Stator: Higher-order expression dependencies finely resolve cell (sub)type and state in single cell data

Yuelin Yao[1, 2], Abel Jansma[2, 3], Jareth Wolfe[2], Luigi Del Debbio[3], Sjoerd Beentjes[2, 4], Chris Ponting[2], Ava Khamseh[1, 2, 3]

[1]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom.
[2]MRC Human Genetics Unit, Institute of Genetics & Cancer, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom.
[3]Higgs Centre for Theoretical Physics, School of Physics Astronomy, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom.
[4]School of Mathematics, University of Edinburgh Edinburgh EH9 3JZ, United Kingdom.

INSTITUTE OF GENETICS & CANCER

Medical Research Council

THE UNIVERSITY of EDINBURGH

CANCER RESEARCH UK

1 Motivation

- Advances in scRNA-seq techniques are resolving cell (sub)types among complex cell populations by clustering in reduced dimensional transcriptome space.
- Cell states representing particular cellular activities such as cell cycle are better described as a continuous spectrum. Identifying cell states is commonly done by extracting activity gene expression program (GEP) with factorisation approaches, i.e., NMF.
- We introduce ***Stator***, a novel method that finely resolves cell types, subtypes and states among cells from higher-order gene expression dependencies.

2 Higher-order interactions

Definition: A pair of genes $\{X_i, X_j\} \in X$ has a pairwise interaction I_{ij} where:

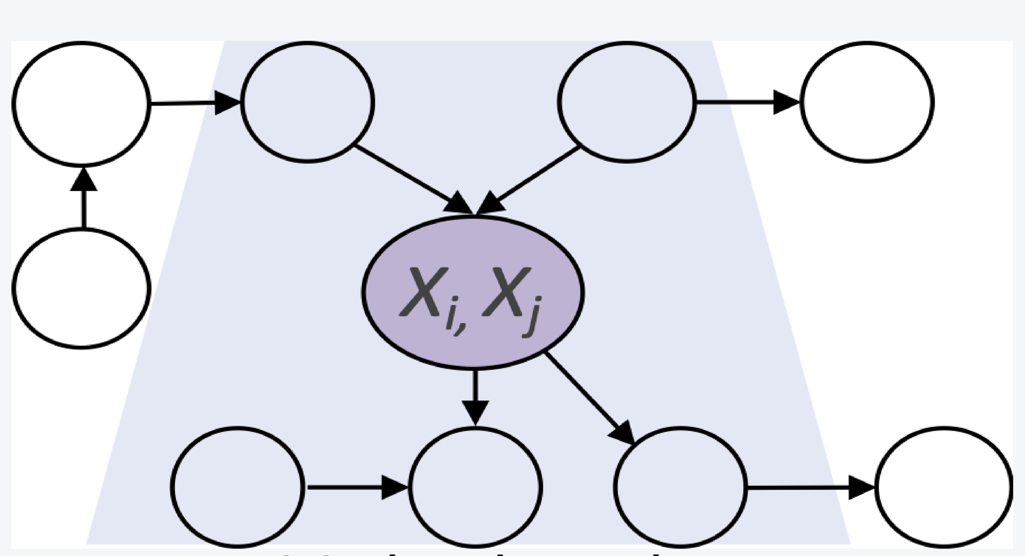
$$I_{ij} = \log \frac{p(X_i=1, X_j=1 \mid \underline{X}=0)\, p(X_i=0, X_j=0 \mid \underline{X}=0)}{p(X_i=1, X_j=0 \mid \underline{X}=0)\, p(X_i=0, X_j=1 \mid \underline{X}=0)}$$

This definition has the following properties:

- It is symmetric: $I_{ij} = I_{ji}$.
- It is model-independent and can be directly estimated from observations [1].
- It conditions on the Markov boundary, a minimal subset conditioned on which X_i and X_j become independent of other genes.
- The Markov boundary is identified by an iterative MCMC method for causal discovery [2].
- If two genes are conditionally independent then I_{ij}=0.
- The method can be extended to higher-order interaction by taking n'th derivatives of $\log p(X)$:

