Learning from Data in Single-Cell Transcriptomics

Sandrine Dudoit
Department of Statistics, Division of Biostatistics, and Center for Computational Biology
University of California, Berkeley
www.stat.berkeley.edu/~sandrine

Bioinformatics Training and Education Program, NIH

April 16, 2026

UC Berkeley College of Computing, Data Science, and Society

Acknowledgments

- Group members (current and former)
  - Philippe Boileau (now at McGill University). [scPCA]
  - Florica Constantine. [Spatial transcriptomics]
  - Fanny Perraudeau (now at Bain & Company). [ZINB-WaVE]
  - Davide Risso (now at University of Padova). [EDASeq, RUVSeq, scone, ZINB-WaVE, RSEC, Dune, Slingshot]
  - Hector Roux de Bézieux (now at Owkin). [Dune, tradeSeq, condiments]
  - Kelly Street (now at USC). [Dune, Slingshot, tradeSeq, condiments]
  - Koen Van den Berge (now at Janssen R&D). [ZINB-WaVE, Dune, tradeSeq, condiments, transfactor]
- John Ngai and his lab, Department of Molecular and Cell Biology, UC Berkeley, and NIH BRAIN Initiative. [Biology: Mouse OE]

Acknowledgments

- Zoltan Laszik and his lab, Department of Pathology, UCSF. [Biology: Kidney transplant]
- Peter Bickel, Department of Statistics and Center for Computational Biology, UC Berkeley. [transfactor]
- Lieven Clement, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium. [ZINB-WaVE, tradeSeq]
- Elizabeth Purdom, Department of Statistics and Center for Computational Biology, UC Berkeley. [scone, RSEC, Dune, Slingshot]
- Jean-Philippe Vert, Owkin, Mines ParisTech, and Institut Curie, Paris, France. [ZINB-WaVE]
- Nir Yosef, Weizmann Institute of Science, Israel. [scone, Slingshot]

Outline

Olfactory Stem Cells and Neural Regeneration

Stem cell differentiation in the mouse olfactory epithelium.
(Fletcher et al., 2017; Gadye et al., 2017; Van den Berge et al., 2026)

- Goal: Elucidate the molecular and cellular mechanisms underlying stem cell-mediated development and regeneration in the olfactory epithelium's (OE) neurogenic stem cell niche.
- Potential applications: Prevention and treatment of neural tissue damage and degeneration, e.g., Alzheimer's disease.
- Focus on the differentiation of horizontal basal cells (HBC), a type of adult tissue stem cells.
- The p63 protein (tumor protein p63, TP63) promotes self-renewal of HBCs by blocking differentiation. When p63 is down-regulated, differentiation proceeds at the expense of self-renewal. Thus, p63 can be viewed as a "molecular switch" that decides between the alternate stem cell fates of self-renewal and differentiation.

Olfactory Stem Cells and Neural Regeneration

- OE p63 dataset. [Fluidigm C1, $\sim$ 700 cells; Fletcher et al. (2017)]
  Investigate the differentiation of HBCs, using single-cell transcriptome sequencing (scRNA-Seq) to measure genome-wide expression levels at the resolution of single cells in wild-type (WT) and p63 knock-out (KO) mice, at five timepoints following tamoxifen treatment.
- OE injury response dataset. [10X Genomics Chromium v2, $\sim$ 25K cells; Van den Berge et al. (2026)]
  Investigate the transcriptional response to injury, using scRNA-Seq to measure gene expression in the OE of adult mice treated with methimazole, at 24h, 48h, 96h, 7d, and 14d after injury.

Olfactory Stem Cells and Neural Regeneration

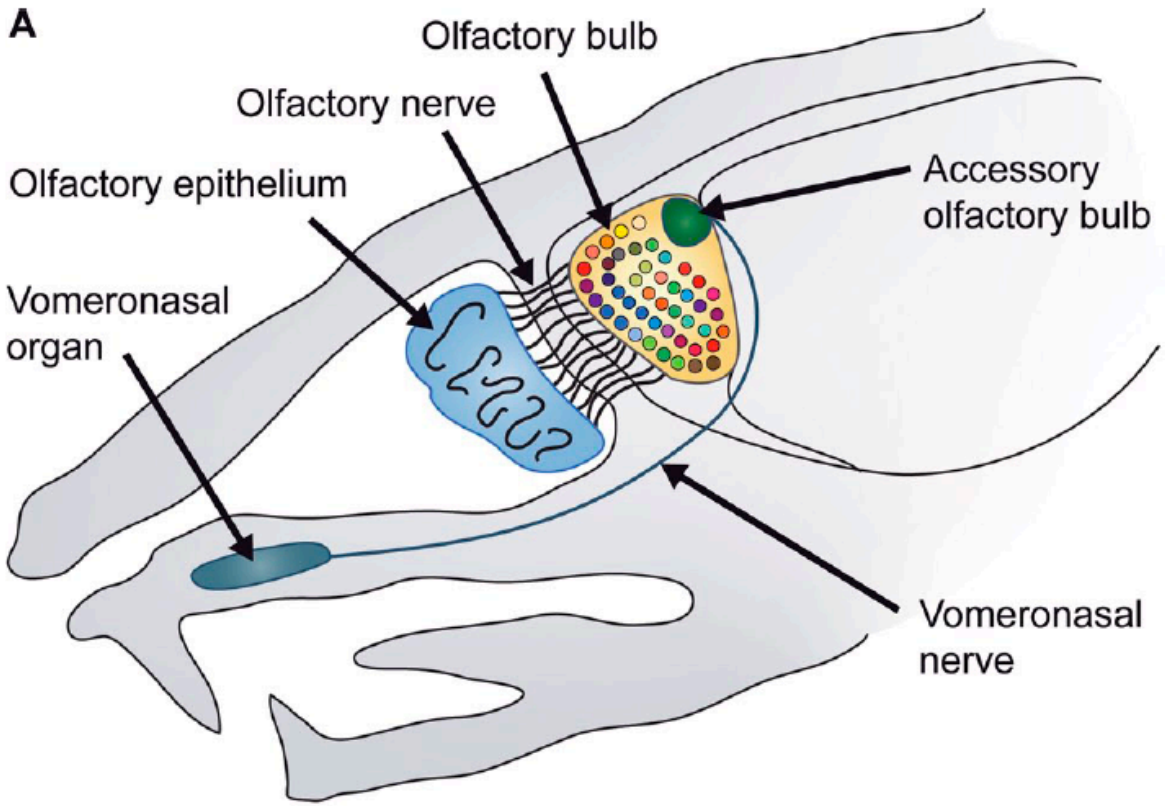


Figure 1: *Mouse olfactory epithelium.*

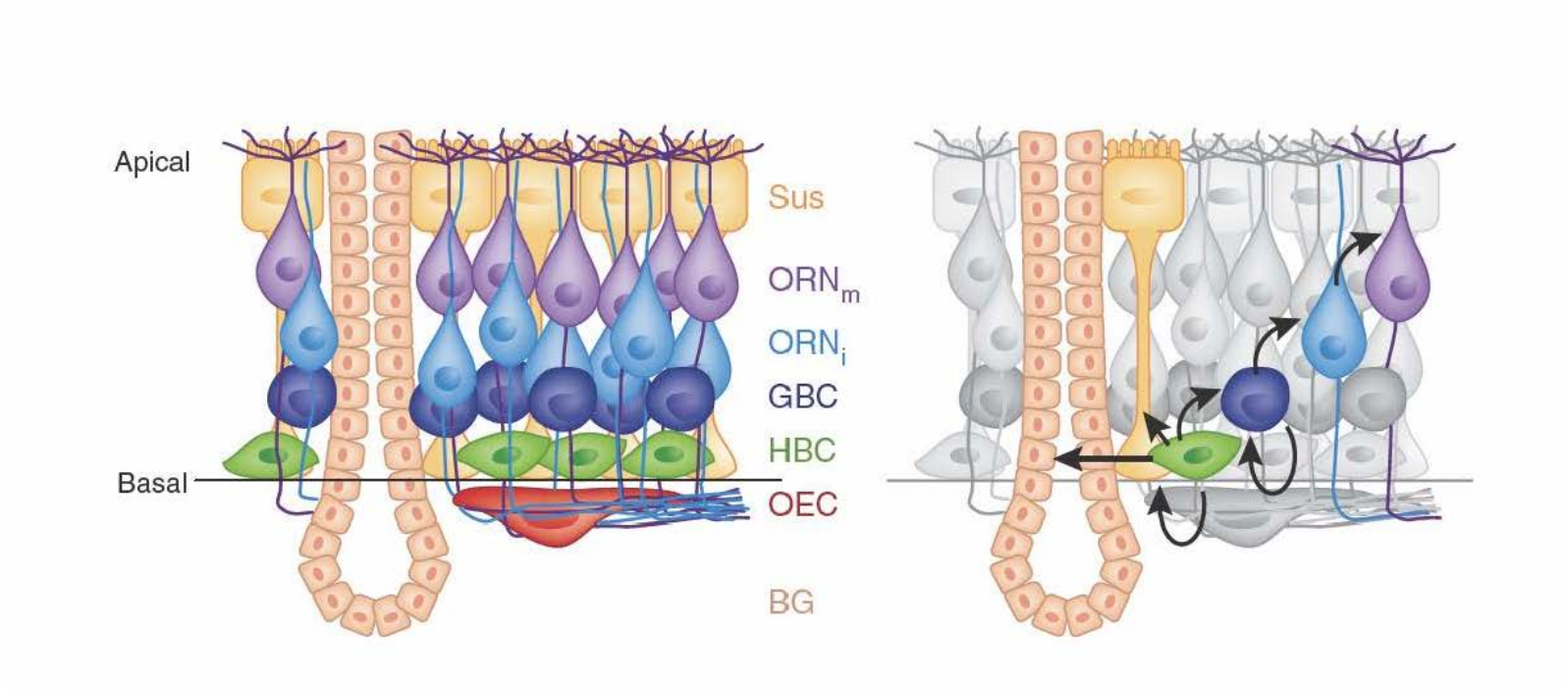


Figure 2: *Olfactory epithelium cell types.* Sus: sustentacular cell, ORN: olfactory receptor neuron, GBC: globose basal cell, HBC: horizontal basal cell, OEC: olfactory ensheathing cell, BG: Bowman gland.

Olfactory Stem Cells and Neural Regeneration

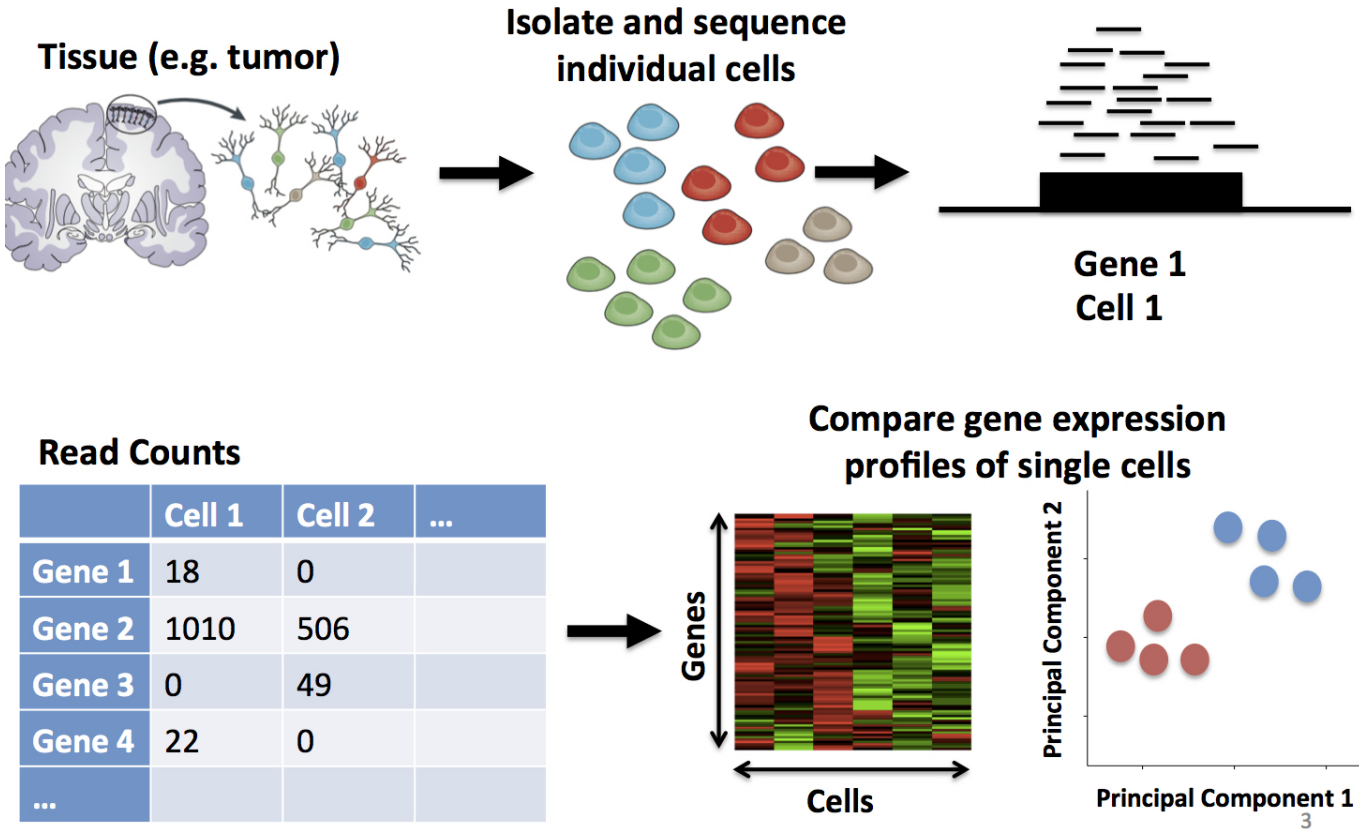


Figure 3: *Singe-cell RNA-Seq.*

- Frame and translate the domain question into a data-enabled statistical question.
  - Essential and far-from-trivial step, yet often overlooked.
  - *"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."* (John W. Tukey)
- Exploratory data analysis and quality assessment/control: EDASeq (Perraudeau et al., 2017).
  Summarize and visualize the data to identify the main features as well as problems with these data.
  - Look at data to avoid "garbage in, garbage out" (GIGO).
- Normalization: RUVSeq, scone (Cole et al., 2019; Risso et al., 2011, 2014a,b; Vallejos et al., 2017).
  Adjust read counts to ensure that observed differences in expression measures between genes or samples reflect biological effects of interest and not unwanted technical effects.

