Statistical Methods and Software for Investigating Stem Cell Differentiation Using Single-Cell Transcriptome Sequencing

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

Department of Statistics and Actuarial Science
Simon Fraser University

March 23, 2018

BIDS
BERKELEY INSTITUTE FOR DATA SCIENCE

Version: 24/03/2018, 00:30

Acknowledgments

- Sandrine Dudoit, Department of Statistics, Division of Epidemiology and Biostatistics, Berkeley Institute for Data Science (BIDS), and Center for Computational Biology (CCB), UC Berkeley.
  - Kevin Benac.
  - Matthieu Doutreligne (Ecole Polytechnique).
  - Josie Hayes.
  - Partow Imani.
  - Amanda Moc.
  - Fanny Perraudeau. [ZINB-WaVE]
  - Davide Risso (now at Weill Cornell Medicine). [scone, ZINB-WaVE, RSEC, Slingshot]
  - Hector Roux de Bézieux.
  - Courtney Schiffman.
  - Kelly Street. [Slingshot]
- John Ngai, Department of Molecular and Cell Biology, UC Berkeley. [Biology]

  - Diya Das (BIDS Doctoral Fellow).
  - Russell Fletcher.
  - Levy Gadye.
- Lieven Clement, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium. [ZINB-WaVE]
  - Koen Van den Berge.
- Elizabeth Purdom, Department of Statistics and Center for Computational Biology, UC Berkeley. [scone, RSEC, Slingshot]
- Jean-Philippe Vert, Mines ParisTech and Institut Curie, Paris, France. [ZINB-WaVE]
  - Svetlana Gribkova.
- Nir Yosef, Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, UC Berkeley. [scone, Slingshot]
  - Michael Cole.

Acknowledgments

BRAIN Initiative. `www.braininitiative.nih.gov`.

Outline

Long and fruitful history of joint development between statistics and biology, with data at the core.

Figure 1: *G. J. Mendel (1822–1884).*

Statistics and Biological Data: 19th Century

Gregor Johann Mendel (1822–1884).

- Founding father of genetics: Mendel (1866), `www.mendelweb.org`.
- Mendel's laws of heredity were entirely based on statistical inference applied to data from carefully designed experiments.
- Statistical analyses allowed Mendel to make the bold leap from experimental results to theoretical conclusions.
- He reverse-engineered the Punnett square: He only observed phenotypes, but inferred genotypes.
- We now know that Mendel's hypothetical factors are genes, i.e., segments of DNA that code for proteins.
- Experimental confirmation came much later.
  - Genes lie on chromosomes (Sutton, 1903; Morgan, 1910).
  - DNA is the genetic material (Avery et al., 1944; Hershey and Chase, 1952).

Statistics and Biological Data: 19th Century

- Double helical structure of DNA (Watson and Crick, 1953).
- Genetic code (Nirenberg, 1961).

Statistics and Biological Data: Early 20th Century

Figure 2: *R. A. Fisher (1890–1962).*

Statistics and Biological Data: Early 20th Century

Fisher (1952). Statistical methods in genetics. *Heredity*, 6:1–12.

- *"Genetics and Statistics have in common that they are both characteristic products of the twentieth century."*
- *"... connection between our two subjects ...the "factorial" method of experimentation ... derives its structure, and its name, from the simultaneous inheritance of Mendelian factors."*
- *"Its [Genetics's] characteristic frequencies are a constant stimulus to statistical inquiry."*
- *"... beautifully randomized by the meiotic process."*
- *"Quite suddenly in the intellectual history of mankind it has become possible to think coherently and confidently about variation ..."*

Statistics and Biological Data: Early 20th Century

- *"Experimental design has become an intelligible subject for discussion ..."*
- *"It is not, I believe, sufficiently realized that this need for absolute realism is particularly required in statistical work when applied to genetic purposes. It is in general the statisticians task to bring theory into a truly organic coherence with objective and verifiable observations."*
- *"Direct contact with what is actually done in experimentation helps the statistician in another very essential way, by leading him to consider variations in procedure, and the reasons why one method is to be preferred to others. The whole wide subject of experimental design is opened out by this consideration."*

... 2018: The reign of -omic data. Today's statistical inference problems in biology and medicine are truly multivariate and involve the joint analysis of multiple, diverse, and high-dimensional datasets.

- Genome/Epigenome/Transcriptome. High-throughput microarray and sequencing measures for
  - identity-by-descent (IBD) states (GMS);
  - single nucleotide polymorphisms (SNP);
  - DNA copy numbers (CGH DNA-Chip/DNA-Seq);
  - transcript (mRNA) levels (RNA-Chip/RNA-Seq);
  - protein-nucleic acid interactions, e.g., transcription factor binding, histone modification (ChIP-Chip/ChIP-Seq);
  - DNA methylation status (methyl-Chip/methyl-Seq).
- Proteome. E.g. Mass spectrometry and protein microarrays.

Statistics and Biological Data: Today

- Metagenome. High-throughput abundance measures for thousands of bacteria in human or environmental samples. E.g. 16S rRNA sequencing.
- Exposome. High-throughput adductomic and metabolomic exhaustive and untargeted measures of chemicals in human biospecimens that can be used in exposome-wide association studies (EWAS).
- Covariates. E.g. Age, sex, environmental exposure, treatment, dose, time.
- Phenotypes. Biological and clinical outcomes, e.g., cell type/state, affectedness/unaffectedness, quantitative trait, (censored) survival time, response to treatment.
- Biological annotation metadata. In-house or WWW, e.g., Gene Ontology (GO), pathway (KEGG), protein structure (PDB), literature (PubMed).

Statistics and Biological Data: Today

Data complexity.