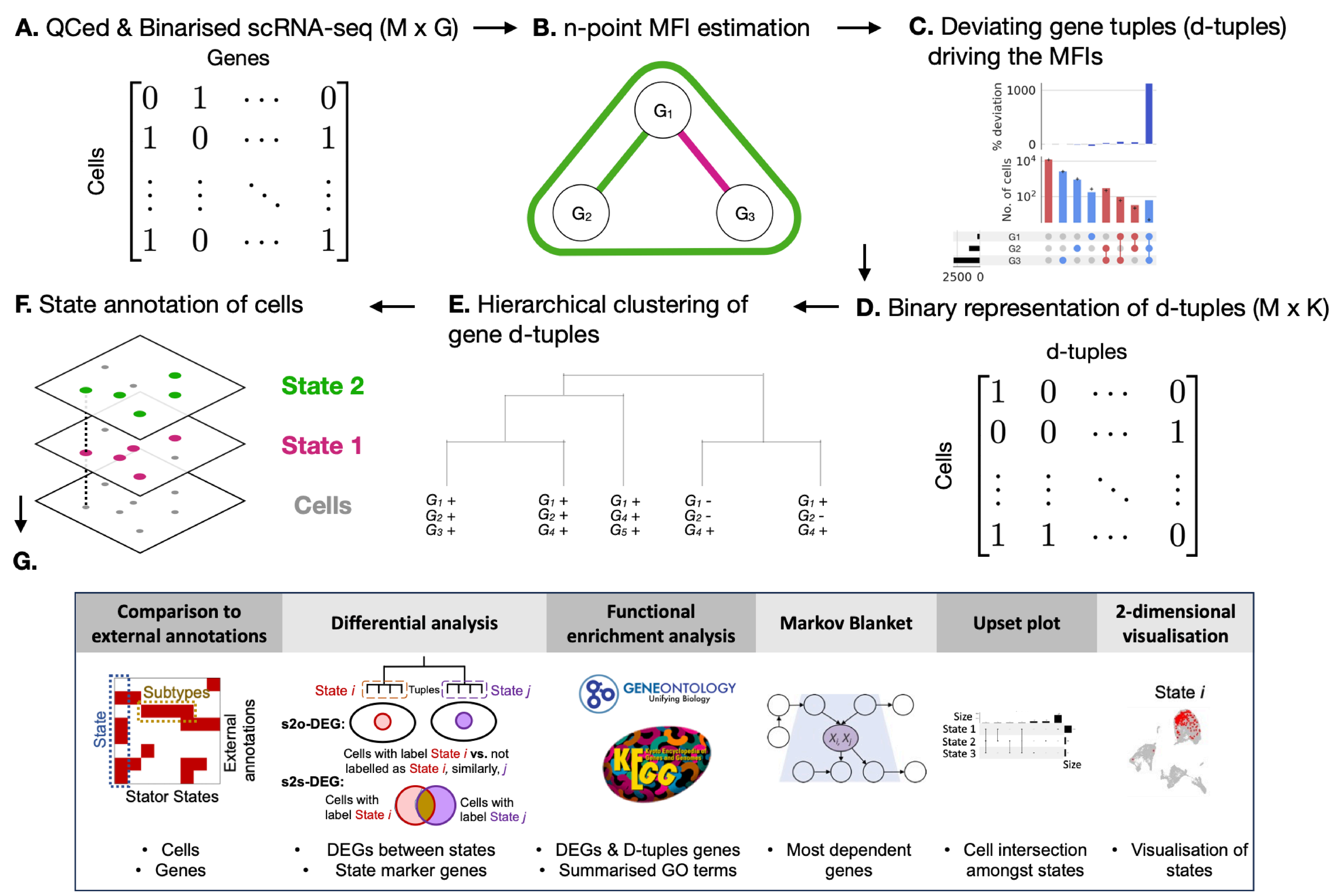
$$I_{ijk} = \log \frac{p(1,1,X_k \mid \underline{X})\, p(0,0,X_k \mid \underline{X})}{p(0,1,X_k \mid \underline{X})\, p(0,1,X_k \mid \underline{X})}$$

$$= \log \left(\frac{p(1,1,1 \mid \underline{X})\, p(0,0,1 \mid \underline{X})\, p(0,1,0 \mid \underline{X})\, p(1,0,0 \mid \underline{X})}{p(0,0,0 \mid \underline{X})\, p(0,1,1 \mid \underline{X})\, p(1,1,0 \mid \underline{X})\, p(1,0,1 \mid \underline{X})} \right)$$



