Computational methods for single-cell omics across modalities

Mirjana Efremova[1], Sarah A. Teichmann[1,2]

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK
[2] Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0EH, UK

Corresponding authors

Correspondence to Sarah Teichmann st9@sanger.ac.uk

Abstract

Single-cell omics approaches provide high resolution data on cellular phenotypes, developmental dynamics and communication networks in diverse tissues and conditions. Emerging technologies now measure different modalities of individual cells, such as genomes, epigenomes, transcriptomes and proteomes, in addition to spatial profiling. Combined with analytical approaches, these data open new avenues for accurate reconstruction of gene regulatory and signaling networks driving cellular identity and function. Here, we summarise computational methods for analysis and integration of single-cell omics data across different modalities and discuss their applications, challenges, and future directions.

Introduction

Single cell technologies are providing a means to generating detailed cellular maps of diverse tissues, in both healthy and pathogenic conditions. But how a cell adopts a certain state and which are the factors driving this decision is challenging to answer with a single modality such as single cell RNA sequencing (scRNA-seq) alone. To disentangle the complexity of gene regulatory and cell-cell communication networks that drive cell functions and responses, measuring multiple modalities of the multivariate phenotypic and genetic cellular state is extremely powerful. These modalities include DNA, chromatin, gene expression and protein, and the spatial location of each cell in its tissue microenvironmental context. However, each technology measures only particular aspects of cellular identity and has unique strengths and weaknesses.

Increased throughput of many new single-cell techniques catalyses development of innovative computational methods that use this multi-modal omics data to integrate and characterize multiple functional measurements in a biologically meaningful manner, allowing not only cell type classification, but also deeper insights into cell phenotype, interactions, and spatial organisation.

A major challenge of integrative analysis lies in reconciling the heterogeneity observed across individual datasets, while overcoming the extensive amount of missing data inherent in single cell sequencing experiments. In the majority of cases, the datasets are unpaired; different modalities are not profiled from the same cells but from cells sampled from the same sample or tissue. Integration methods for unpaired single-cell omics data aim to project multiple measurements of molecular information into a common latent space in order to assemble multiple modalities into an integrated reference, or use transfer learning to fill in missing modalities. Different inference algorithms for latent space projection are currently employed, for instance canonical correlation analysis[1], non-negative matrix factorization[2], or variational autoencoders[3]. Often, scRNA-seq serves as a common reference, facilitating integration across multiple technologies and modalities (Figure 1).

Here, we highlight state-of-the-art computational methods for multi-modal omics analysis and discuss their applications, including lineage relationships, gene regulatory and signaling networks inference and spatial context. We anticipate that single-cell omics technologies will bring exiting possibilities to dissect the regulatory and functional relationships of molecules within and between cells and construct more holistic maps of cells in health and disease.

Genotype to phenotype

With recent advances in single-cell technologies, simultaneous measurement of transcriptome and genome is becoming possible, which offers an opportunity to link genotype and phenotype. Methods for integration of DNA and transcriptomics data can determine the phenotypic impact of genetic variants.

A direct application of the DNA and RNA link is to enable inference of DNA-based cellular lineages and hierarchies while providing cell type and state identity from the same cells, for instance from CRISPR scarring experiments[4]. Beyond this, two recent studies exploit somatic mitochondrial DNA (mtDNA) mutations tracked by scRNA-seq and ATAC-seq sequencing as natural genetic barcodes to reconstruct cell lineages and infer clonal dynamics[5,6]. Moreover, integrated DNA-seq and scRNA-seq data was recently used to reconstruct clonal substructure for single-cell transcriptomes and characterize phenotypic and functional variations between genetically distinct subclonal populations[7]. Intrinsic T and B cell receptors can also serve as natural markers to trace clonality and lineage relationship for these lymphocytes. Combined with other omics measurements, for instance ATAC-seq[8], these approaches could be used to uncover clone-specific epigenetic patterns.