  - Normalization procedures: Global scaling, quantile matching, regression on known factors of unwanted variation (supervised), regression on unknown factors of unwanted variation (unsupervised, RUV).
  - Normalization performance assessment and selection.
- Dimensionality reduction: scPCA (Boileau et al., 2020a,b).
  Sparse contrastive principal component analysis (scPCA) to remove unwanted variation and extract sparse, stable, interpretable, and relevant biological signal.
- Expression quantitation: zinbwave (Risso et al., 2018a; Van den Berge et al., 2018).
  Zero-inflated negative binomial-based wanted variation extraction method (ZINB-WaVE):
  - account for zero inflation and over-dispersion;
  - accommodate experimental design (e.g., batch, nesting);

  - adjust for known and unknown factors of unwanted variation (normalization);
  - quantify biological effects of interest;
  - perform dimensionality reduction;
  - provide weights to be used in standard bulk RNA-Seq differential expression (DE) methods (e.g., edgeR, DESeq2, and limma).

- Resampling-based sequential ensemble clustering (RSEC): clusterExperiment (Risso et al., 2018b).
  General and flexible framework for applying and comparing a variety of different clustering algorithms and associated tuning parameters and aggregating multiple candidate clusterings into a stable consensus clustering.
- Cluster merging procedure to navigate the trade-off between cluster resolution and replicability across datasets: Dune (Roux de Bézieux et al., 2020, 2024b).

- Inference of cell lineages and pseudotimes: slingshot (Street et al., 2018).
  - Infer the global lineage structure (i.e., the number of lineages and where they branch) using a cluster-based minimum spanning tree (MST).
  - Infer cell pseudotimes along each lineage using simultaneous principal curves.
  - Can identify any number of lineages.
  - May incorporate subject-matter knowledge to supervise parts of the inference process (e.g., known terminal states).
- Trajectory-based differential expression: tradeSeq (Van den Berge et al., 2020).
  Identify differentially expressed (DE) genes, both within- and between-lineages.

  - Rely on a negative binomial (NB) generalized additive model (GAM) to exploit the continuous resolution provided by the pseudotimes from trajectory inference (vs. DE between discrete cell clusters).
  - Identify different types of DE patterns based on contrasts for the NB-GAM coefficients.

- Trajectory inference across multiple conditions: condiments (Roux de Bézieux et al., 2024a).
  Identification of differences between conditions (e.g., wild-type/knock-out) at the trajectory (differential topology), cell population (differential progression and fate selection), and gene (differential expression) levels.
- Inference of transcription factor activity: transfactor (Van den Berge et al., 2025).

  - Deconvolve transcription factor-specific gene expression from overall gene expression by leveraging gene regulatory network (GRN).
  - Investigate regulatory differences in TF activity within and between lineages in a trajectory.

- Software. The above methods are implemented in open-source R (`www.r-project.org`) software packages released through the Bioconductor Project (`www.bioconductor.org`).

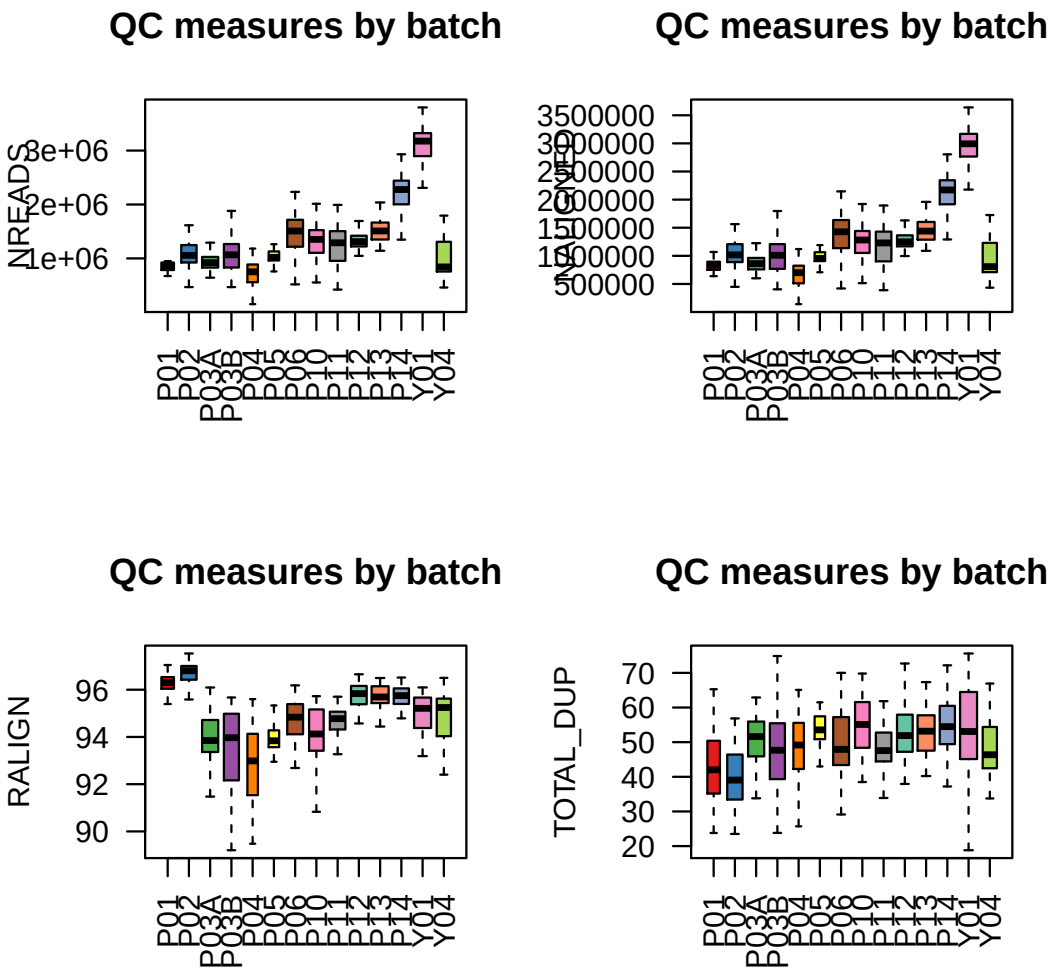


Figure 4: *Sample-level QC: OE p63 dataset.* Boxplots of QC measures, by batch.

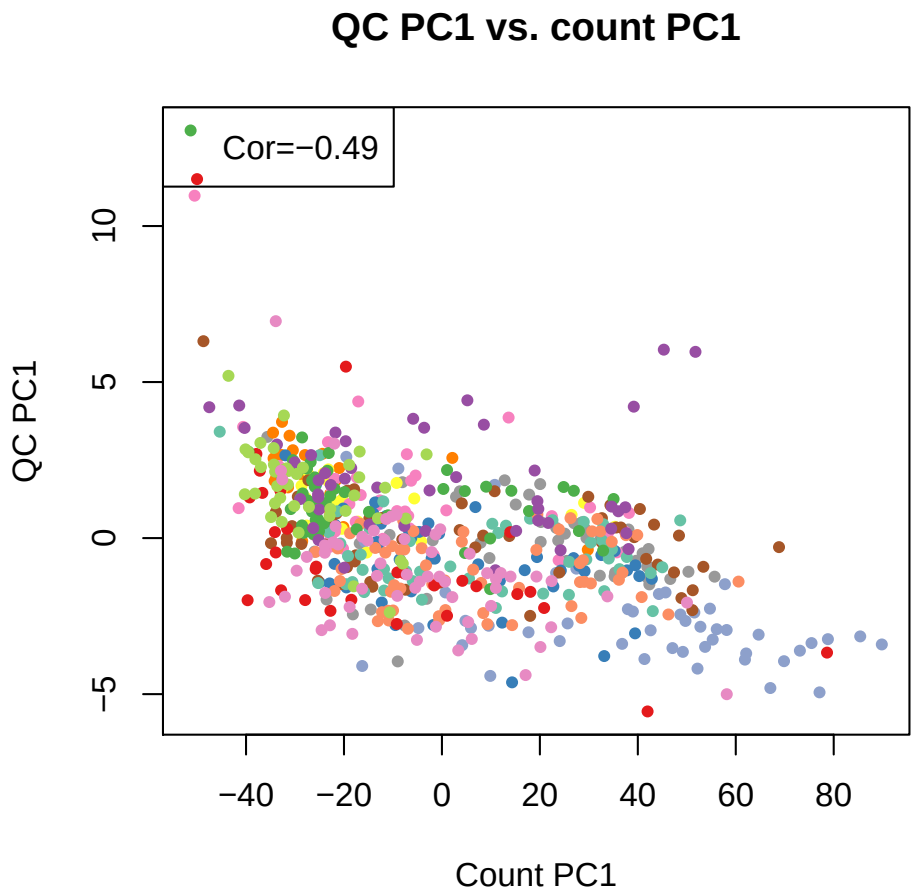
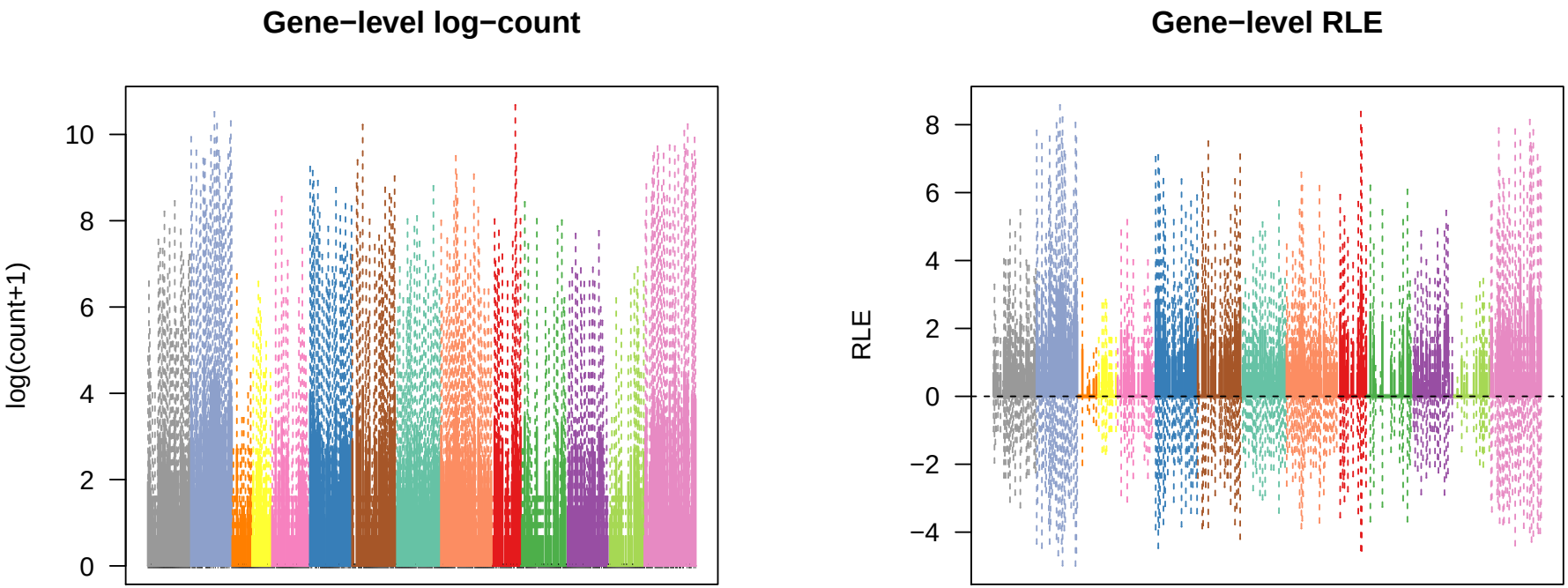


Figure 5: *Sample-level QC: OE p63 dataset.* Association of counts and sample-level QC measures. QC PC1 vs. count PC1, color-coded by batch.

EDA: Gene-Level Counts



(a) Log-count

(b) RLE

Figure 6: *Gene-level counts: OE p63 dataset.* Boxplots of gene-level log-count and relative log expression (RLE = log-ratio of read count to median read count across cells), color-coded by batch.