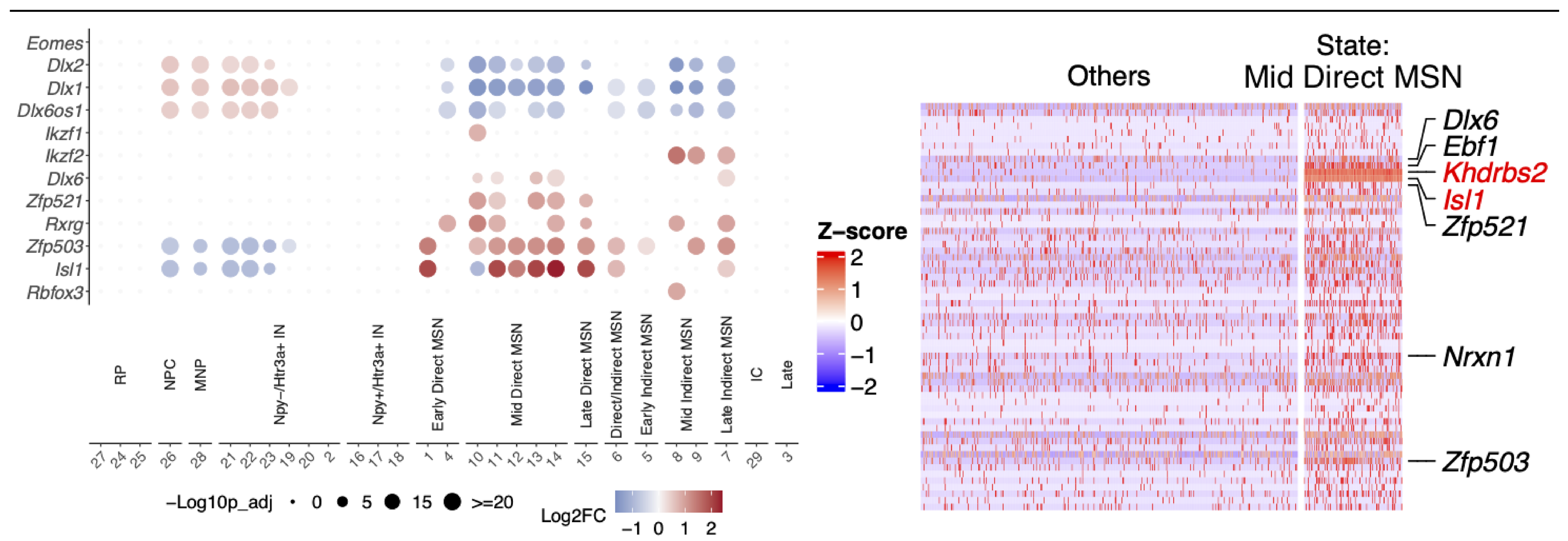
Markov boundary

3 *Stator* workflow



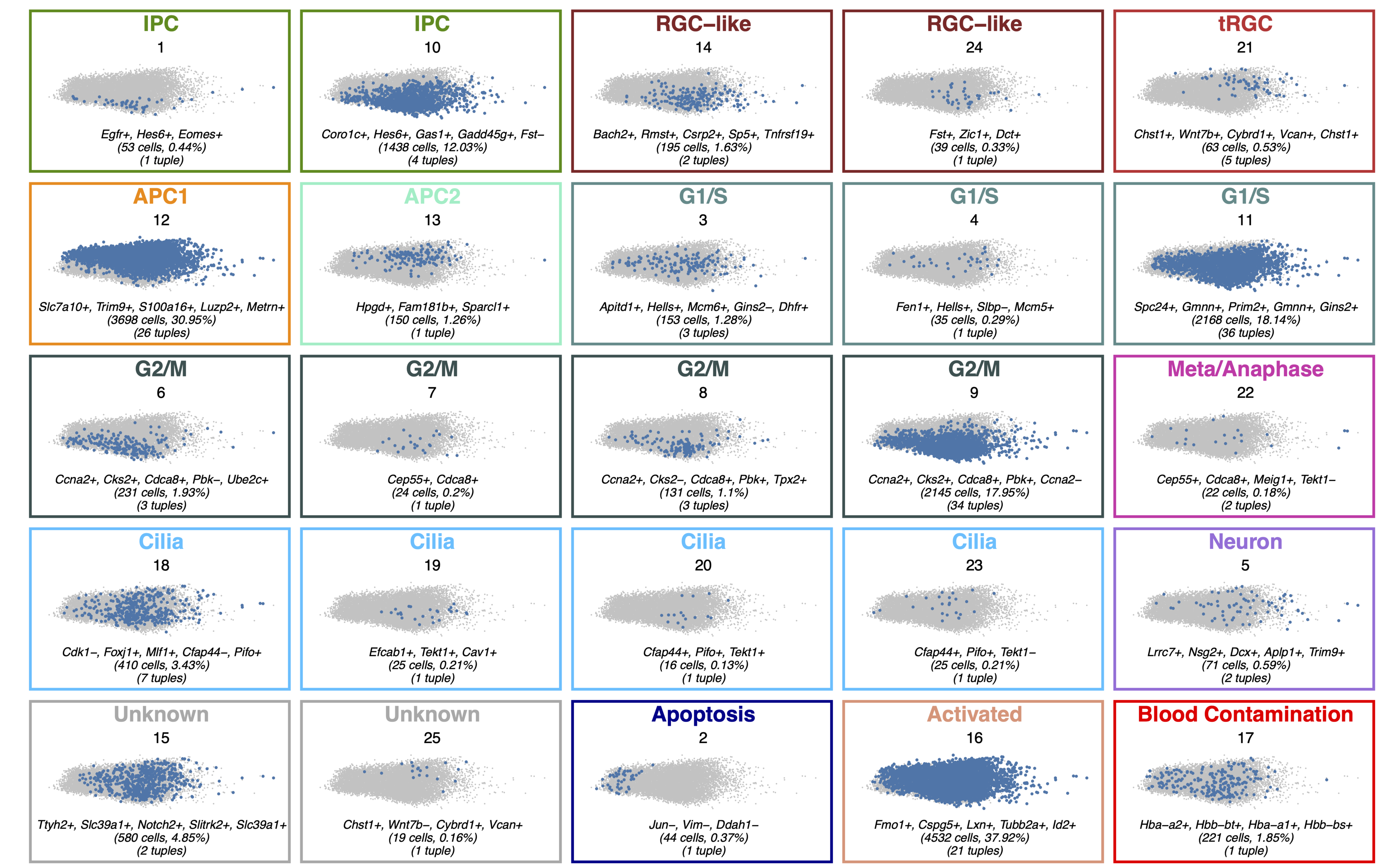
Downstream analysis with interactive Shiny App