- Dimensionality.
  - "Small n, large p", e.g., microarrays.
  - "Large n, large p", e.g., 10x Genomics scRNA-Seq.
- Multiple types. Quantitative (continuous, discrete), qualitative, text, graph, image.
- Censored, erroneous, missing, sparse (cf. zero inflation).
- Various levels of processing. E.g. Microarray and sequencing data.
- Dynamic and evolving. E.g. DNA sequence (GenBank), Gene Ontology (GO), literature (PubMed).
- Multiple sources and locations. In-house, WWW.

No longer just numerical data or $X_{n \times p}$!

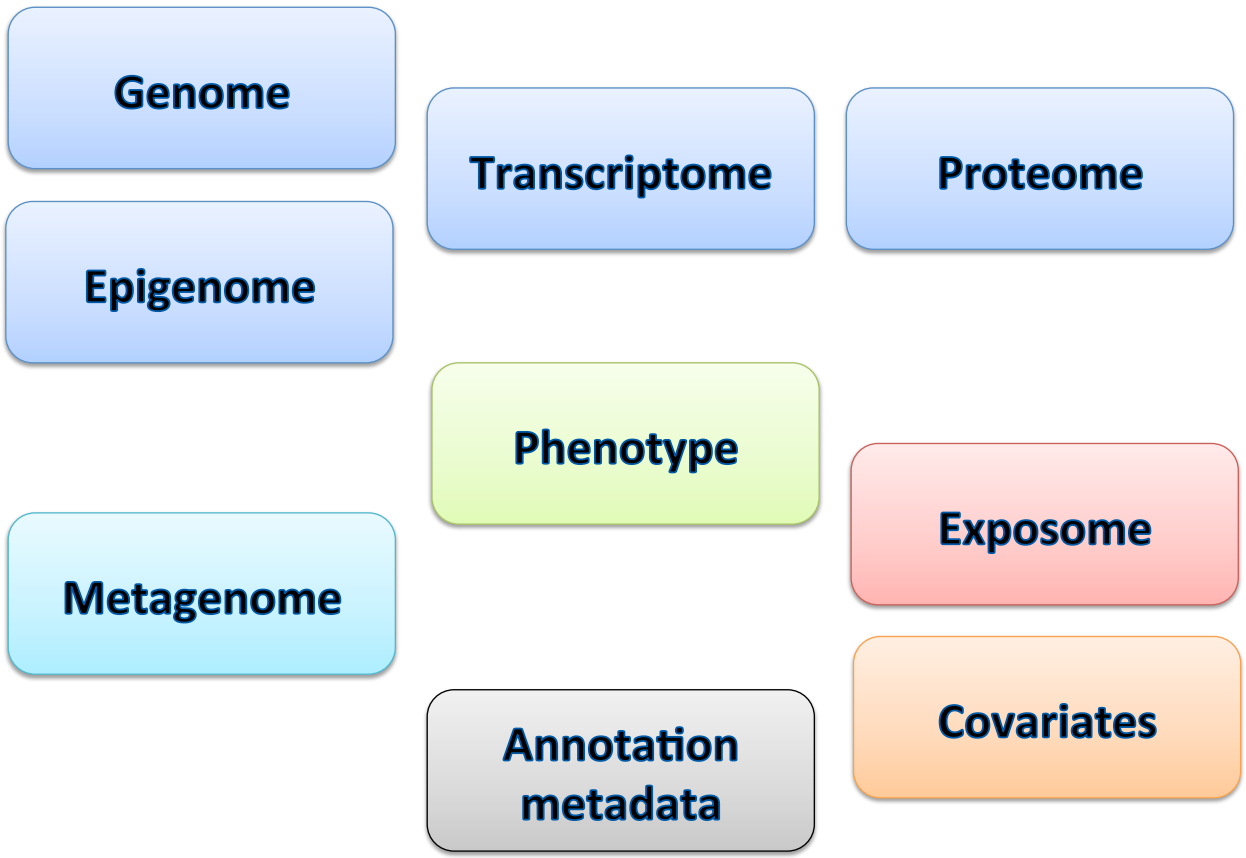


Figure 3: *Biological and medical -omic data.*

Olfactory Stem Cells and Neural Regeneration

Fletcher et al. (2017). Stem cell differentiation in the mouse olfactory epithelium.

- Adult olfactory stem cells support the replacement of olfactory sensory neurons and non-neuronal supporting cells (e.g., sustentacular) over postnatal life and can reconstitute the entire olfactory epithelium (OE) following injury.
- The horizontal basal cell (HBC) is an adult tissue stem cell.
- The p63 protein (tumor protein p63, TP63) promotes self-renewal of HBC by blocking differentiation.
- When p63 is down-regulated, this "brake" is removed, allowing differentiation to proceed 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 vs. differentiation.

Olfactory Stem Cells and Neural Regeneration

- In order to investigate the differentiation of HBC, single-cell transcriptome sequencing (scRNA-Seq) was performed for wild-type (WT) and p63 knock-out (KO) mice, at five timepoints following tamoxifen treatment.
- By combining single-cell extraction and high-throughput sequencing capabilities, single-cell RNA-Seq allows the transcriptome of large numbers of individual cells to be assayed efficiently.