Normalization: Motivation

- The goal of normalization is to adjust read counts for gene-level (e.g., length, GC-content) and sample-level (e.g., sequencing depth, batch, QC) unwanted technical effects, in order to allow meaningful comparison of expression measures between genes or samples.
- Normalization is essential before any clustering or differential expression analysis, to ensure that observed differences in expression measures between genes or samples reflect biological effects of interest and not technical artifacts.
- Normalization is even more important for single-cell RNA-Seq than bulk RNA-Seq due to increased technical noise and zero inflation.

Normalization: Motivation

- Does normalization matter? Yes!
  The choice of normalization method can have a greater impact on the results than the choice of downstream method for inferring differential expression (Bullard et al., 2010).
- Which method is best? Not obvious, depends on dataset.
  Need a data-driven approach and controls for selecting a suitable normalization procedure.
  $\longrightarrow$ scone.

Cole et al. (2019). General framework for the normalization of scRNA-Seq (and other) data, scone.

- Implementation of a range of normalization methods.
  - Global-scaling, e.g., DESeq, TMM, upper-quartile (UQ).
  - Full-quantile (FQ).
  - Regression on known factors of unwanted variation (supervised): E.g. QC PC, batch.
  - Regression on unknown factors of unwanted variation (unsupervised): Remove unwanted variation (RUV) (Risso et al., 2014a,b).
- Normalization performance metrics.
  - Clustering of samples according to factors of wanted and unwanted variation.
  - Association of expression measures with factors of wanted and unwanted variation.
  - Between-sample distribution of expression measures.

Normalization: scone

- Numerical and graphical summaries of normalized read counts and performance metrics.
- Shiny app.
- We've used the scone framework for the normalization other types of -omic data, including adductomics and metabolomics data.
- Bioconductor R package scone: `www.bioconductor.org/packages/release/bioc/html/scone.h`

Application to OE p63 dataset.

- Apply and evaluate 172 normalization procedures using main `scone` function.
  - `scaling_method`: None, DESeq, TMM, FQ.
  - `uv_factors`: None; RUVg $k = 1, \cdots, 5$; QC PC $k = 1, \cdots, 5$.
  - `adjust_biology`: Yes/no.
  - `adjust_batch`: Yes/no.
- Among best-performing methods:
  `none,fq,qc_k=4,bio,no_batch`,
  `none,fq,qc_k=2,no_bio,no_batch`.

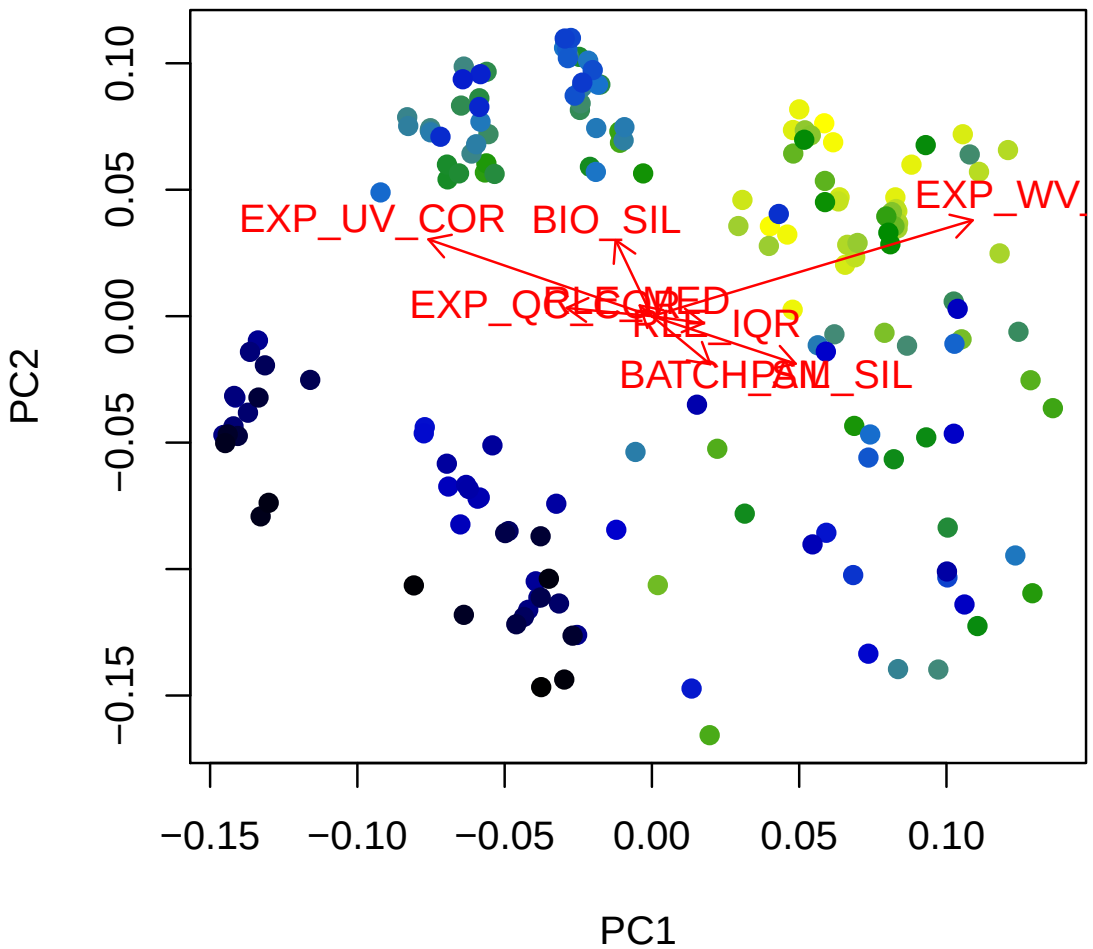
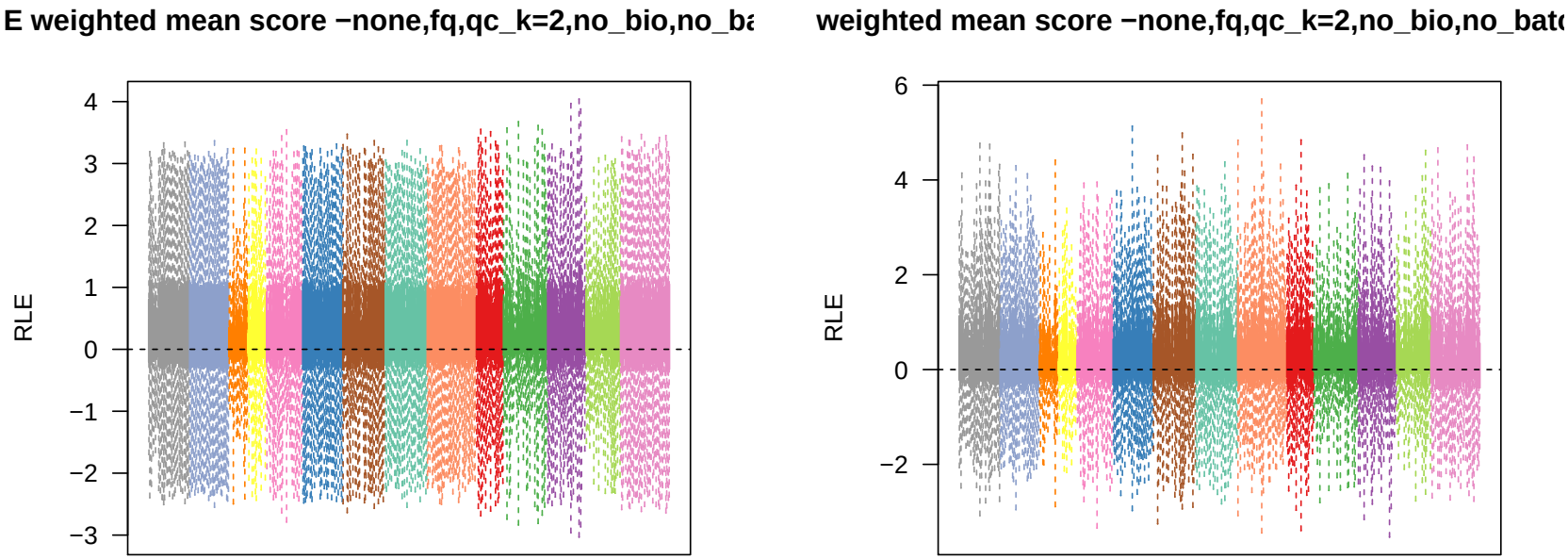


Figure 7: *scone: OE p63 dataset.* Biplot of performance scores, colored by mean score (yellow high/good, blue low/bad).

Normalization: scone



(a) All genes

(b) Housekeeping genes

Figure 8: *scone: OE p63 dataset.* Gene-level relative log expression (RLE = log-ratio of read count to median read count across samples), color-coded by batch, `none,fq,qc_k=2,no_bio,no_batch`.

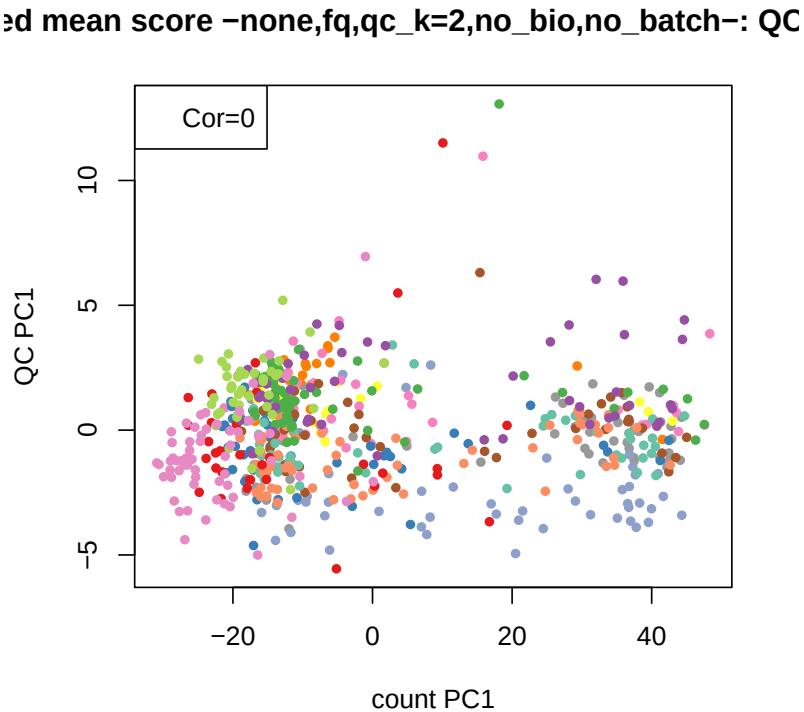
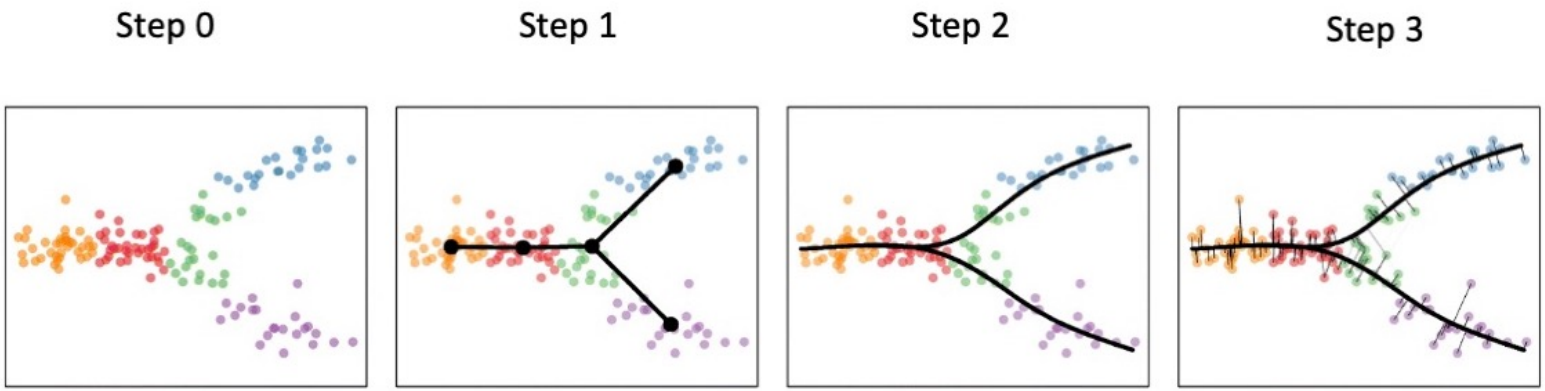


Figure 9: *scone: OE p63 dataset.* Association of counts and sample-level QC measures. QC PC1 vs. count PC1, color-coded by batch (`none,fq,qc_k=2,no_bio,no_batch`).

Inference of Cell Lineages and Pseudotimes: Slingshot

- Slingshot provides a flexible and robust framework for inferring cell lineages and pseudotimes (Street et al., 2018).
- The method comprises two main steps:
  1. the inference of the global lineage structure (i.e., the number of lineages and where they branch) using a cluster-based minimum spanning tree (MST);
  2. the inference of cell pseudotimes along each lineage using a novel method of simultaneous principal curves.
- Slingshot allows the identification of any number of lineages, with the option of incorporating subject-matter knowledge to supervise parts of the inference process (e.g., known terminal states).
- Bioconductor R package slingshot: `www.bioconductor.org/packages/release/bioc/html/slingsh`

Inference of Cell Lineages and Pseudotimes: Slingshot



0. Clustering and dimensionality reduction
1. Infer global trajectory structure using MST on clusters
2. Fit simultaneous principal curves to cells
3. Infer pseudotimes by orthogonal projection onto the curves

Figure 10: *slingshot*: *Main steps.*

Inference of Cell Lineages and Pseudotimes: Slingshot

Application to OE p63 dataset.

