Using Random Matrix Theory to Analyze Power Law Signatures in scRNA-seq data

Archishman Raju

NCBS

24 April 2025

With Suvranil Ghosh and Shaon Chakrabarti

S. Ghosh, S. Chakrabarti and AR bioRxiv:2025.01.15.633183

Inheritance of cell state heterogeneity

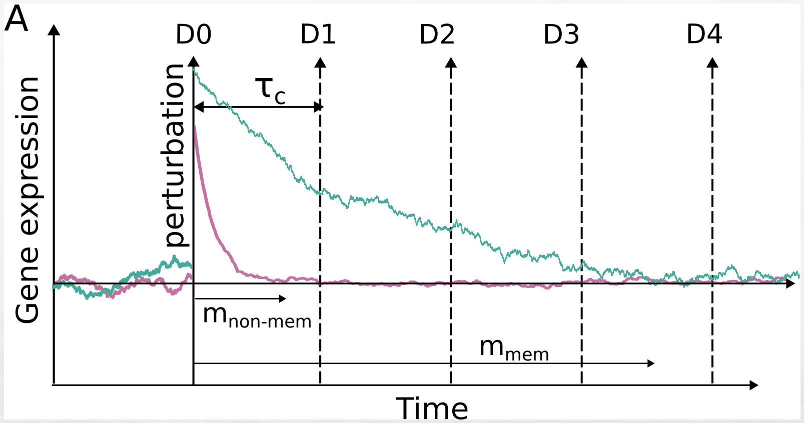
Inheritance

DNA

RNA

Protein

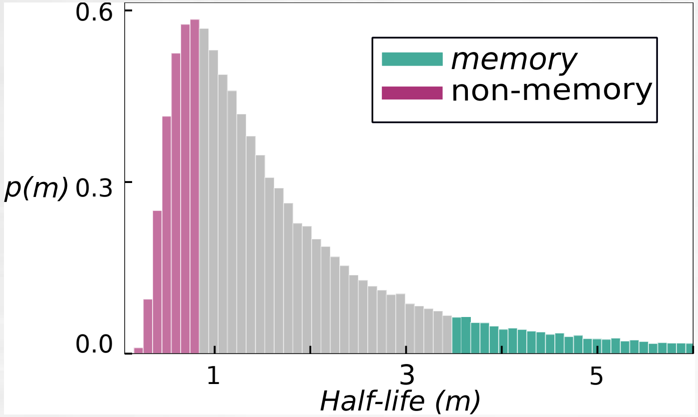
"Central Dogma of Molecular Biology"