A different application of linking genotype to phenotype is to dissect genetics at the single cell level. Integrating scRNA-seq data with genotypes obtained from DNA sequencing facilitates the detection of functional genetic variants driving cell type-specific gene expression variation and the cellular contexts in which they affect gene expression. To identify the downstream effects of disease-associated genetic risk factors, SNPs can be linked with quantitative trait loci (QTLs), using statistical models to predict their effects on gene expression (eQTLs)[9], protein abundance (pQTL) or DNA methylation (meQTLs). We believe

that these type of analysis will bring deeper insights into the molecular and cellular mechanisms involved in disease risk and will inform therapeutic strategies.

Reconstructing gene regulatory networks from multi-modal single cell omics data

Gene expression is tightly regulated by a complex interplay of regulatory interactions with other genes and signaling molecules. Although scRNA-seq data in principle allows inference of gene regulatory networks[10], regulatory processes are often too complex to predict reliably from the transcriptome alone. Epigenetic modifications such as chromatin accessibility, DNA methylation and histone marks play an essential role in establishing and maintaining cellular phenotypes, but the mechanisms of this regulation are not well understood. Chromatin profiling techniques such as single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) and cleavage under targets and release using nuclease (CUT&RUN)[11] provide information of both TF binding and the regulatory potential of a genetic locus and can identify functional genomic elements that determine cellular state.

Joint analysis of transcriptomics and chromatin accessibility using the integration methods mentioned above[1,2] can reveal the existence of novel cell states and enable investigation of TF activity and enhancer elements that underlie those states. In addition, corrected expression matrix inferred from the joint projection can be used as input to additional methods such as pseudotime or network reconstruction. Moreover, another approach[12] using manifold alignment for inferring a shared pseudotime latent variable, can reveal connections among transcriptomics and epigenetic changes and the underlying regulatory mechanisms driving these changes. This can be applied to study many dynamic processes such as differentiation, development and tumorigenesis. Using scRNA-seq and paired bulk ATAC-seq data, another method applies a Dirichlet process mixture model that jointly learns clusters of cells and their underlying gene regulatory networks [13]. We anticipate gene regulatory network inference from integrated single-cell omics as a future direction for methods developments. These approaches will need to address the unique challenges of single-cell data, such as sparsity and increased dimensionality, as well as variability across technologies and studies.

Finally, combining single-cell omics with perturbation experiments at the level of TFs or enhancers provides the most direct way to infer causal regulatory programs. Several recent methods used pooled CRISPR/Cas9 genome editing to introduce a library of genetic perturbations into a population of cells, followed by single-cell transcriptomic or epigenomic profiling (reviewed in [14]). These approaches allow inferring regulatory relationships by correlating the phenotype or response of a cell to a perturbation of a certain regulator and can be a powerful strategy to infer gene regulatory networks and uncover molecular mechanisms that govern cell fate and function. The data produced by these methods can also be interpreted using a regression model, followed by clustering applied to the regression coefficient matrix to identify ‘modules’ of TFs with similar phenotypes, and modules of coregulated target genes.

Reconstruction of signaling networks using paired protein and transcriptomic quantification

Gene regulatory networks are directed by the proteins they encode. Proteomic methods that measure protein abundance and state (phosphorylation and other modifications) on a single cell level, provide quantification of ligands, receptors, downstream signaling molecules and lineage-specific TFs. This opens up opportunities for reconstructing inter- and intracellular signaling networks. Methods for inference of signaling pathway architecture are already being applied on single-cell mass cytometry (CyTOF) data [15,16].

Moreover, techniques for paired quantitation of protein epitopes and RNAs can correlate the state of cell signaling proteins with gene expression[17,18]. This allows for identifying cases where gene and protein expression levels are poorly correlated, suggesting post-transcriptional modifications. In addition, joint clustering on both gene expression and cell surface proteins from the same cells could achieve considerably higher resolution in defining cell states. A new generative model specifically designed for CITE-seq data[19] combines both RNA and protein measurements into one joint latent representation of cell state, while addressing the unique technical biases of each modality. The previously mentioned integration methods[1] can also be employed to identify shared cell states present across different modalities, as well as predict expression patterns of proteins that are not assayed. Going forwards, machine learning techniques such as nested effects models[20,21] coupled with perturbations (small molecules, CRISPR, RNA interference) that can computationally reconstruct signaling networks could be adapted to paired transcriptome and protein data.

