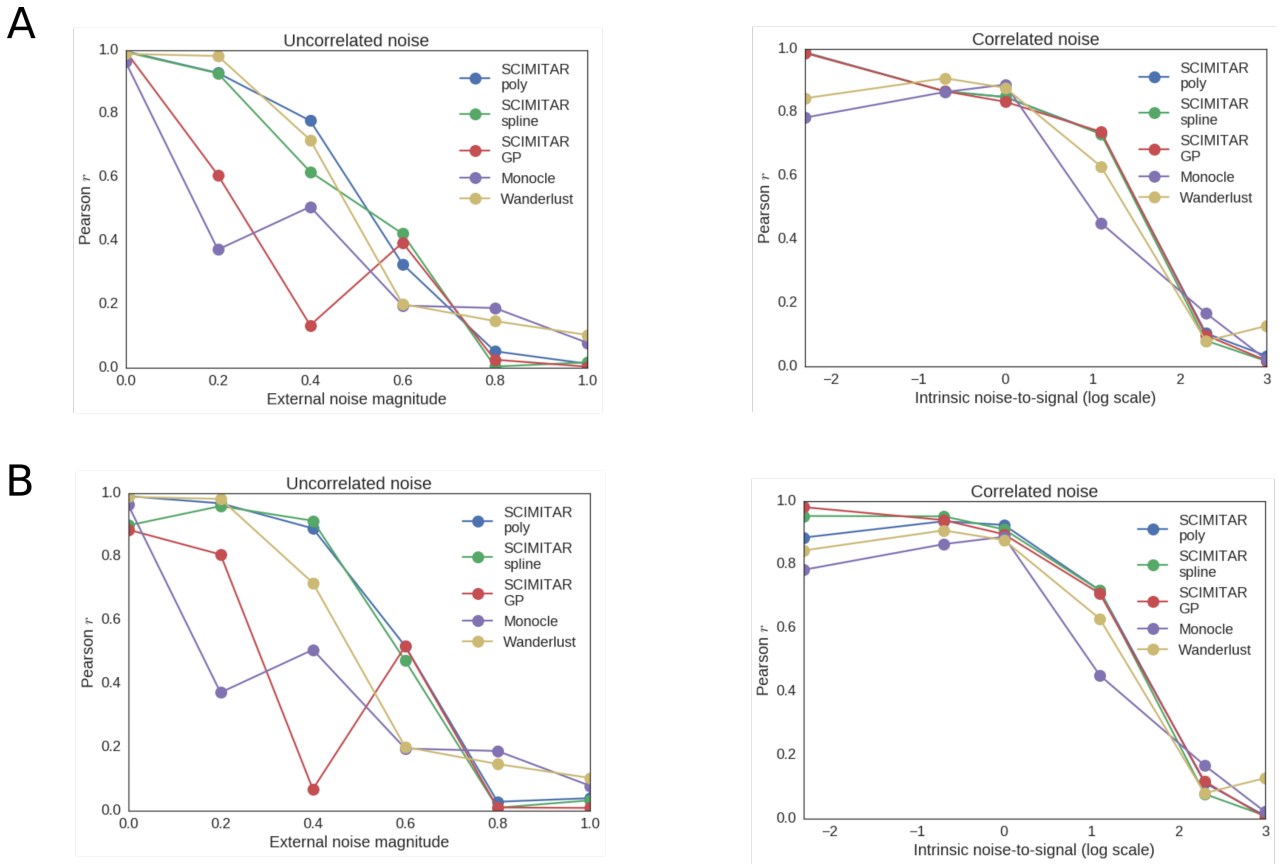

Figure 2. *In silico* benchmark results with non-heteroscedastic covariance structure. A. Benchmark results when covariances at each pseudo time-point are set to the identity matrix. B. Benchmark results when covariances at each pseudo time-point are set to a global average of local covariances. Results are equal or inferior to the full heteroscedastic models (see Figure 2 in main text).