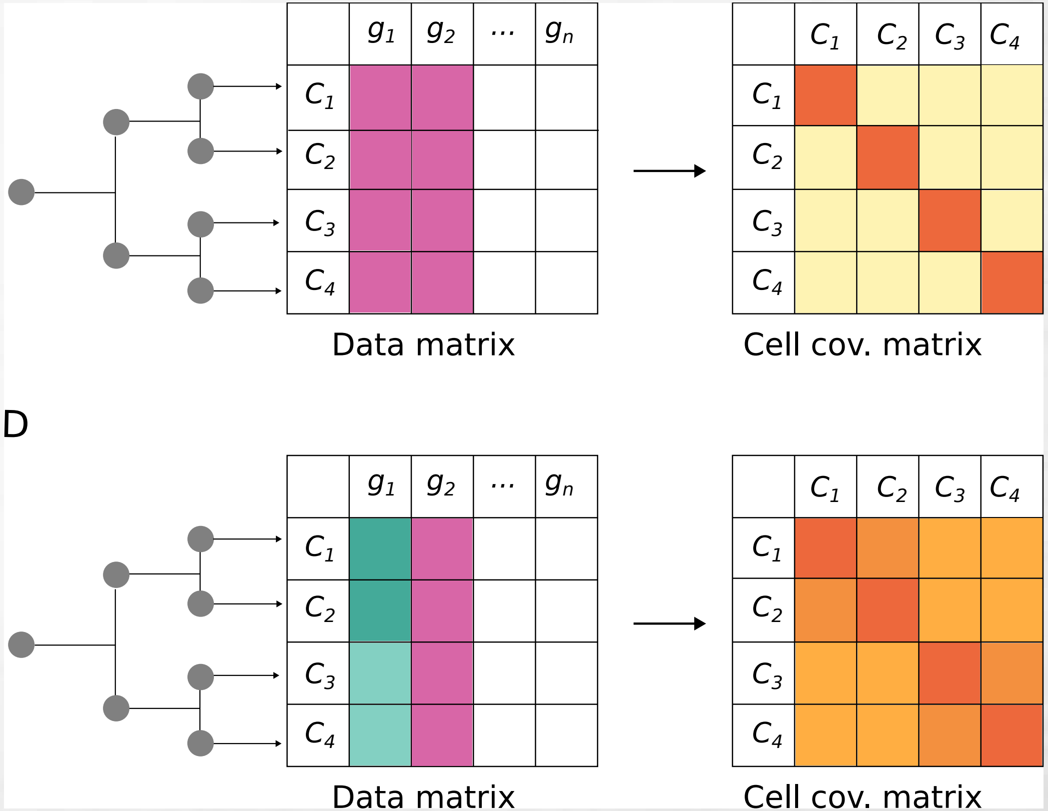


Genetically identical cells can be in different cell states depending on their RNA or protein concentration.

This expression can be inheritable

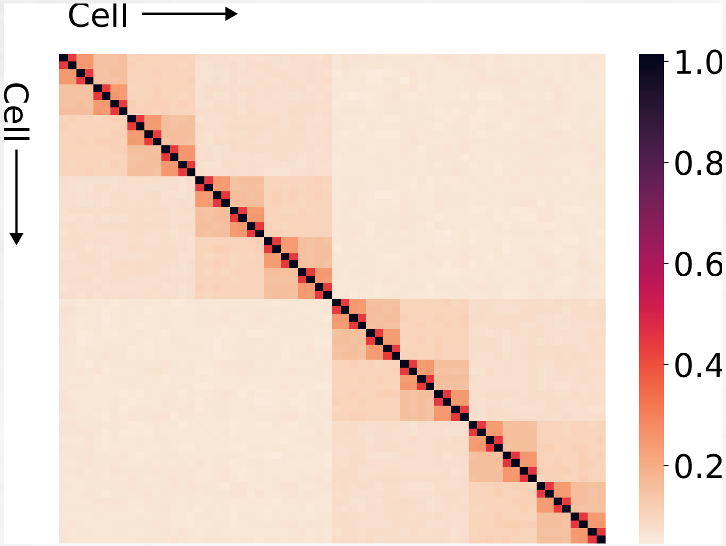
Lineage Correlations are in Cell-Covariance



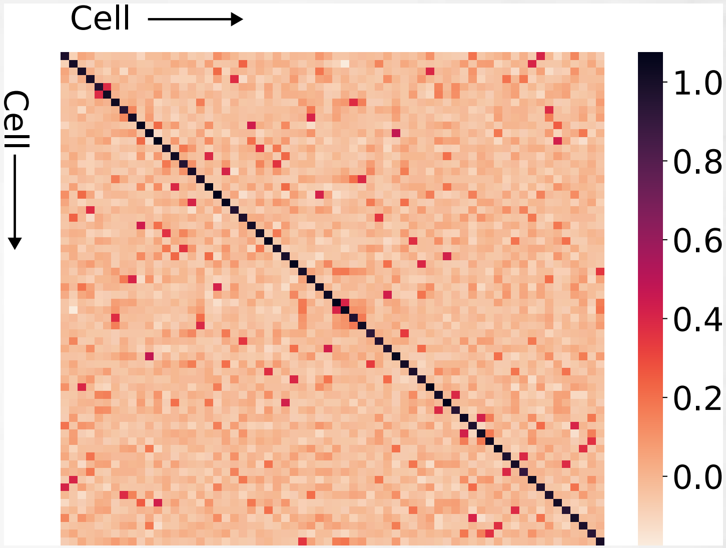


Inheritance is dependent on time-scales. The cell covariance matrix carries signatures of “memory genes”

Can a sample identify memory genes?