https://shiny.igc.ed.ac.uk/MFIs/

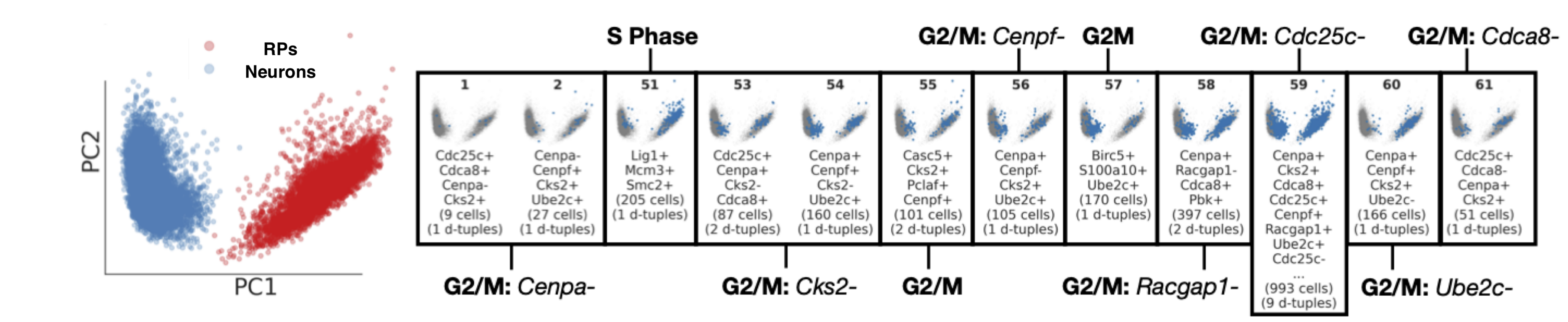


- Stator successfully distinguished striatal medium spiny neurons (MSN) from interneurons by known marker genes' expression.
- It further separated MSNs into their two known sub-types, Direct or Indirect pathway cells.

4 Results



Stator identified states in seemingly homogeneous embryonic radial glial precursors.



Stator can also identify cells in G1/S or G2/M phases within an admixture of two cell types, neurons and RPs.

5 Conclusion

- Stator differentiates cells by primary (cell type), secondary (sub-type) and tertiary (cell state, activity, cell cycle phase, or maturity) markers.
- Stator results show that a wealth of biological information can be inferred from the higher-order statistics of single-cell expression data.

References

1. Beentjes, S. V. & Khamseh, A. Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium. *Physical Review E* **102**, 053314 (2020).
2. Kuipers, J. *et al.* Efficient sampling and structure learning of Bayesian networks. *Journal of Computational and Graphical Statistics*, 1–12 (2022).