The potential of a cell to orchestrate a response to a ubiquitously used signaling pathway is determined by other factors in addition to expression of receptors and TFs, such as pre-existing chromatin state or its spatial context and communication with neighbouring cells. We envision that adding additional multi-omic assays, like single-cell chromatin accessibility profiling and single-cell methods that preserve spatial position will substantially advance our understanding of signaling networks in diverse tissues and conditions in the coming years.

Mapping single cell omics into a spatial context in tissues

Rapid advances in spatial technologies, either imaging-based or sequencing-based[22], offer highly multiplex profiling of RNAs, while preserving spatial context in the tissue. Combined with computational approaches for mapping single cells to spatial reference maps[23,24,25,26], enables the construction of whole-transcriptome 3D atlases of at single-cell resolution. Moreover, using transfer learning approaches[1] or deep learning methods[3], spatial expression patterns of genes that are not assayed can also be imputed. This offers the possibility to infer spatial expression patterns of signaling ligands, receptors and downstream targets and complement cell-cell communication analysis[27] from standard scRNA-seq techniques. We

foresee that the diverse spatial technologies will in future be mapped to each other, with exciting prospects for new computational methods to achieve this.

Going forwards, this implies the possibility of mapping all features currently measured in disaggregated single cells into two and three dimensions. This includes cell lineage relationships in healthy and tumour tissue, cellular differentiation trajectories, T and B cell clonality, epigenetic states etc. In summary, these approaches will allow quantitative and phenotypic description of many facets of each cell in a tissue context. This includes intercellular communication across neighbouring cells, opening up the future possibility to study the complete interactome in a spatially resolved manner.

Outlook

Existing methods for integration of different modalities require at least partial correspondence between profiled features across omics; a limitation is their inability to incorporate different types of features, for instance gene expression and intergenic methylation. One way to overcome this is by constructing single cell maps based on a joint kernel that incorporates all multi-modal omic layers with the idea of creating a multi-space single similarity measure between them, which will allow for standard analyses on the integrated cell-cell distances, such as t-SNE or UMAP based visualization, clustering and trajectory inference[28].

Moving forward, combined multi-modal omics measurements of genomes, transcriptomes, epigenomes, proteomes and chromatin organization in the same single-cells will open up new avenues to link multiple aspects of cellular identity. Even simple correlation and regression models that leverage the scale and different readouts of single cell multi-modal omics data will reveal association between genotype and gene expression, transcriptional and DNA-methylation heterogeneity, and distal elements and target genes on the basis of covariance of accessibility and expression[29]. More sophisticated approaches, such as the recently developed MOFA+[30], provide a integrative framework for joint modeling of variation across both multiple modalities and multiple conditions and groups.

We envision that multi-view learning methods including kernel learning, network-based fusion methods, matrix factorization and multimodal deep learning will help disentangle the causal relationship between different -omics layers[29]. These models can be applied to uncover genotype-phenotype interactions and associate transcriptional states with epigenetic signatures. This will enable prediction of gene expression dynamics dependent on TF activity at specific sequences of regulatory DNA and identification of factors that drive cell fate. In addition, it will be possible to link regulatory networks with clustering and trajectory inference, which will in turn increase the power of causal inference. Follow-up validation of candidate enhancers with genetic deletions or CRISPRi would be crucial.

Finally, integration with spatial methods will enable identification of context specific functionally relevant relationships and how these shape cellular phenotypes. Novel computational methods offer an opportunity to exploit the full potential of single cell multi-

modal omics sequencing techniques and will deepen our understanding of cellular identity and responses in both health and disease.