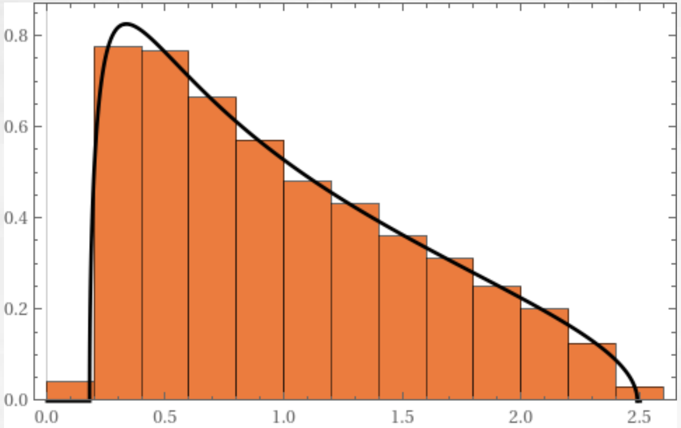
True Covariance Matrix

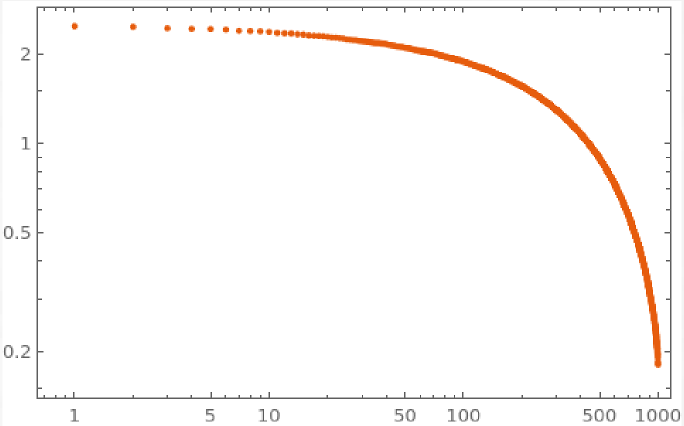


Sample Covariance Matrix

Can we identify the time-scale of the dynamics of gene expression from a single snapshot?

The null model is M-P





$$\rho(x) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}$$

Marchenko-Pastur Distribution

Parameters are a function only of m/n for an m x n rectangular matrix

Deriving Eigenvalues of Cell Covariance

True Covariance Matrix

$$\begin{pmatrix} n_g & \sum_i \exp(-2\mu\tau_c) & .... & \sum_i \exp(-2b\mu_i\tau_c) \\ \cdots & \cdots & & \\ \cdots & \cdots & & \\ \sum_i \exp(-2b\mu_i\tau_c) & \sum_i \exp(-2b\mu_i\tau_c) & \cdots & n_g \end{pmatrix}$$

Eigenvalues of True Covariance Matrix

$$\lambda_i = \begin{cases} 1 + \sum_{j=1}^{b} 2^{j-1}\alpha_j & \text{when } i = 0, \\ 1 + \sum_{j=1}^{b-i} 2^{j-1}\alpha_j - 2^{b-i}\alpha_{b-i+1} & \text{when } 0 < i < b, \\ 1 - \alpha_1 & \text{when } i = b, \end{cases}$$

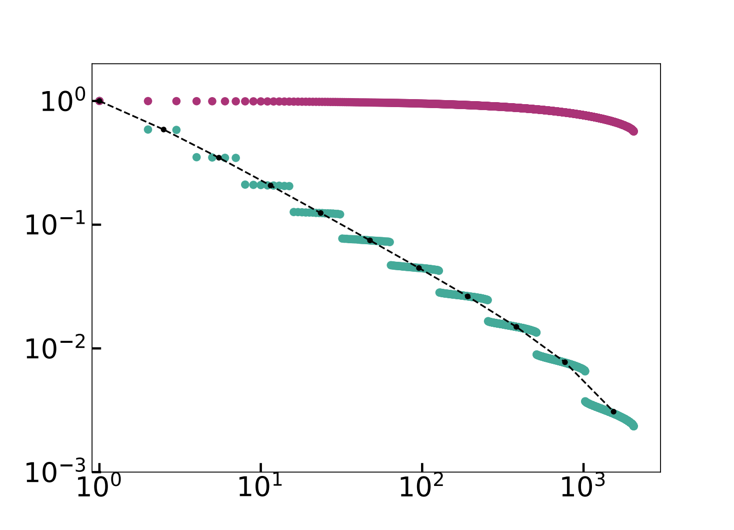
Stieltjes Transform can be used to derive sample covariance