- Cell clusters. We use the RSEC clustering to define states in the differentiation of HBCs to neuronal and sustentacular cells.
  - horizontal basal cells (HBC),
  - globose basal cells (GBC),
  - microvillous cells (MV),
  - immediate neuronal precursors (INP),
  - immature and mature olfactory sensory neurons (iOSN, mOSN),
  - immature and mature sustentacular cells (iSus, mSus).

Inference of Cell Lineages and Pseudotimes: Slingshot

- Leaf-node supervision. Known terminal clusters were provided to Slingshot: Mature sustentacular cells (mSus), microvillous cells (MV), and mature olfactory sensory neurons (mOSN) (only the first had an effect).
  Without leaf-node supervision, we draw the (known) false conclusion that sustentacular cells may develop into GBC.
- Slingshot identifies three lineages:
  HBC–mSus,
  HBC–GBC–MV,
  HBC–GBC–mOSN.

Inference of Cell Lineages and Pseudotimes: Slingshot

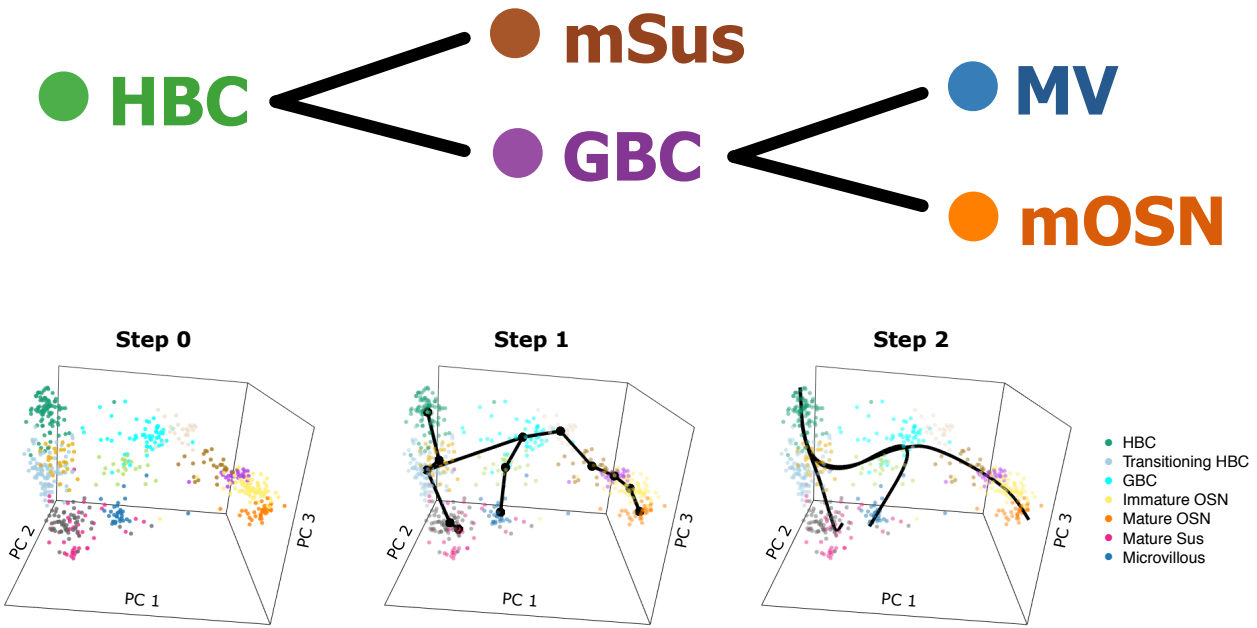


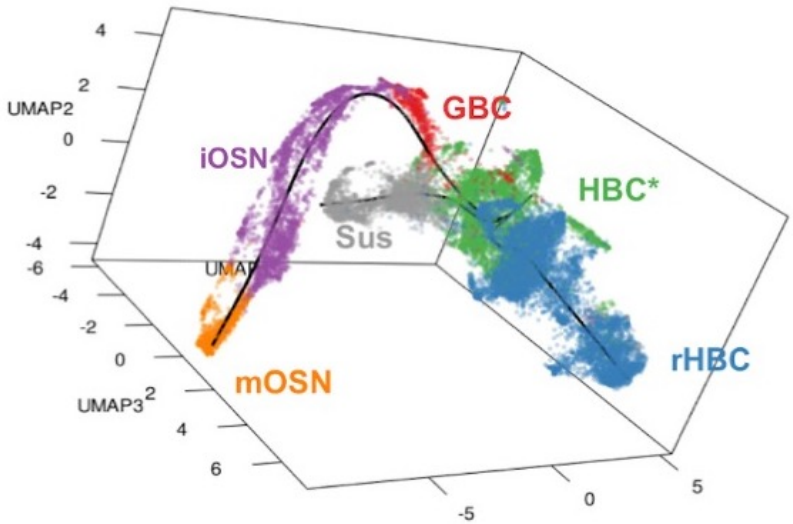
Figure 11: *slingshot: OE p63 dataset.* Step 1: MST on cell clusters. Step 2: Simultaneous principal curves.
Slingshot identifies three lineages: HBC–mSus, HBC–GBC–MV, HBC–GBC–mOSN.

Inference of Cell Lineages and Pseudotimes: Slingshot

Application to OE injury response dataset.

- Cell clusters. We use the clustering from Brann et al. (2020) to define states in the differentiation of HBCs to neuronal and sustentacular cells.
- Leaf-node supervision. Known terminal clusters were provided to Slingshot: Sus, mOSN, and rHBC.
- Upon injury of the OE, HBCs are activated in order to rebuild the tissue and Slingshot identifies three lineages:
  HBC*–Sus,
  HBC*–GBC–iOSN–mOSN,
  HBC*–rHBC.

Inference of Cell Lineages and Pseudotimes: Slingshot



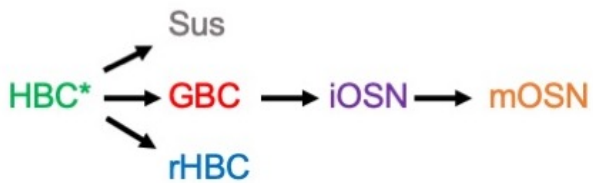


Figure 12: *slingshot: OE injury response dataset.*

Trajectory-Based Differential Expression: tradeSeq

- Downstream of trajectory inference, it is of interest to identify genes that are associated with the lineages, in order to gain insight into the biological processes underlying differentiation.
- The typical approach is to assess differential expression between (discrete) cell clusters, which fails to exploit the continuous resolution of the trajectory.
- In Van den Berge et al. (2020), we introduce tradeSeq, a negative binomial generalized additive model (NB-GAM) framework, that allows flexible inference of
  - within-lineage differential expression, by detecting associations between gene expression and pseudotime over an entire lineage or between points/regions within the lineage;
  - between-lineage differential expression, by comparing gene expression between lineages over the entire lineages or at specific points/regions.

Trajectory-Based Differential Expression: tradeSeq

- Different types of DE patterns are identified by based on linear combinations of the NB-GAM coefficients.
- The NB-GAM can also be used to cluster genes according to their expression patterns.
- Bioconductor R package tradeSeq: `www.bioconductor.org/packages/release/bioc/html/tradeSe` (GAM fit using Simon Wood's R package mgcv).

Trajectory-Based Differential Expression: tradeSeq

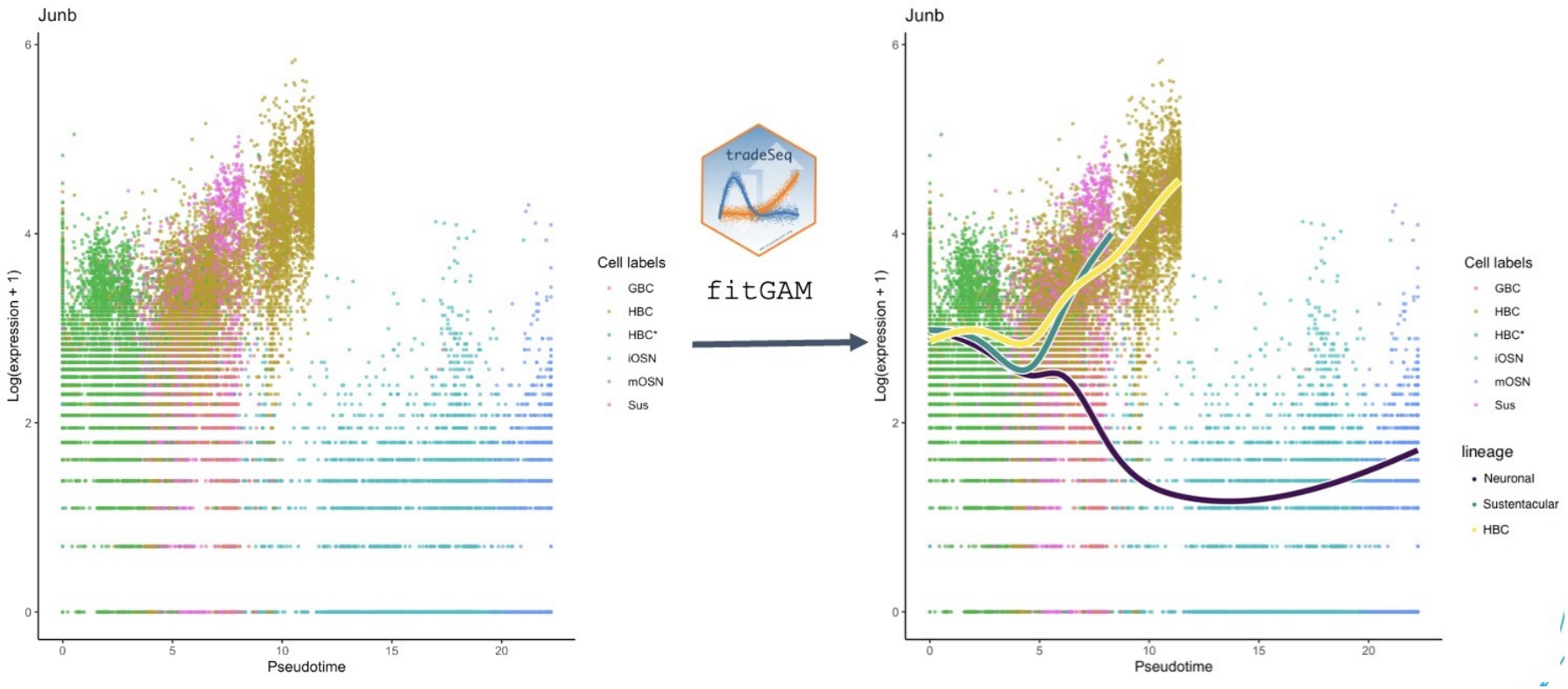


Figure 13: *tradeSeq: OE injury response dataset.* Use NB-GAM to relate gene expression to pseudotime for each lineage and to detect DE both within and between lineages.

Trajectory-Based Differential Expression: tradeSeq

Gene-wise negative binomial generalized additive model (NB-GAM)

$$\begin{cases} Y_{gi} & \sim & NB(\mu_{gi}, \phi_g) \\ \log(\mu_{gi}) & = & \eta_{gi} \\ \eta_{gi} & = & \sum_{l=1}^{L} s_{gl}(T_{li}) Z_{li} + \mathbf{U}_i \alpha_g + \log(N_i) \end{cases}$$

g : gene
i : cell
l : lineage

Smooth function of pseudotime for each lineage

Assigns cells to lineages

Covariates (fixed effects)