Text box: Open questions

- Can we reconstruct gene regulatory networks more accurately by using integration of transcriptome and chromatin profiling? Will this improve methods for inferring regulatory networks associated with dynamic trajectories?
- Will methods for inference of cell-cell communication networks work well with imputed spatial data? Can the spatial distances between pairs of single cells be used as a prior to construct more accurate network models assuming that cells which are located closer together are more likely to signal to each other?
- Can methods for multi-omics data integration for bulk data be adapted to single cell data, considering single-cell data challenges such as sparsity and increased dimensionality?
- How can we apply these methods to elucidate mechanisms of complex diseases such as cancer?

Acknowledgments

We thank Emma Dann and Mika Sarkin Jain for careful and critical reading of the manuscript. We are grateful to Jana Eliasova for help with the illustrations.

Author contribution

M.E and S.A.T wrote the manuscript.

Competing interests

The authors declare no competing interests.

Figures

Figure 1. Schematic representation of different computational methods for single cell genomics across modalities.

References:

1. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
2. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).
3. Lopez, R. *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv [cs.LG]* (2019).
4. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166–179 (2018).
5. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325–1339.e22 (2019).
6. Xu, J. *et al.* Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *Elife* **8**, (2019).
7. McCarthy, D. J. *et al.* Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants. doi:10.1101/413047
8. Satpathy, A. T. *et al.* Transcript-indexed ATAC-seq for precision immune profiling. *Nat. Med.* **24**, 580–590 (2018).
9. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. doi:10.1101/630996
10. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
11. Hainer, S. J., Bošković, A., McCannell, K. N., Rando, O. J. & Fazzio, T. G. Profiling of

Pluripotency Factors in Single Cells and Early Embryos. *Cell* **177**, 1319–1329.e11 (2019).

12. Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 138 (2017).
13. Burdziak, C., Azizi, E., Prabhakaran, S. & Pe'er, D. A Nonparametric Multi-view Model for Estimating Cell Type-Specific Gene Regulatory Networks. *arXiv [stat.ML]* (2019).
14. Henriksson, J. CRISPR Screening in Single Cells. *Methods in Molecular Biology* 395–406 (2019). doi:10.1007/978-1-4939-9240-9_23
15. Krishnaswamy, S. *et al.* Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. *Science* **346**, 1250689 (2014).
16. Qin, X. *et al.* Single-Cell Signalling Analysis of Heterocellular Organoids. doi:10.1101/659896
17. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
18. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
19. Gayoso, A. *et al.* A Joint Model of RNA Expression and Surface Protein Abundance in Single Cells. doi:10.1101/791947
20. Markowetz, F., Kostka, D., Troyanskaya, O. G. & Spang, R. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* **23**, i305–12 (2007).
21. Pirkl, M. & Beerenwinkel, N. Single cell network analysis with a mixture of Nested Effects Models. *Bioinformatics* **34**, i964–i971 (2018).
22. Mayr, U., Serra, D. & Liberali, P. Exploring single cells in space and time during tissue development, homeostasis and regeneration. *Development* **146**, (2019).

23. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

24. Karaiskos, N. *et al.* The Drosophila Embryo at Single Cell Transcriptome Resolution. doi:10.1101/117382

25. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

26. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).

27. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. doi:10.1101/680926

28. Colomé-Tatché, M. & Theis, F. J. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology* **7**, 54–59 (2018).

29. Packer, J. & Trapnell, C. Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation. *Trends Genet.* **34**, 653–665 (2018).

30. Argelaguet, R. *et al.* MOFA : a probabilistic framework for comprehensive integration of structured single-cell data. doi:10.1101/837104