Olfactory Stem Cells and Neural Regeneration

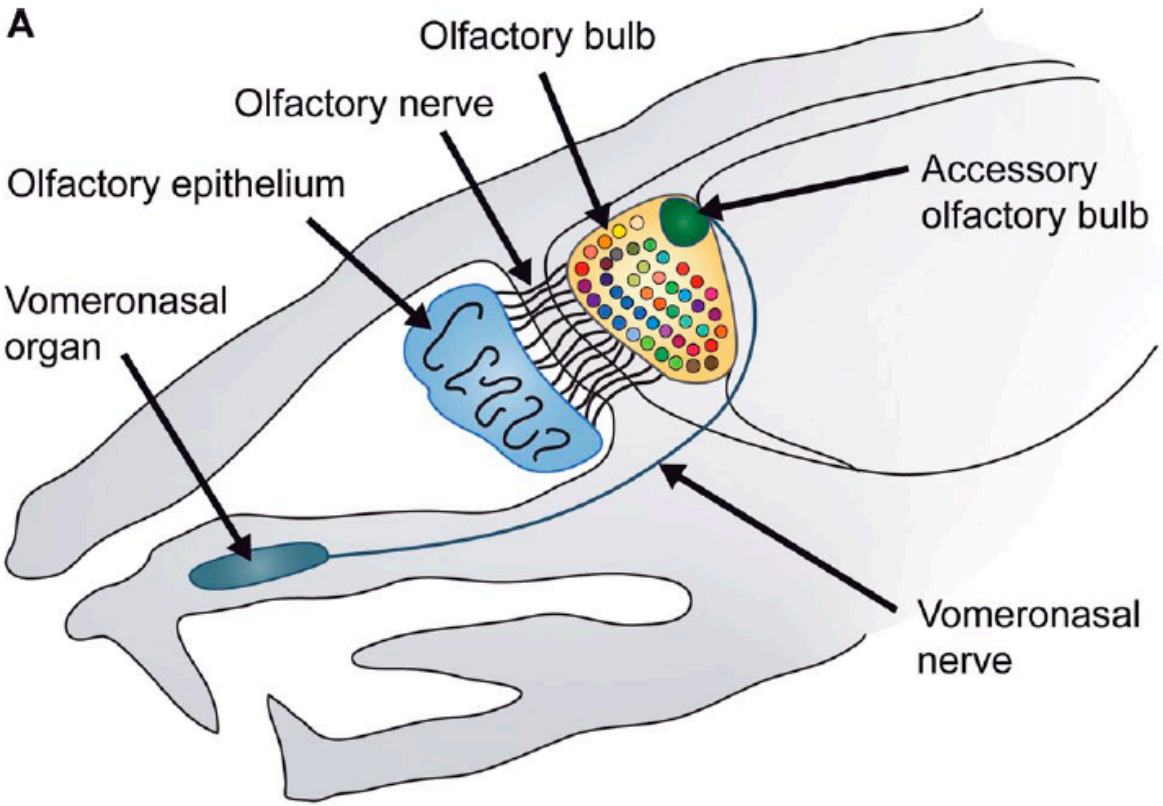


Figure 4: *Mouse olfactory epithelium.*

Olfactory Stem Cells and Neural Regeneration

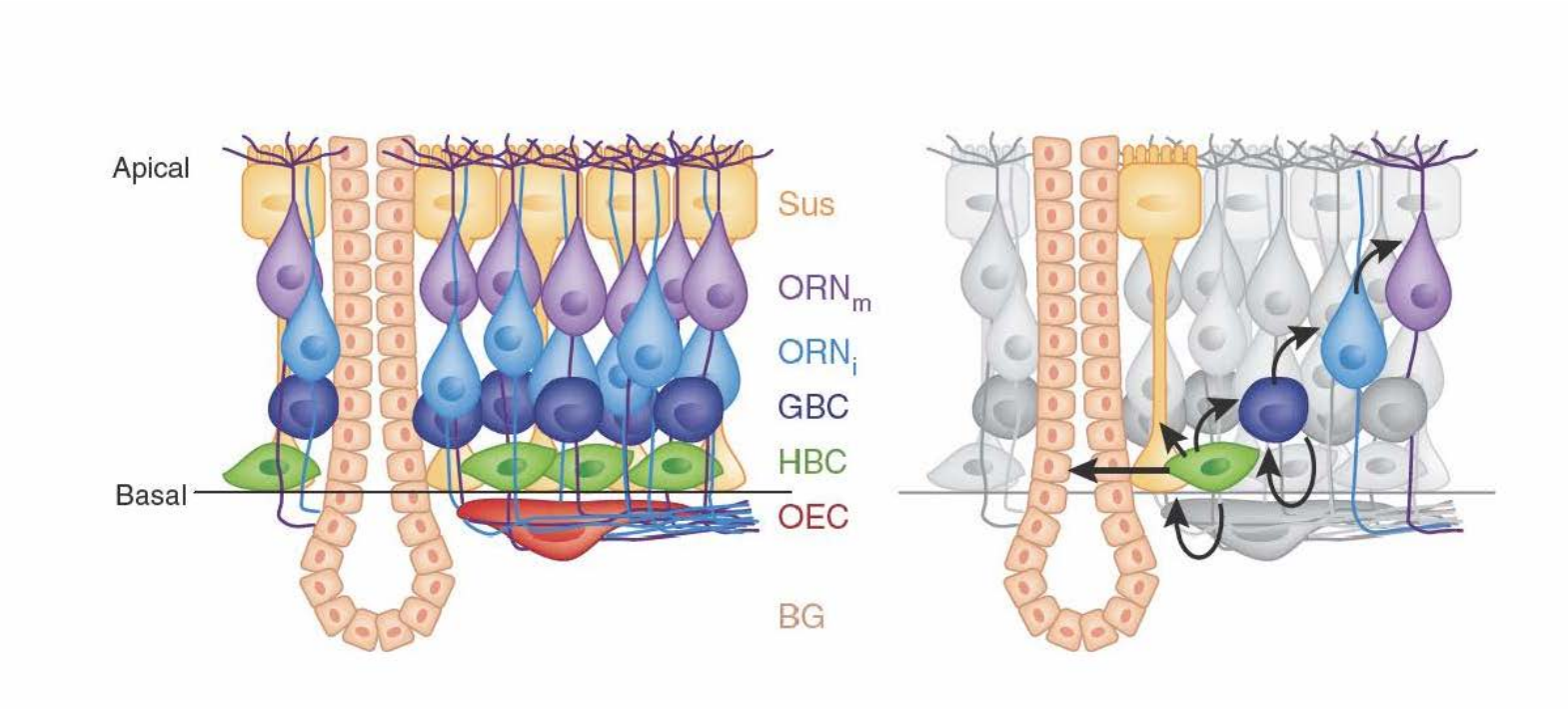


Figure 5: *Olfactory epithelium cell types.* Sus: sustentacular cell, ORN: olfactory receptor neuron, GBC: globose basal cell, HBC: horizontal basal cell, OEC: olfactory ensheathing cell, BG: Bowman gland.