Normalization offsets

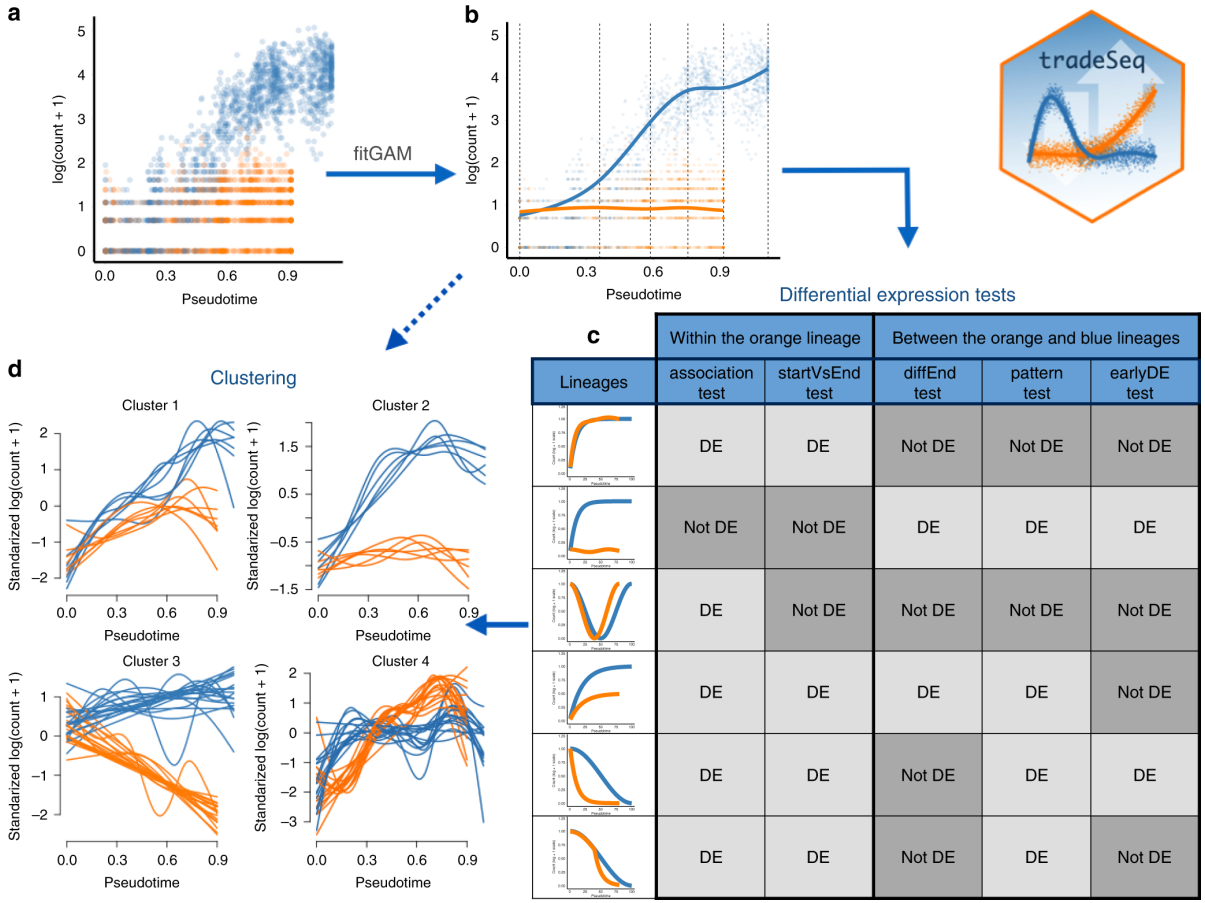
Lineage-specific smoothing splines, with K cubic basis functions

$$s_{gl}(t) = \sum_{k=1}^{K} b_k(t) \beta_{glk}$$

Identify DE patterns based on contrasts of β's.

Figure 14: *tradeSeq: NB-GAM.* Gene-wise NB-GAM relates gene expression measures Y to pseudotimes T; different types of DE patterns are identified based on contrast for coefficients β.

Trajectory-Based Differential Expression: tradeSeq



Lineages	Within the orange lineage		Between the orange and blue lineages		
	association test	startVsEnd test	diffEnd test	pattern test	earlyDE test
	DE	DE	Not DE	Not DE	Not DE
	Not DE	Not DE	DE	DE	DE
	DE	Not DE	Not DE	Not DE	Not DE
	DE	DE	DE	DE	Not DE
	DE	DE	Not DE	DE	DE
	DE	DE	Not DE	DE	Not DE

Figure 15: *tradeSeq: Overview of functionality.*

Trajectory-Based Differential Expression: tradeSeq

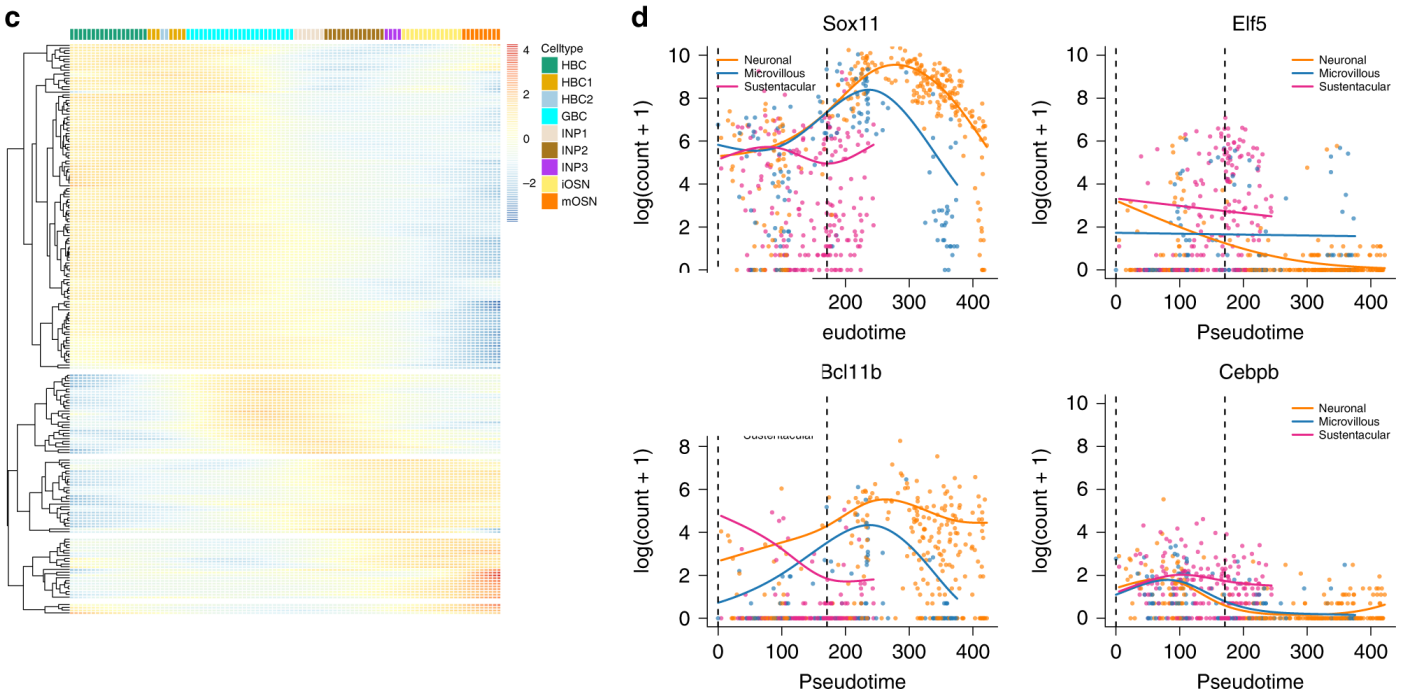


Figure 16: *tradeSeq: OE p63 dataset.* Left: Heatmap of fitted values for top 200 DE genes within neuronal lineage (associationTest). Right: Four DE TFs between lineages (earlyDETest).

Trajectory-Based Differential Expression: tradeSeq

OE p63 dataset.

- A heatmap of the fitted values for the top 200 DE genes for the neuronal lineage reveals five gene clusters, each with a different region of activity during the developmental process (associationTest).
- Four of the transcription factors (TF) that are DE between lineages are involved in epithelial cell differentiation (earlyDETest).
- tradeSeq uncovers transcriptional programs that are active in each of the three lineages, identifying both known and novel marker genes.
- Sustentacular cells are produced via direct conversion of HCB (without cell division). By contrast, microvillous and neuronal cells are produced via an intermediate, proliferative state (GBC).

Trajectory-Based Differential Expression: tradeSeq

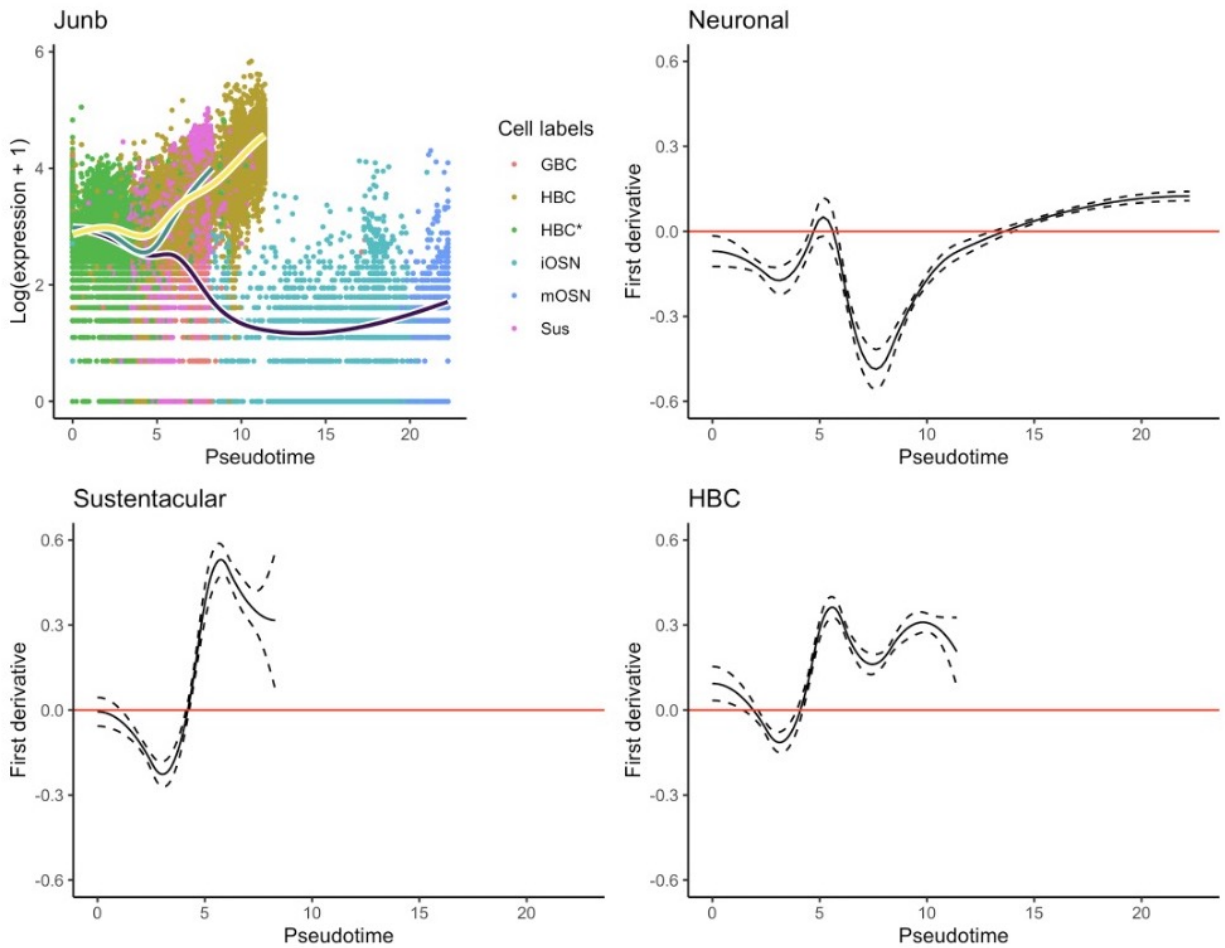


Figure 17: *tradeSeq: OE injury response dataset.* First derivatives of NB-GAM fits for each lineage. Purple: Neuronal, Green: Sus, Yellow: rHBC.

Trajectory-Based Differential Expression: tradeSeq

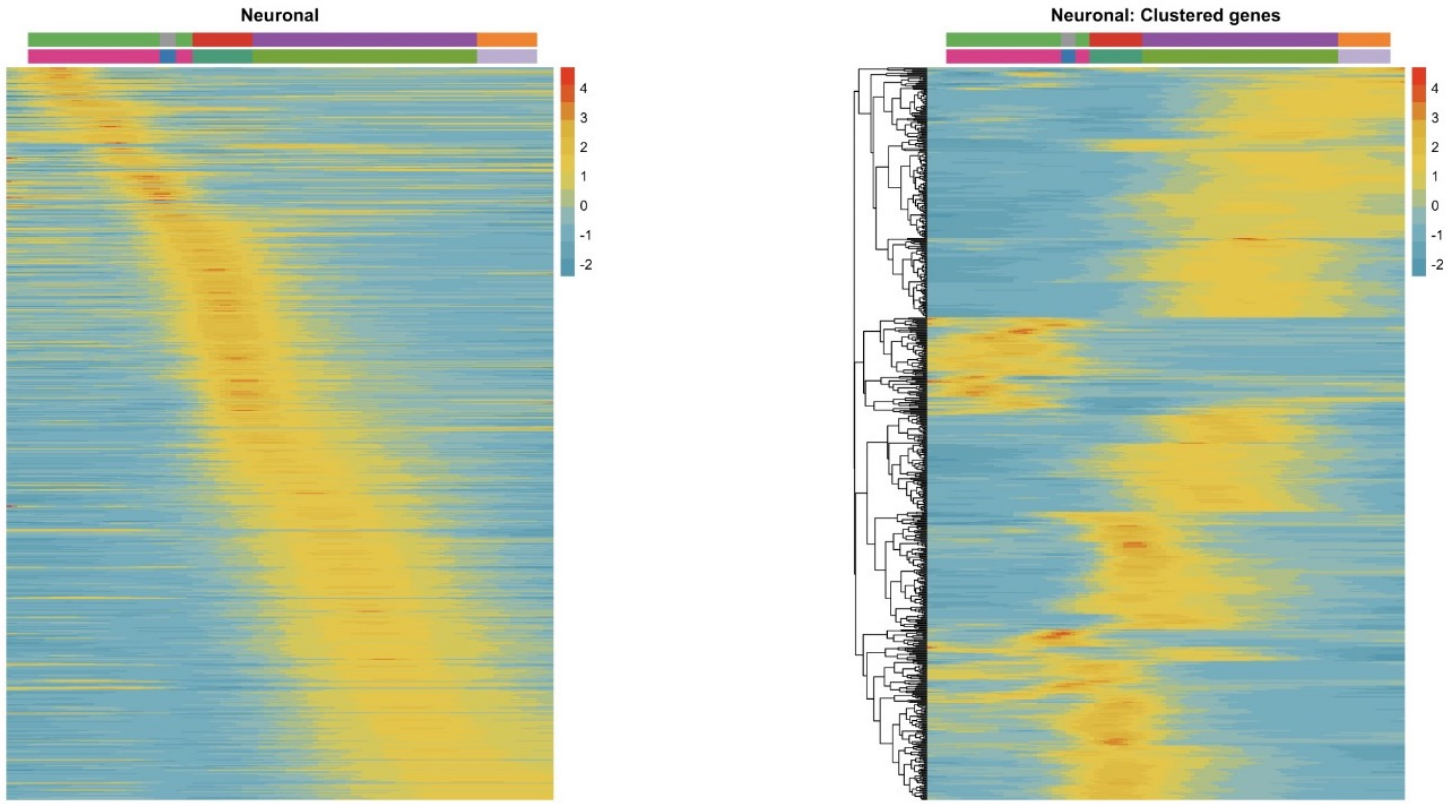


Figure 18: *tradeSeq: OE injury response dataset, TF activity cascade for neuronal lineage.* Heatmap of tradeSeq fitted values, cells binned by pseudotime and most abundant cell type indicated in color bar. Left: TFs ordered according to pseudotime of most significant peak. Right: Hierarchical clustering of TFs.