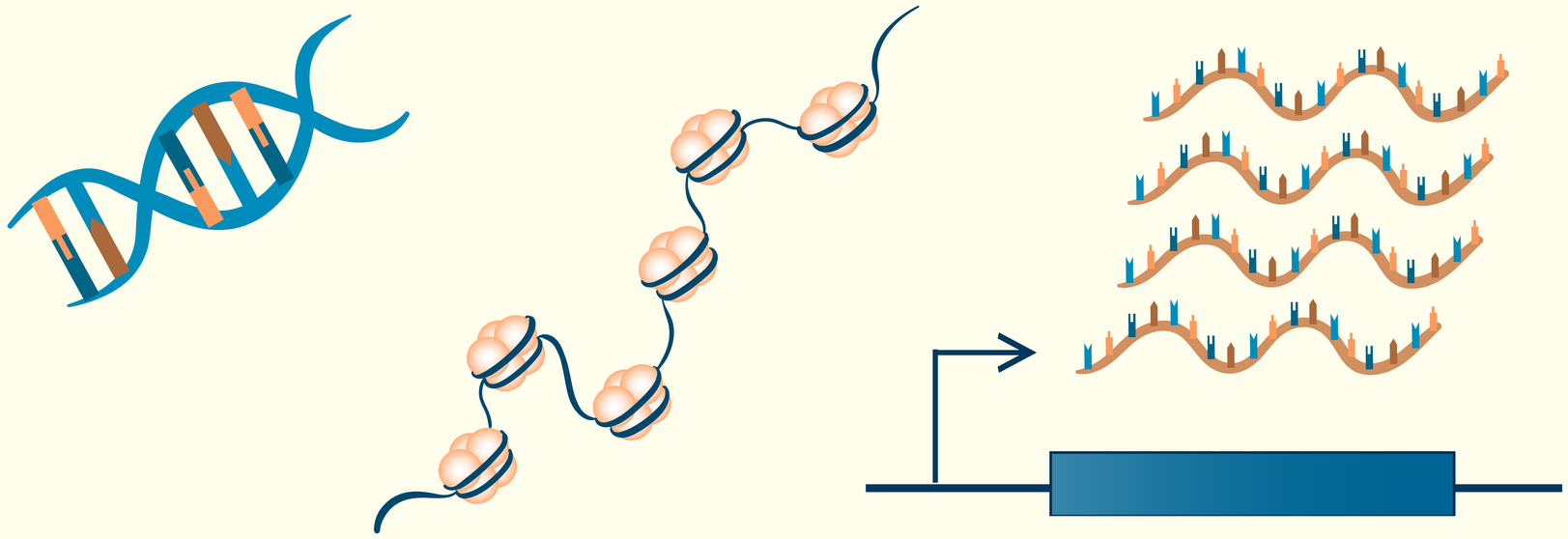
Genome

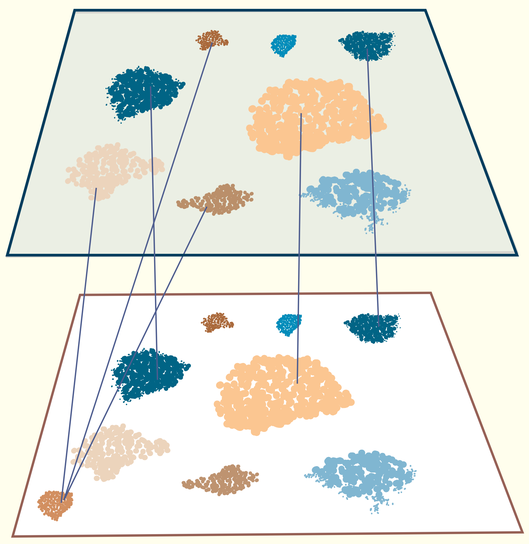
Chromatin accessibility

Gene expression

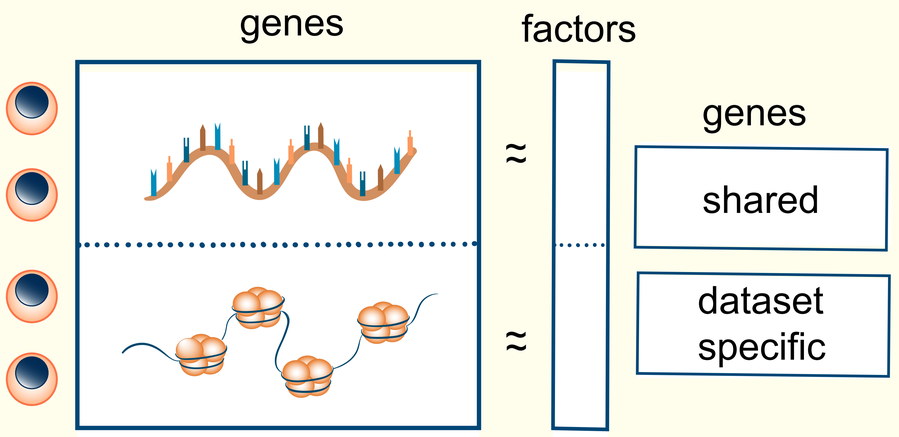
Protein abundance

Spatial transcriptomics

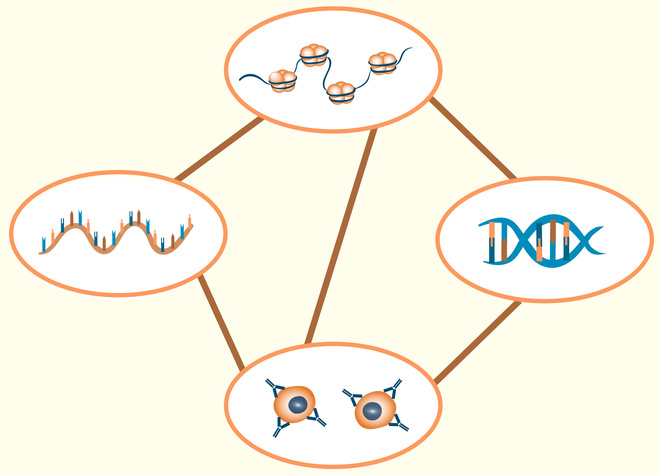




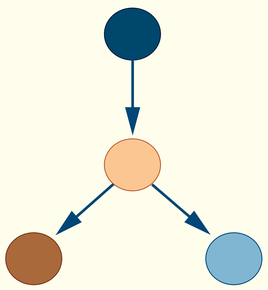