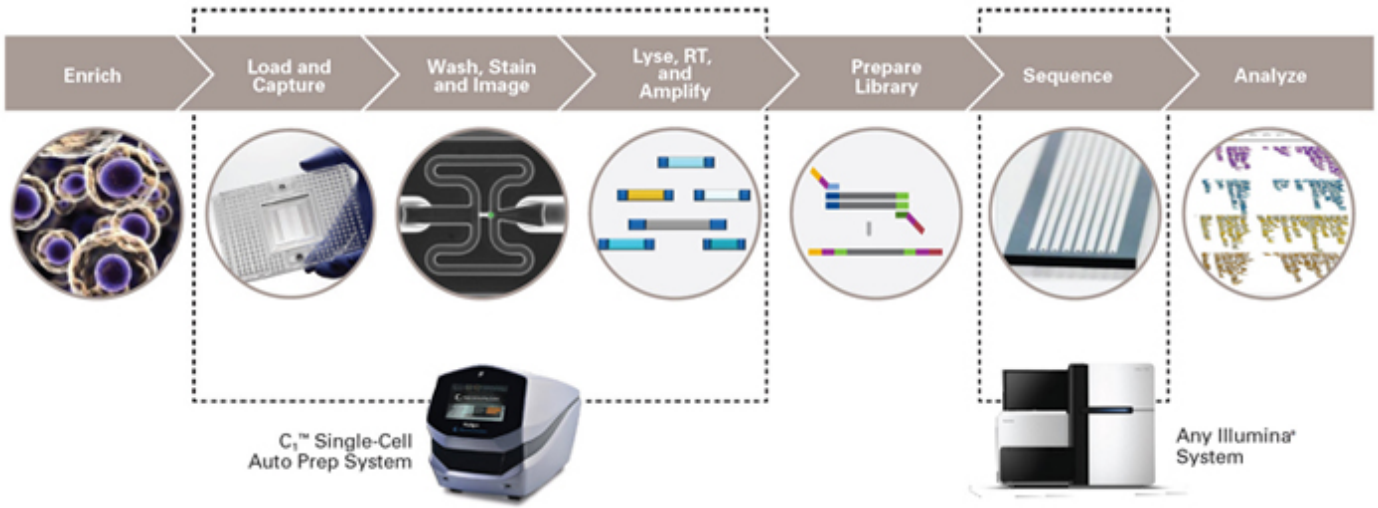


Figure 6: *Single-cell RNA-Seq.* Fluidigm C1 Single-Cell Auto Prep System.