Trajectory-Based Differential Expression: tradeSeq

OE injury response dataset.

- Using the first derivatives of the NB-GAM, we identified transcription factor activity cascades in each lineage, highlighting the moment at which TFs are most active.
- This allows a grouping of TFs based on their sequential activation profile and the functional annotation of the biological processes that are active along development.
- Gene set enrichment analysis of the TF clusters indicates the processes activated by the TFs at various stages of differentiation.
  - Neuronal lineage: TFs involved in stress response at the early HBC* stage, then cell cycle regulation and neuron differentiation during the GBC and iOSN stages, and finally processes such as dendrite development, cell projection, and calcium-mediated signaling at the iOSN and mOSN stages.

  - Sustentacular lineage: TFs involved in cell growth and the regulation of proliferation early in lineage, then regulation of differentiation later on.
  - rHBC lineage: TFs involved in initial stress response, followed by neural precursor cell proliferation and circulatory system development.

- Overall, Slingshot and subsequent DE analysis with tradeSeq revealed that olfactory stem cells use divergent strategies to generate the major cell types of the epithelium. There are numerous step-like transitions in the neuronal lineage, but fewer gradual changes in the sustentacular lineage.

Inference of Transcription Factor Activity: transfactor

- We would like to examine transcription factor (TF) activity to gain insight into regulatory differences underlying differential gene expression along a trajectory.
- Why not use transcript abundance for a TF gene to measure the TF's activity?
  - While TF protein abundance is typically high in single cells, the mRNA abundance of the corresponding TF gene is often low.
  - TFs that are highly active, i.e., producing many mRNA molecules from their downstream target genes, may have genes with relatively low mRNA abundances.
- Instead, we define transcription factor activity in terms of the number of mRNA molecules produced across all the genes that a TF is regulating.
- Gene regulation by transcription factors may be summarized by a gene regulatory network (GRN), with
  - nodes representing genes and TFs,

Inference of Transcription Factor Activity: transfactor

  - edges representing regulatory interactions between genes and TFs (induction or repression).

- We have developed an approach, transfactor, to leverage GRNs to infer transcription factor activity.

Inference of Transcription Factor Activity: transfactor

- Use a hierarchical Poisson model for the number of transcripts produced by each TF for a given gene, where prior information on the GRN may be incorporated via a Dirichlet distribution.
- Use the EM algorithm to fit the model and deconvolve TF-specific gene expression from overall gene expression for each gene.
- Assess differences in TF activity within and between lineages in a trajectory using tradeSeq.

Inference of Transcription Factor Activity: transfactor

- The paradigm shift from investigating differences in gene expression to investigating regulatory differences in TF activity provides a more parsimonious way to interpret gene expression and allows the identification of a limited number of TFs that are driving gene expression differences.
- Applying tradeSeq to TF activity estimates for the OE injury response dataset allowed us to identify TFs involved in neurogenesis.

Inference of Transcription Factor Activity: transfactor

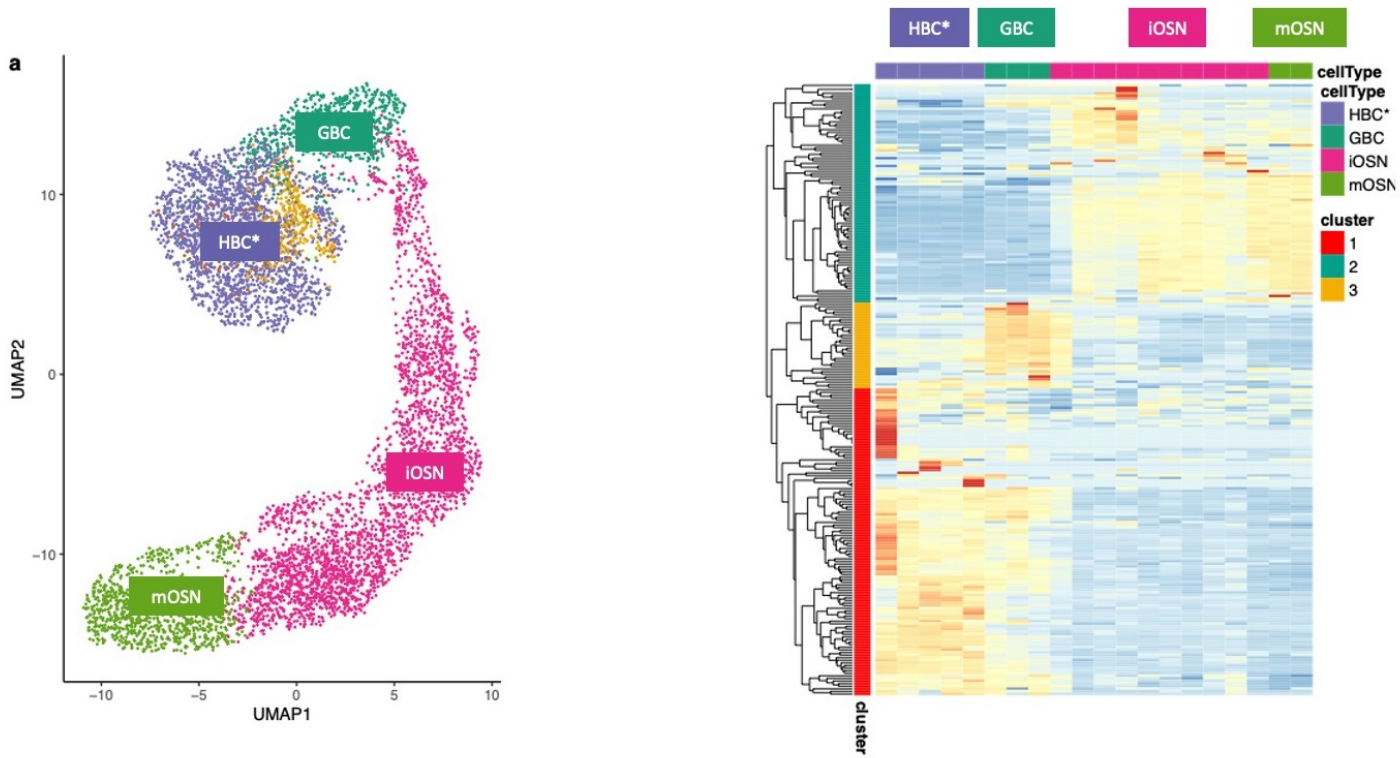


Figure 19: *transfactor: OE injury response dataset.* Left: UMAP representation of gene expression in neuronal lineage. Right: Heatmap of TF activity for differentially active TFs in neuronal lineage (tradeSeq associationTest), cells binned by pseudotime and most abundant cell type indicated in color bar.

Trajectory-Based Differential Transcription Factor Activity: transfactor

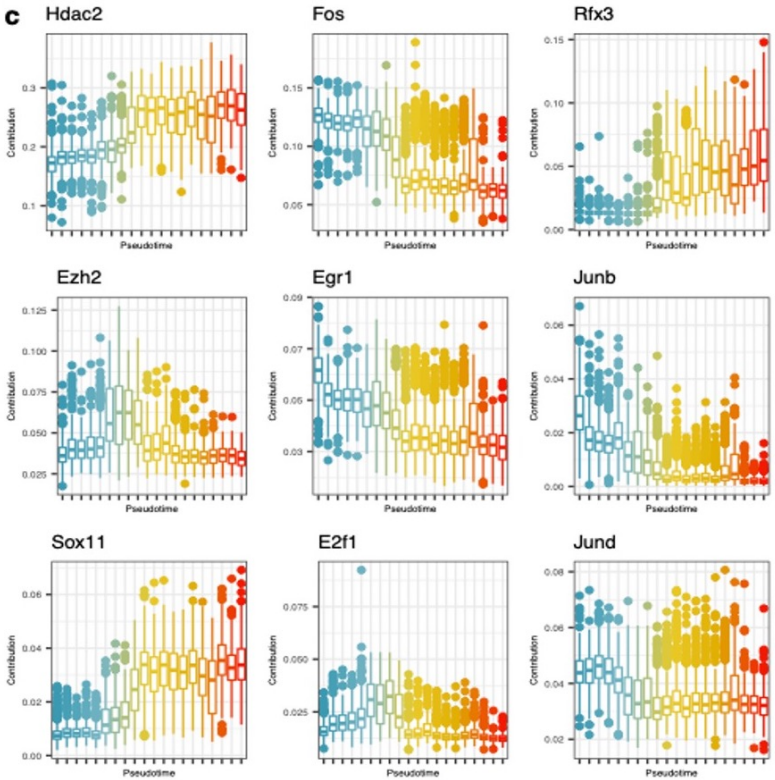


Figure 20: *transfactor: OE injury response dataset.* TF activity stratified by pseudotime for 9 TFs found to be most differentially active in the neuronal lineage (tradeSeq associationTest).

Spatial Multi-Omics

- Computational pathology involves the joint analysis of digitized pathology slides and spatially-resolved multi-omics data to improve our understanding of disease etiology, prevention, diagnosis, and treatment.
- Spatial transcriptomics provides high-throughput measures of genome-wide expression (i.e., mRNA) levels at (near-)cellular resolution while retaining a cell's (or group of cells') spatial location in a tissue.
- Spatial transcriptomics allows the identification of genes whose expression varies across spatial location, cell type, tissue architecture, or patient condition.
- Identifying differentially expressed genes can help understand a particular disease mechanism and identify therapeutic targets.

Kidney Transplant Spatial Multi-Omics Study

- Goal: Identify genes that are DE across kidney transplant conditions, tissue structures, or cell types while accounting for the spatial structure in the data.
- Data: Spatially-resolved gene and protein expression measures from kidney transplant biopsies (tissue slice).
  - Four conditions: Acute cellular rejection (ACR), antibody-mediated rejection (AMR), polyomavirus nephropathy (PNP), and normal morphology.
  - Three patients per condition.
  - One specimen per patient and five samples per specimen.
  - Expression measures for 960 genes.
  - Within-sample cell location coordinates, for a total of 219,462 cells across all samples.
  - For each sample, immunofluorescence images for 4 protein markers.
  - Platform: Nanostring CosMx imaging.

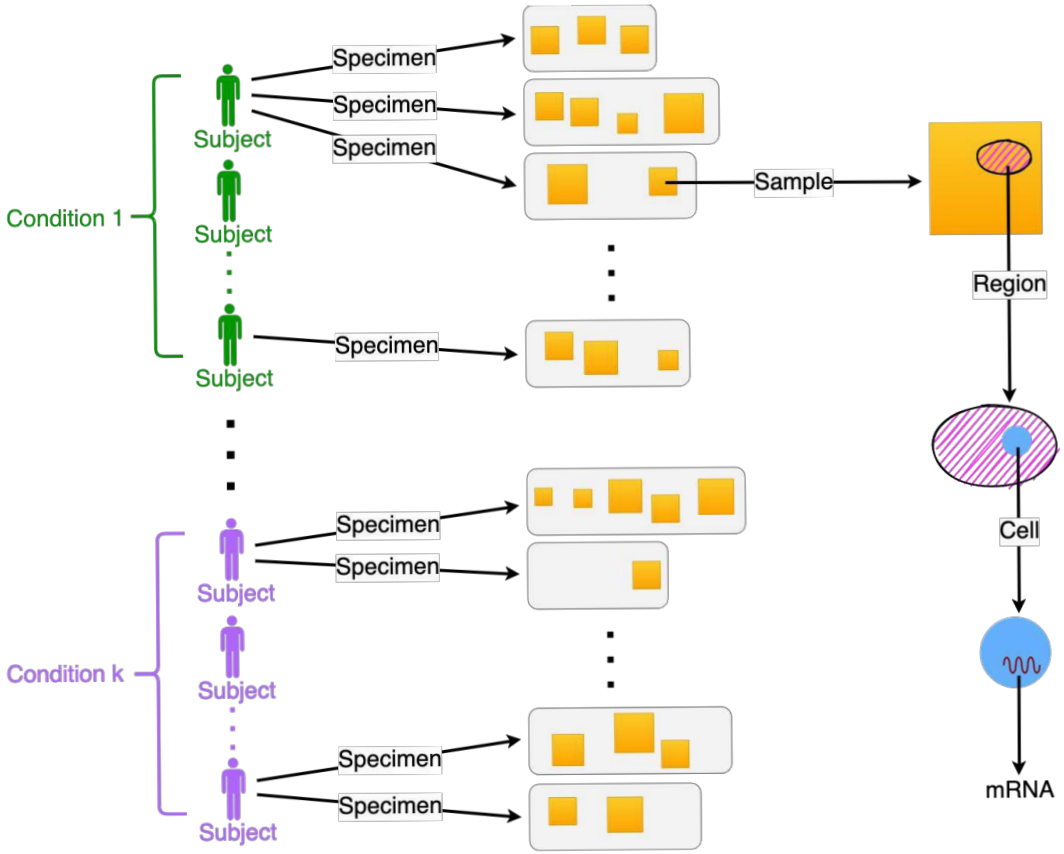


Figure 21: *Spatial multi-omics.* Design.