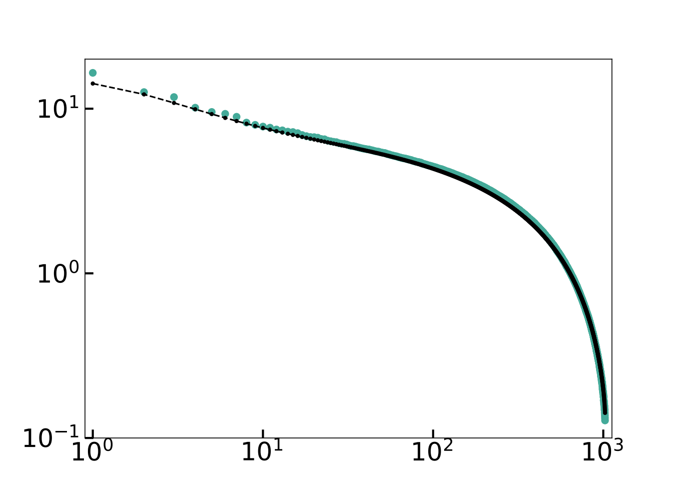
$$G(z) = \int_{-\infty}^{\infty} \frac{P_S(\lambda)}{\lambda - z}\, d\lambda$$

$$\frac{1}{-G(z)} = z - \frac{n_c}{n_g} \int_{-\infty}^{\infty} \frac{\lambda}{1 + \lambda G(z)}\, dP_T(\lambda)$$

Marchenko & Pastur (1967)
Qin & Colwell (2018)

Eigenvalues show a power-law form





True Covariance Matrix

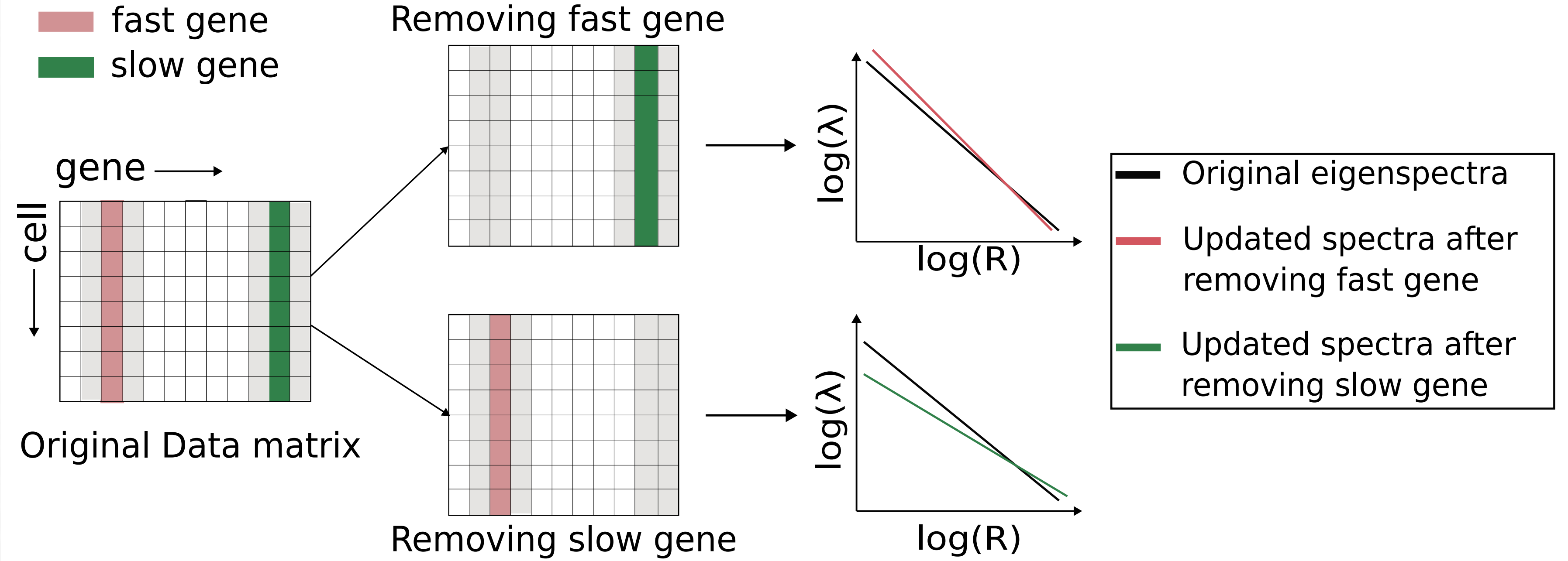
Green: Degenerate Power-Law distributed eigenvalues