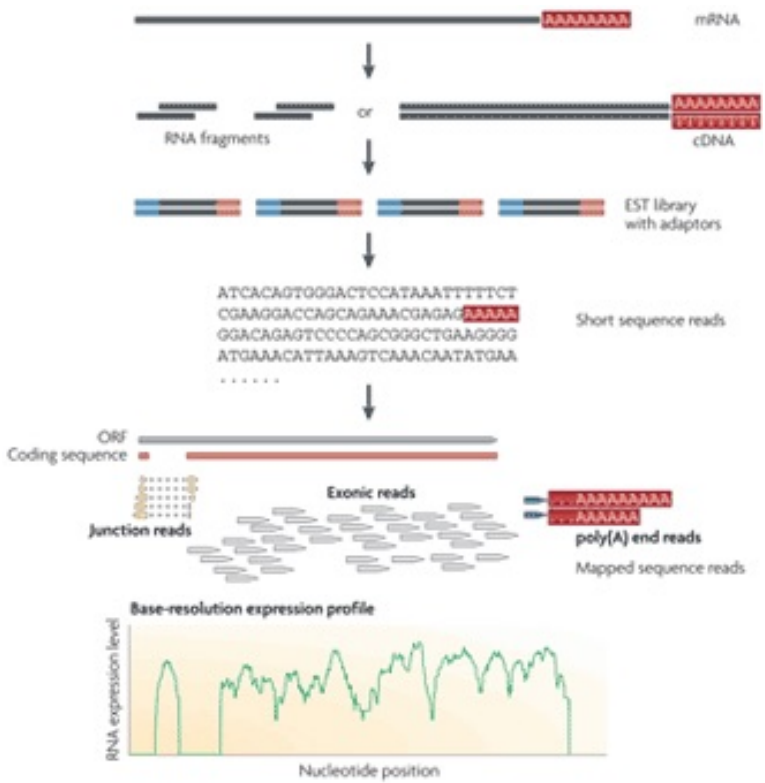


Figure 7: *RNA-Seq.*

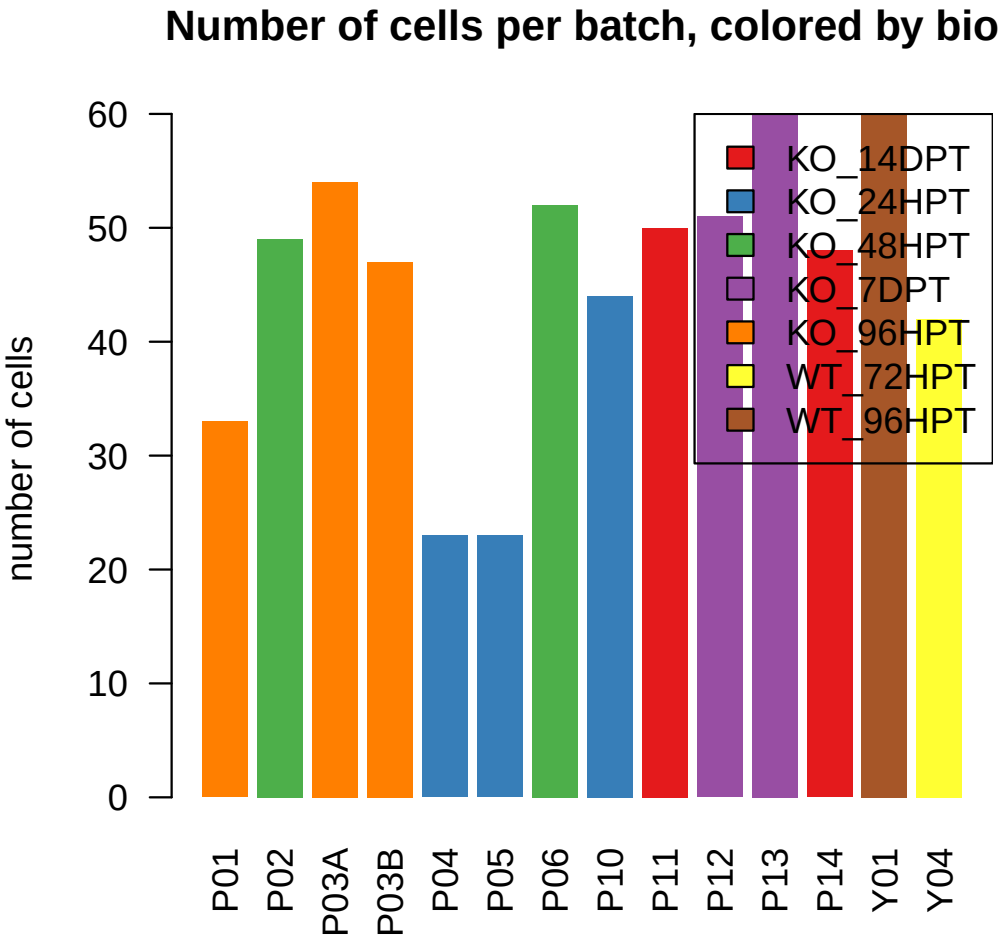


Figure 8: *OE dataset: Experimental design.* Number of cells per batch (C1 run), color-coded by biological condition (WT/KO, timepoint). Batch effects are nested within biological effects.

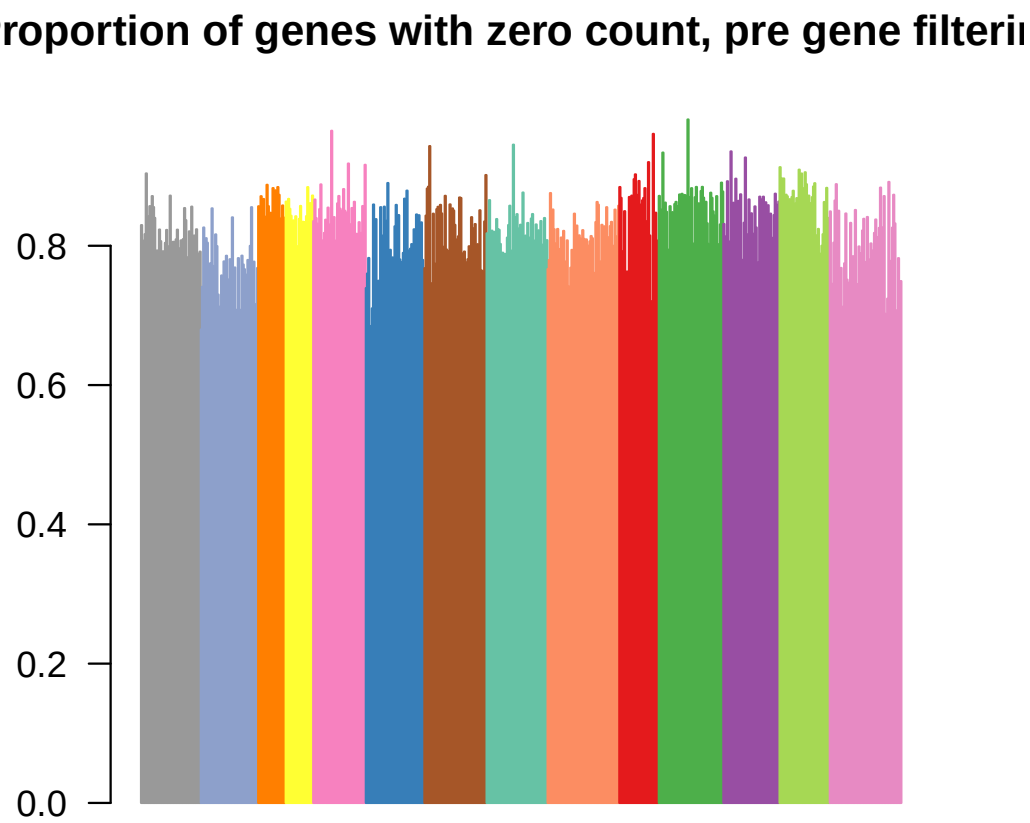


Figure 9: *Zero inflation.* Per-cell proportion of genes with zero count, color-coded by batch.