Transfer learning



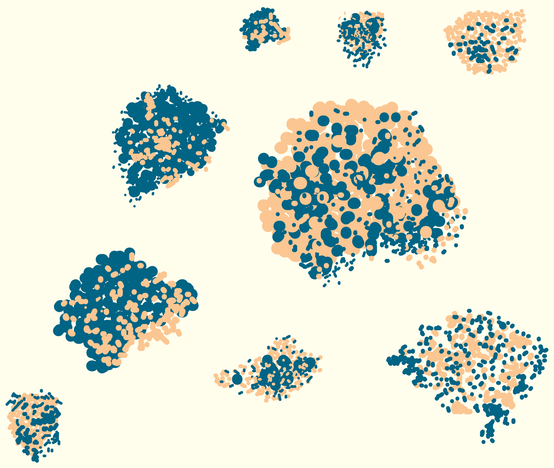
Matrix factorisation



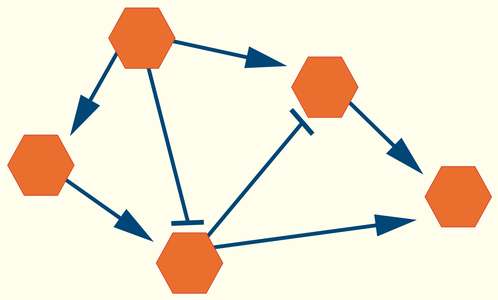
Correlation/regression models



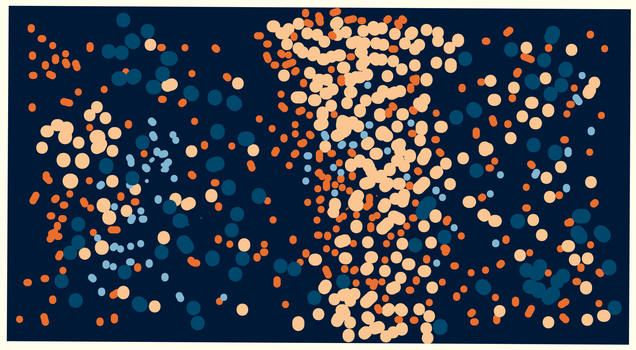
Cell lineage



Joint clustering



Gene regulatory network



Spatial context