Kidney Transplant Spatial Multi-Omics Study

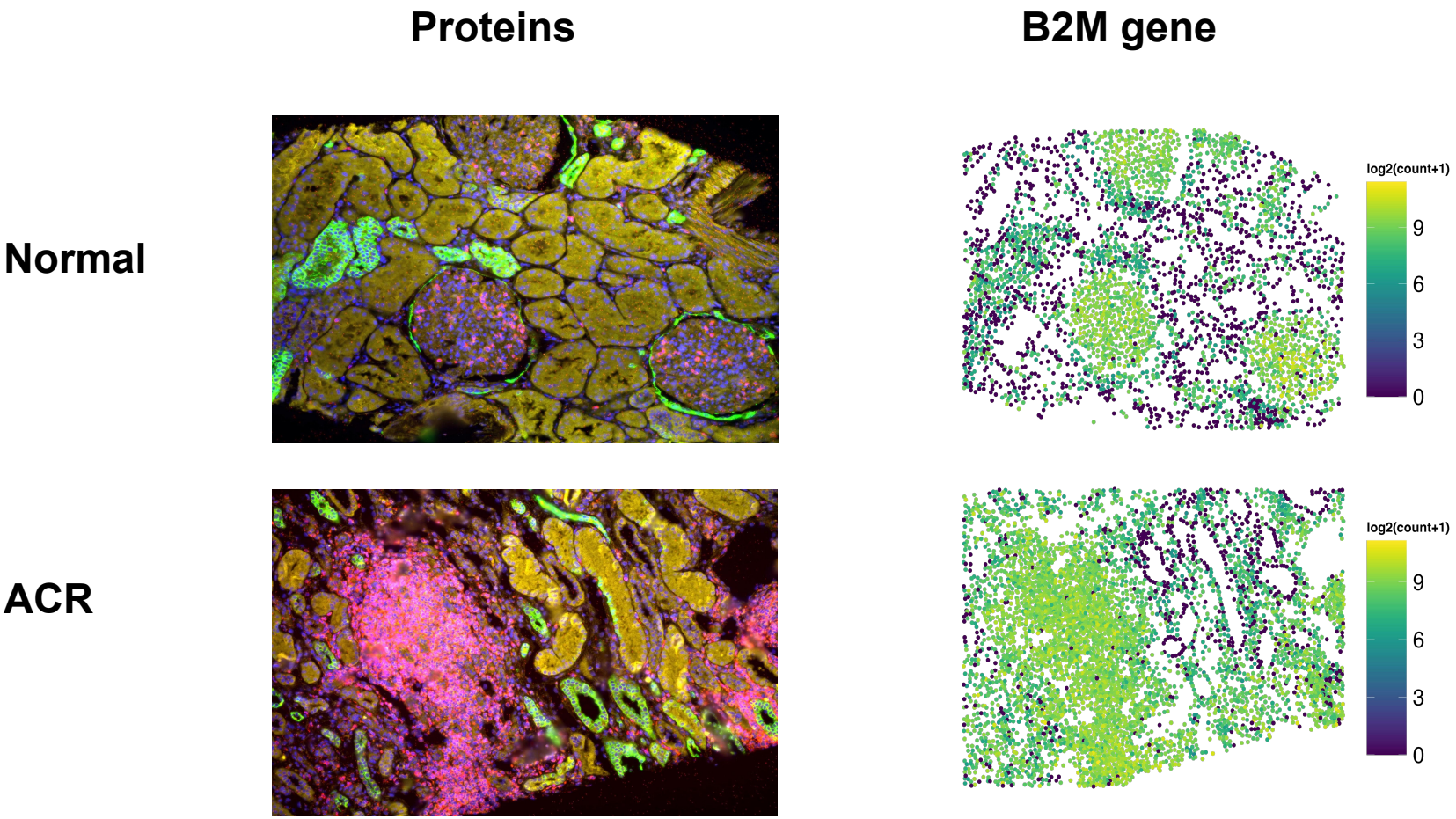


Figure 22: *Kidney transplant study.* Protein and gene expression measures.

Differential Expression in Spatial Transcriptomics

- For spatial transcriptomics, framing the question of differential expression is complex due to the range of dimensions along which a gene can be differentially expressed, including spatial location, cell type, and condition.
- Within-sample DE
  - Between two different cell types, e.g., immune cell subtypes.
  - Within a specific cell type, between two different tissue structures, e.g., between endothelial cells located in the glomeruli vs. interstitium kidney compartments.
- Between-sample DE
  - In the same tissue structure across two different conditions, e.g., glomeruli in acute cellular rejection and antibody-mediated rejection.
  - In the same cell type across two different conditions.

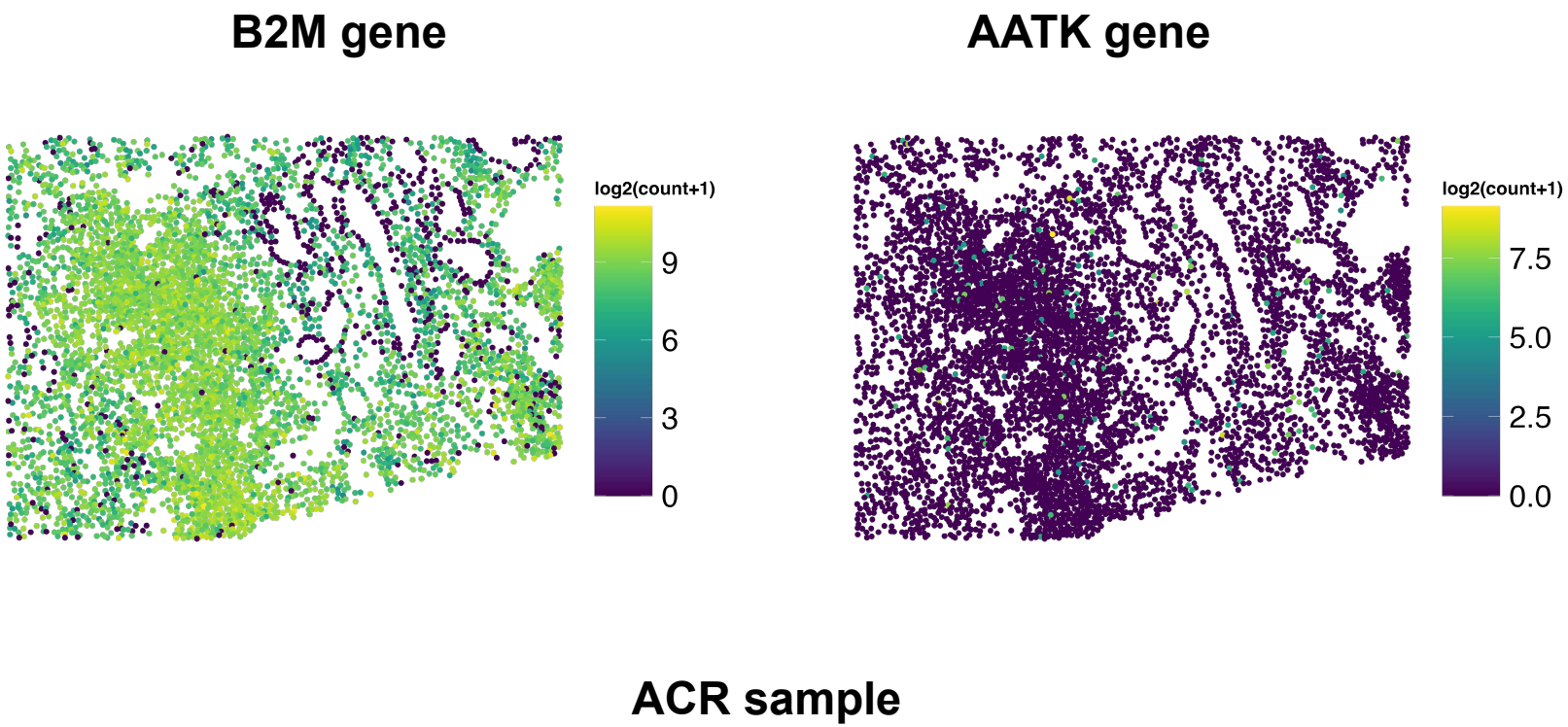


Figure 23: *Kidney transplant study: Within-sample DE.* Same sample, one spatially varying gene and the other not.

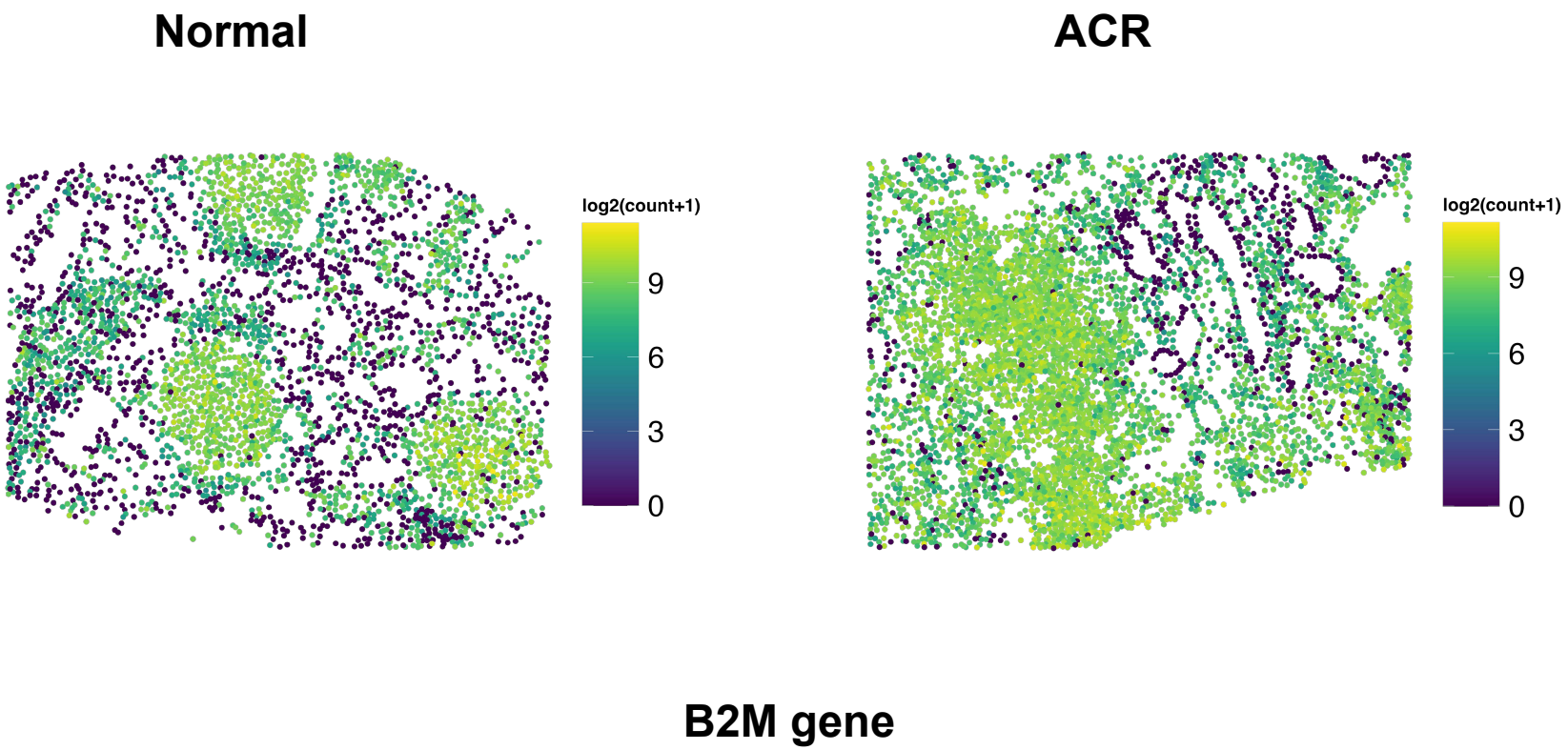


Figure 24: *Kidney transplant study: Between-sample DE.* Same gene, two different samples.