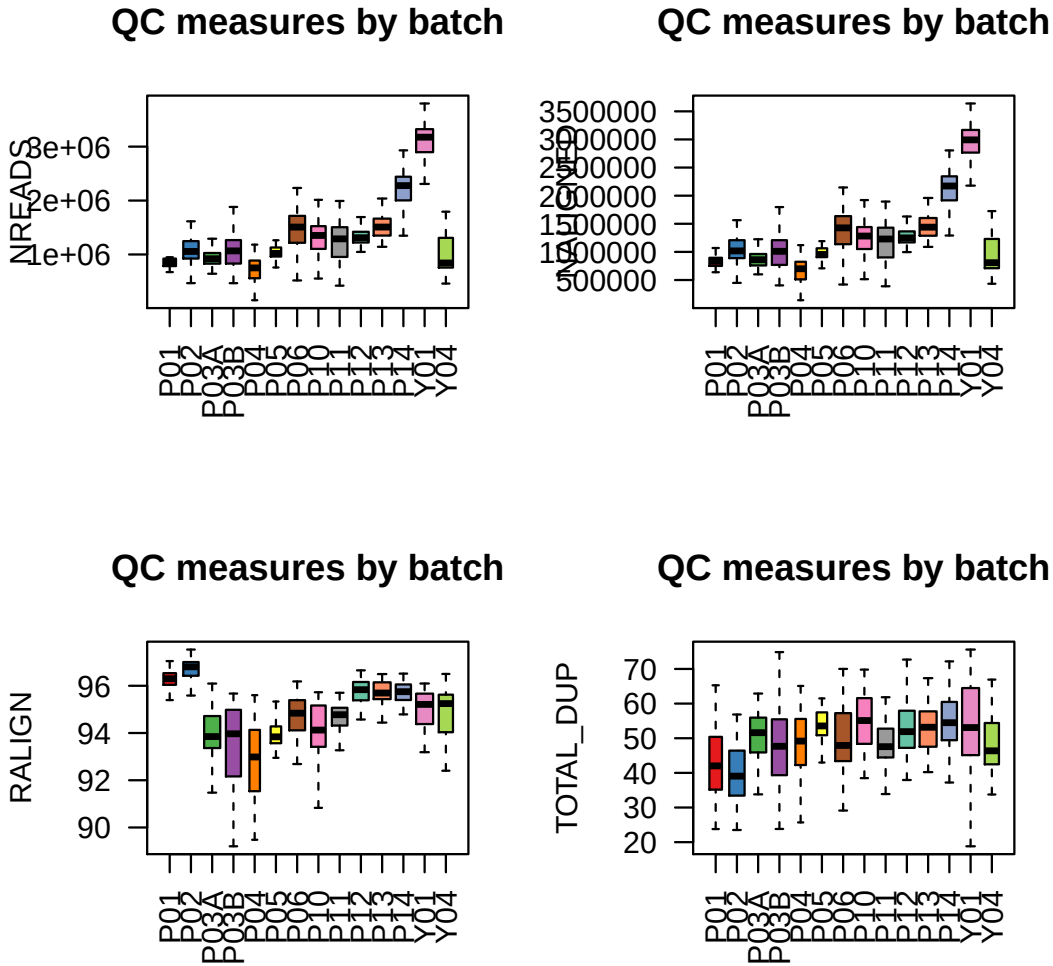


Figure 10: *Sample-level QC.* Boxplots of QC measures, by batch.