Purple: Slow division gives back MP

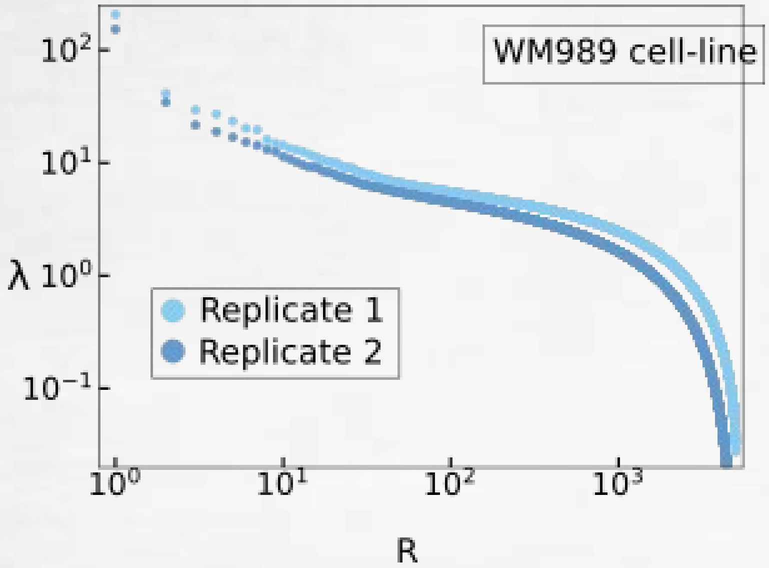
Black: Analytics

Sample Covariance Matrix numerics (green) and analytics (black) from ST + Order Statistics

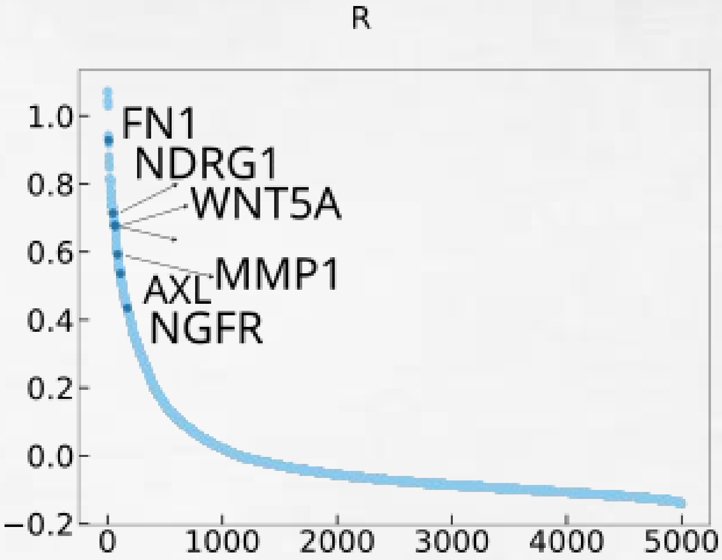
Simple Algorithm to Identify Memory Genes



Applying to scRNA seq data



Real Data is Distinctly non MP and looks power-law like in a regime



We can identify previously suggested memory genes in the literature from single snapshots

Shaffer et al. Cell 2020