- Data
  - $Z_{i,j}$: For a given gene, observed expression measure in cell $j = 1, \ldots, J_i$ within sample $i = 1, \ldots, n$.
  - $\mathbf{X}_{i,j} \in I\!R^p$: Cell-level covariates, biological (e.g., biopsy, cell type, tissue structure, condition) or technical (e.g., slide).
- Gene-wise Poisson GLM

$$\begin{aligned} Z_{i,j} &\sim \text{Poisson}\left(C_{i,j}\theta_{i,j}\right) \\ \theta_{i,j} &= \exp\left(\mathbf{X}_{i,j}^{\top}\boldsymbol{\beta} + \phi_{i,j}\right), \end{aligned} \tag{1}$$

  - $C_{i,j}$ is a known location-dependent scaling factor (e.g., library size);
  - $\boldsymbol{\beta} \in I\!R^p$ captures (fixed) biological effects of interest, as well as technical effects, across samples;
  - $\boldsymbol{\phi}_i = (\phi_{i,j} : j = 1, \ldots, J_i)$ are within-sample spatial random effects, assumed to be independent of the covariates $\mathbf{X}_{i,j}$.

- The within-sample spatial random effects are modeled as

$$\boldsymbol{\phi}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_i\right),$$

where the covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{J_i \times J_i}$ captures spatial dependence between locations.

  - Lattice-based models: $\Sigma = \tau^2 Q^{-1}$, where $Q \in \mathbb{R}^{J \times J}$ depends on a user-defined adjacency matrix that specifies the neighbors of each observation and a parameter γ that controls the strength of the covariance between neighbors. τ^2 is a variance parameter.
    E.g. Conditional Autoregressive (CAR), Simultaneous Autoregressive (SAR), and the Leroux.
  - Sparse nearest-neighbor gaussian process models: $\boldsymbol{\Sigma} = K(\boldsymbol{\psi})$, for some kernel matrix K that is parameterized by $\boldsymbol{\psi}$.

- Fit by Expectation Conditional Maximization (ECM) algorithm.

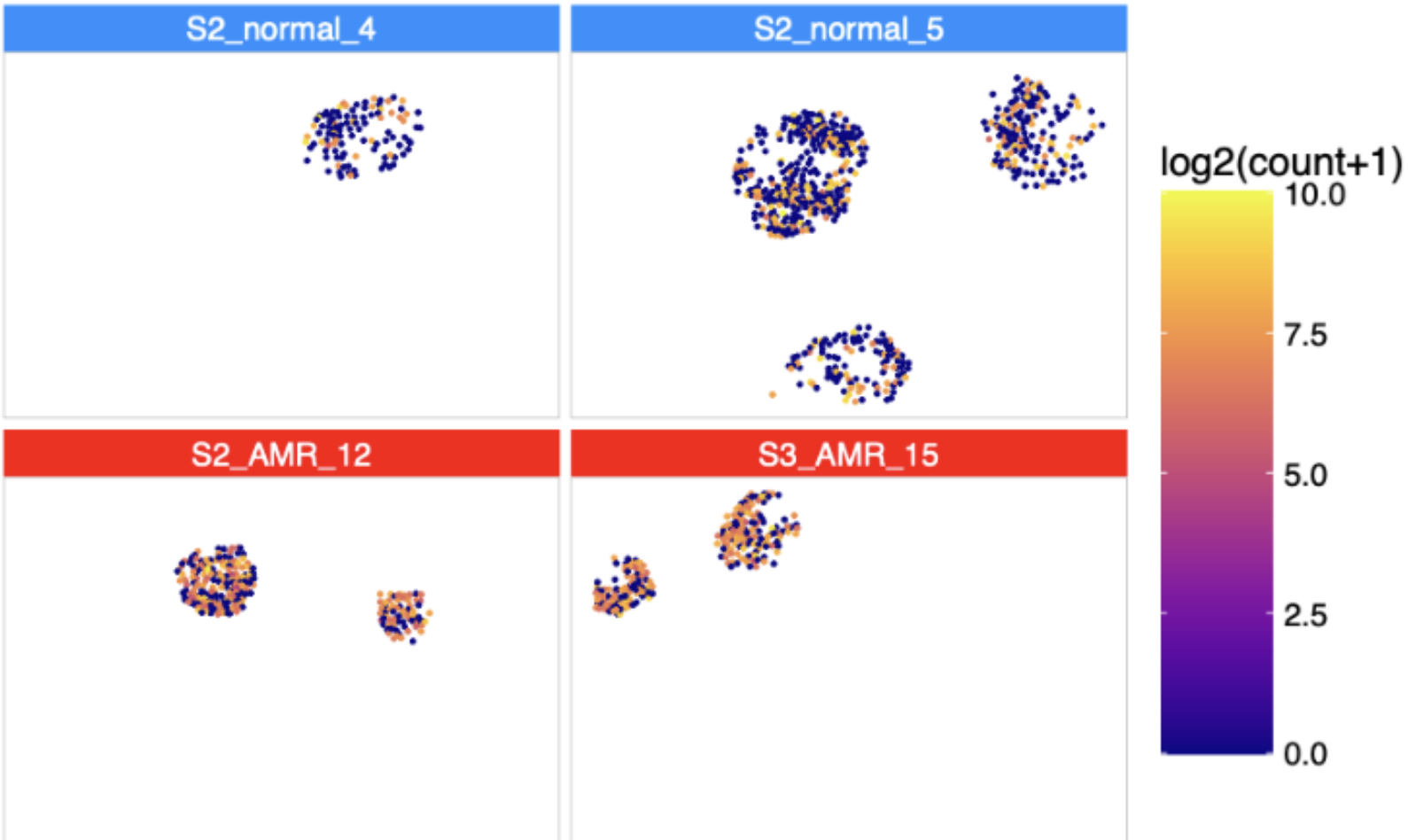


Figure 25: *Kidney transplant study: Between-sample DE.* DE gene (HLA-B) between AMR and normal samples in glomeruli cells.

Software

- R Project: `www.r-project.org`.
- Bioconductor Project: `www.bioconductor.org`.
- clusterExperiment: Resampling-based sequential ensemble clustering (RSEC).
  `www.bioconductor.org/packages/release/bioc/html/cluster`
- Dune: Cluster merging procedure to navigate the resolution-replicability trade-off.
  `www.bioconductor.org/packages/release/bioc/html/Dune.ht`
- EDASeq: Exploratory data analysis and normalization for RNA-Seq.
  `www.bioconductor.org/packages/release/bioc/html/EDASeq.`
- RUVSeq: Remove unwanted variation for RNA-Seq.
  `www.bioconductor.org/packages/release/bioc/html/RUVSeq.`

Software

- scone: Normalization procedures and performance assessment.
  www.bioconductor.org/packages/release/bioc/html/scone.h
- scPCA: Sparse contrastive principal component analysis.
  www.bioconductor.org/packages/release/bioc/html/scPCA.h
- slingshot: Cell lineage and pseudotime inference.
  www.bioconductor.org/packages/release/bioc/html/slingsh
- tradeSeq: Trajectory-based differential expression.
  www.bioconductor.org/packages/release/bioc/html/tradeSe
- condiments: Trajectory inference across multiple condition.
  www.bioconductor.org/packages/release/bioc/html/condime
- zinbwave: Zero-inflated negative binomial-based wanted variation extraction (ZINB-WaVE).
  www.bioconductor.org/packages/release/bioc/html/zinbwav
- Other packages listed at: www.bioconductor.org.

- F1000 Bioconductor workflow (Perraudeau et al., 2017): f1000research.com/articles/6-1158/.

See www.stat.berkeley.edu/~sandrine for publications, presentations, and software.

References

P. Boileau, N. S. Hejazi, and S. Dudoit. Exploring high-dimensional biological data with sparse contrastive principal component analysis. Bioinformatics, 03 2020a. ISSN 1367-4803. doi: https://doi.org/10.1093/bioinformatics/btaa176. btaa176.

P. Boileau, N. S. Hejazi, and S. Dudoit. scPCA: A toolbox for sparse contrastive principal component analysis in R. Journal of Open Source Software, 5(46):2079, 2020b. doi: https://doi.org/10.21105/joss.02079.

D. H. Brann, T. Tsukahara, C. Weinreb, M. Lipovsek, K. Van den Berge, B. Gong, R. Chance, I. C. Macaulay, H. Chou, R. Fletcher, D. Das, K. Street, H. Roux de Bézieux, Y.-G. Cho, D. Risso, S. Dudoit, E. Purdom, J. S. Mill, R. Abi Hachem, H. Matsunami, D. W. Logan, B. J. Goldstein, M. S. Grubb, J. Ngai, and S. R. Datta. Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. Science Advances, 6(31): eabc5801, 2020. doi: https://doi.org/10.1126/sciadv.abc5801.

References

J. H. Bullard, E. A. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 11:Article 94, 2010. URL `http://www.biomedcentral.com/1471-2105/11/94/abstract`. (Highly accessed).

M. B. Cole, D. Risso, A. Wagner, D. DeTomaso, J. Ngai, E. Purdom, S. Dudoit, and N. Yosef. Performance assessment and selection of normalization procedures for single-cell RNA-seq. Cell Systems, 8(4): 315–328, 2019. doi: https://doi.org/10.1016/j.cels.2019.03.010.

R. B. Fletcher, D. Das, L. Gadye, K. N. Street, A. Baudhuin, A. Wagner, M. B. Cole, Q. Flores, Y. G. Choi, N. Yosef, E. Purdom, S. Dudoit, D. Risso, and J. Ngai. Deconstructing olfactory stem cell trajectories at single-cell resolution. Cell Stem Cell, 20(6):817–830, 2017. doi: https://doi.org/10.1016/j.stem.2017.04.003.

L. Gadye, D. Das, M. A. Sanchez, K. Street, A. Baudhuin, A. Wagner, M. B. Cole, Y. G. Choi, N. Yosef, E. Purdom, S. Dudoit, D. Risso, J. Ngai, and R. B. Fletcher. Injury activates transient olfactory stem cell states with diverse lineage capacities. Cell Stem Cell, 21(6):775–790, 2017. doi: https://doi.org/10.1016/j.stem.2017.10.014.

References

F. Perraudeau, D. Risso, K. Street, E. Purdom, and S. Dudoit. Bioconductor workflow for single-cell RNA sequencing: Normalization, dimensionality reduction, clustering, and lineage inference. F1000Research, 6:1158, July 2017. doi: https://doi.org/10.12688/f1000research.12122.1.

D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. GC-content normalization for RNA-Seq data. BMC Bioinformatics, 12:Article 480, 2011. URL `http://www.biomedcentral.com/1471-2105/12/480/abstract`. (Highly accessed).

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology, 32 (9):896–902, 2014a. URL `http://www.nature.com/nbt/journal/vaop/ncurrent/full/nbt.2931.html`.

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. The role of spike-in standards in the normalization of RNA-seq. In S. Datta and D. Nettleton, editors, Statistical Analysis of Next Generation Sequencing Data, Frontiers in Probability and the Statistical Sciences, chapter 9, pages 169–190. Springer International Publishing, 2014b.

References

D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications, 9(1):284, 2018a. doi: https://doi.org/10.1038/s41467-017-02554-5.

D. Risso, L. Purvis, R. Fletcher, D. Das, J. Ngai, S. Dudoit, and E. Purdom. clusterExperiment and RSEC: A Bioconductor package and framework for clustering of single-cell and other large gene expression datasets. PLOS Computational Biology, 14(9):e1006378, 2018b. doi: https://doi.org/10.1371/journal.pcbi.1006378.

H. Roux de Bézieux, K. Street, S. Fischer, K. Van den Berge, R. Chance, D. Risso, J. Gillis, J. Ngai, E. Purdom, and S. Dudoit. Improving replicability in single-cell RNA-Seq cell type discovery with Dune. bioRxiv, 2020. doi: https://doi.org/10.1101/2020.03.03.974220.

H. Roux de Bézieux, K. Van den Berge, K. Street, and S. Dudoit. Trajectory inference across multiple conditions with condiments. Nature Communications, 15:833, 2024a. doi: https://doi.org/10.1038/s41467-024-44823-0.

References

H. Roux de Bézieux, K. Street, S. Fischer, K. Van den Berge, R. Chance, D. Risso, J. Gillis, J. Ngai, E. Purdom, and S. Dudoit. Improving replicability in single-cell RNA-Seq cell type discovery with Dune. BMC Bioinformatics, 25:198, 2024b. doi: https://doi.org/10.1186/s12859-024-05814-6.

K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics, 19(1):477, 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4772-0. URL https://doi.org/10.1186/s12864-018-4772-0.

C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nature Methods, 14(6):1–7, 2017. doi: https://doi.org/10.1038/NMETH.4292.

K. Van den Berge, F. Perraudeau, C. Soneson, M. I. Love, D. Risso, J.-P. Vert, M. D. Robinson, S. Dudoit, and L. Clement. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biology, 19(1):24, 2018. ISSN 1474-760X. doi: https://doi.org/10.1186/s13059-018-1406-4.

References

K. Van den Berge, H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement. Trajectory-based differential expression analysis for single-cell sequencing data. Nature Communications, 11:1201, 2020. doi: https://doi.org/10.1038/s41467-020-14766-3.

K. Van den Berge, P. Bickel, and S. Dudoit. transfactor: Transcription factor activity estimation via probabilistic gene expression deconvolution. bioRxiv, 2025. doi: https://doi.org/10.1101/2025.03.19.644088.

K. Van den Berge, D. Bakalar, H.-J. Chou, D. Kunda, D. Risso, K. Street, E. Purdom, S. Dudoit, J. Ngai, and W. Heavner. A latent activated olfactory stem cell state revealed by single-cell transcriptomic and epigenomic profiling. Stem Cell Reports, 2026. Accepted.