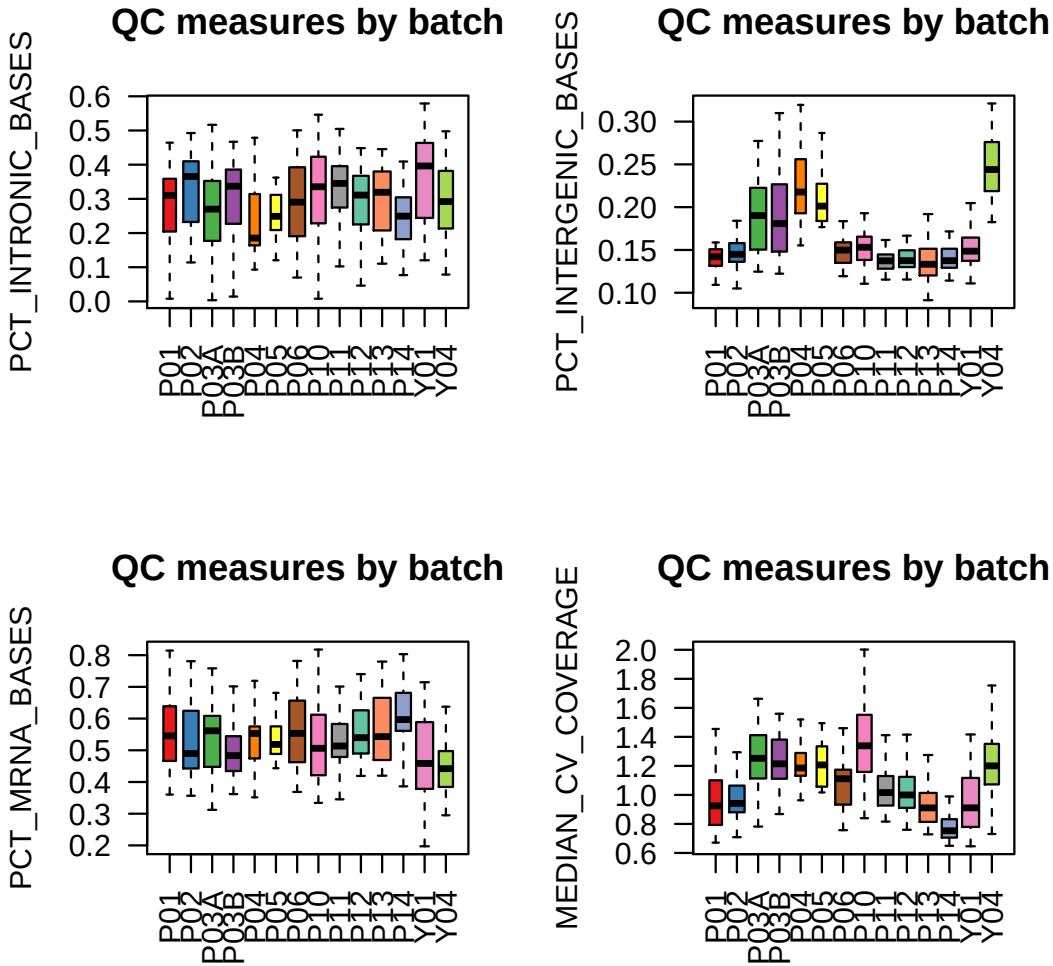
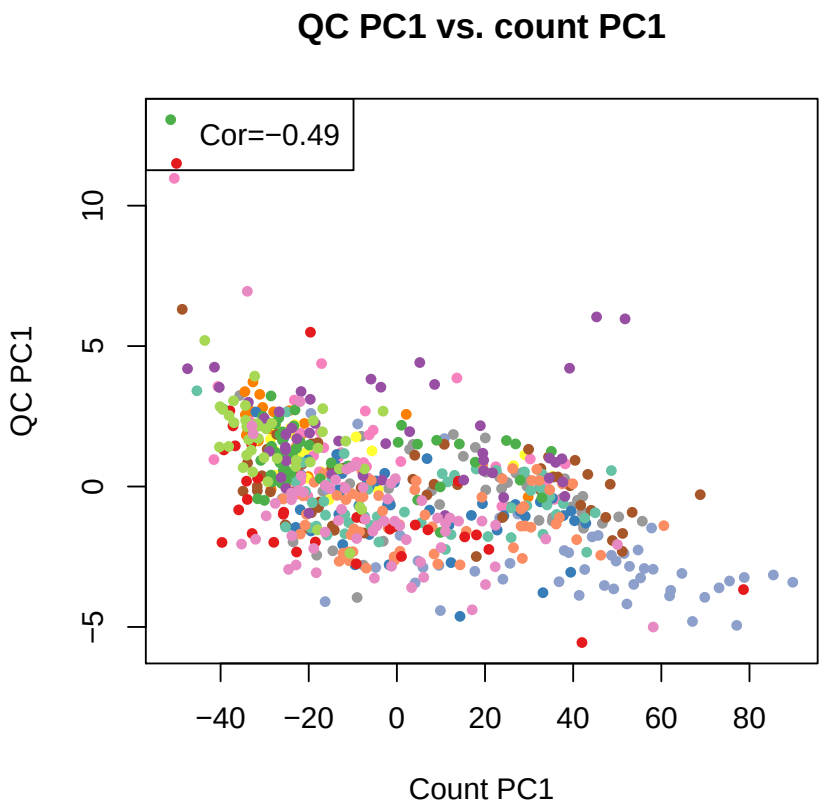
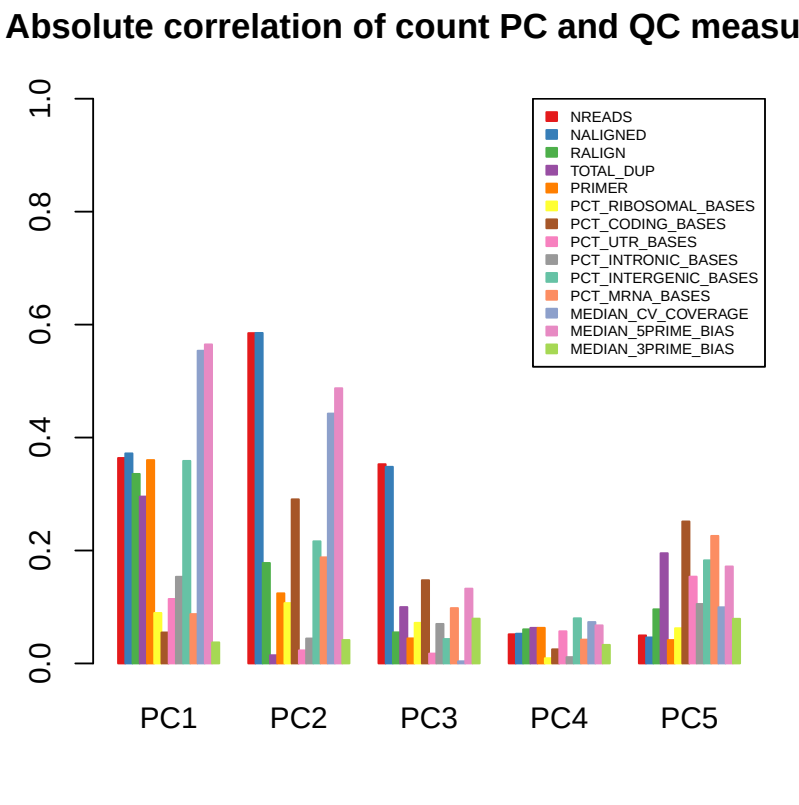


Figure 11: *Sample-level QC.* Boxplots of QC measures, by batch.

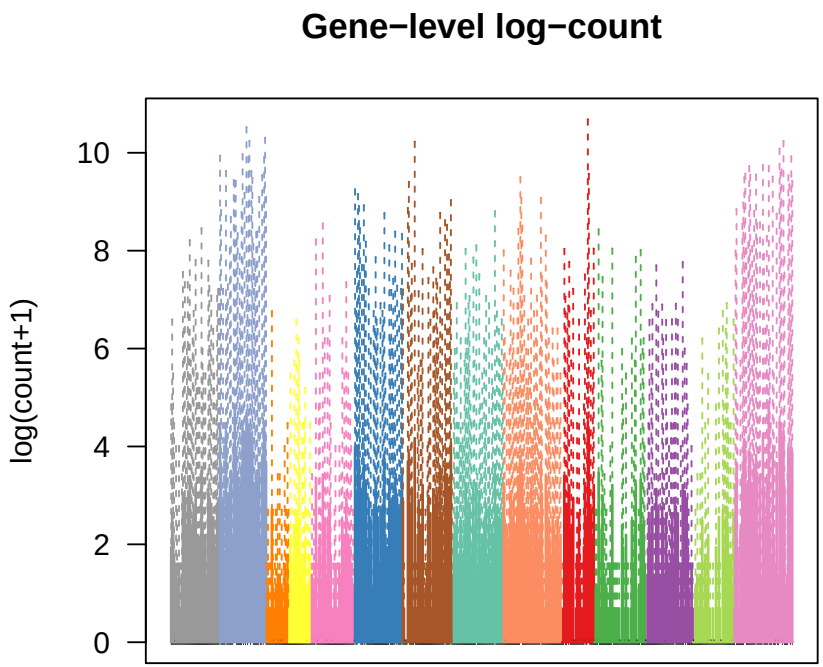


(a) QC PC1 vs. count PC1, color-coded by batch

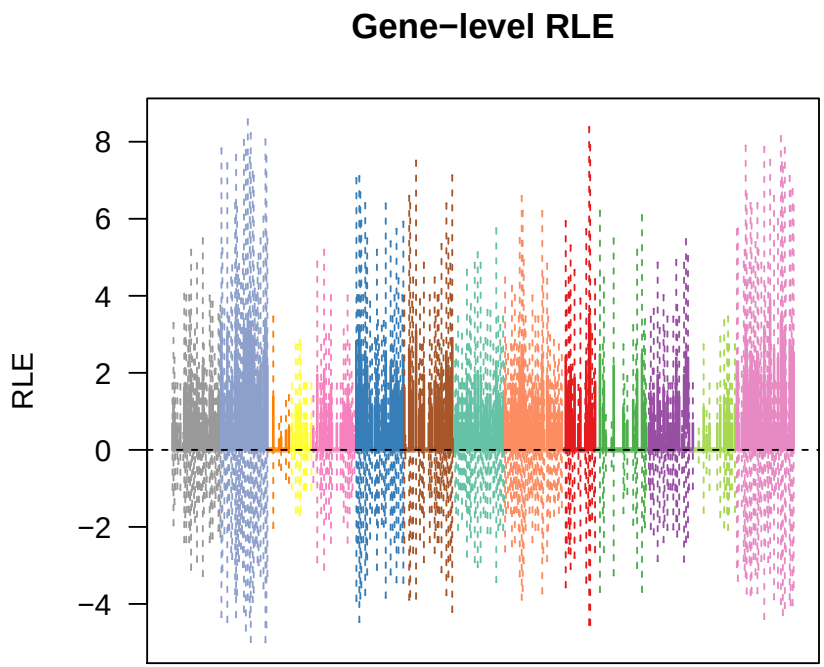


(b) Correlation of count PC and QC measures

Figure 12: *Sample-level QC.* Association of counts and sample-level QC measures.



(a) Log-count



(b) RLE

Figure 13: *Gene-level counts.* 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.

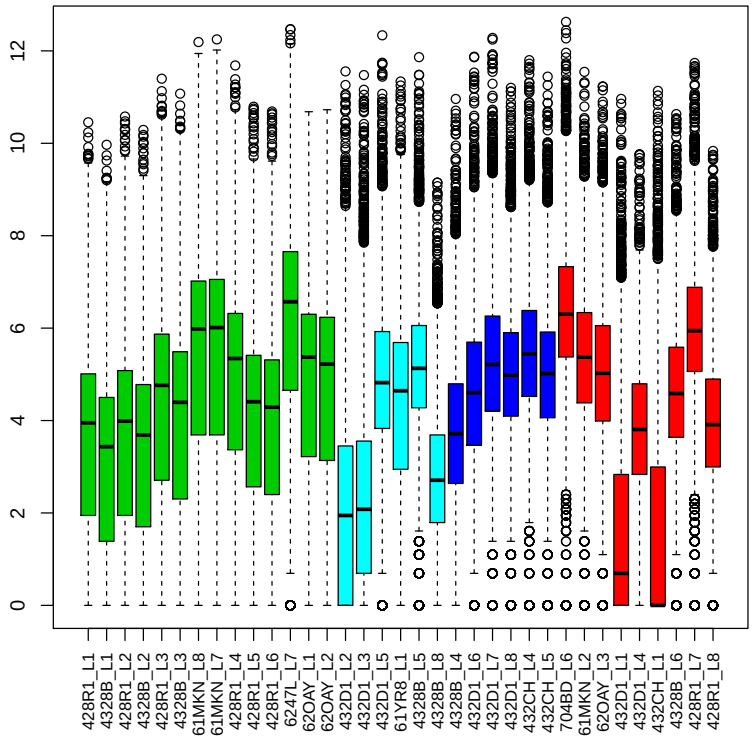


Figure 14: *QA/QC: Mislabeled species, yeast bulk RNA-Seq dataset.* Boxplots of per-sample unnormalized log(count+1), color-coded by species, green: *S. cerevisiae*, cyan: *S. paradoxus*, blue: *S. mikatae*, red: *S. bayanus*.

Exploratory Data Analysis: Summary

- It is essential to look at data to identify the main features of a dataset and for quality assessment/control – Garbage in, garbage out.
- Zero inflation.
  - Single-cell RNA-Seq data have many more genes with zero read counts than bulk RNA-Seq data.
  - This zero inflation could occur for biological reasons (i.e., the gene is simply not expressed, transcriptional bursting) or technical reasons (e.g., low mRNA capture efficiency, dropouts).
  - Most RNA-Seq normalization methods involve scaling or ranking and perform poorly when many genes have zero counts.

Exploratory Data Analysis: Summary

- In particular, the global-scaling method of Anders and Huber (2010), implemented in the Bioconductor R package DESeq, discards any gene having zero count in at least one sample. In practice, the scaling factors are therefore estimated based on only a handful of genes, e.g., 5/22,054 genes for OE dataset.
- Full-quantile (FQ) normalization also doesn't behave properly due to ties from the large number of zeros.
- Zero-count gene filtering is advisable before normalization and downstream analyses.
- We apply the following zero-count gene filtering to the OE dataset: Retain only the genes with at least $n_r = 20$ reads, in at least $n_s = 40$ cells.
  This yields 9,133/22,054 genes.

- Sample-level QC measures.
  - The distribution of QC measures can vary substantially within and between batches.

Exploratory Data Analysis: Summary

- Some QC measures clearly point to low-quality samples, e.g., low percentage of mapped reads (RALIGN).
- There can be a strong association between QC measures and read counts (cf. PCA).
- Filtering samples based on QC measures is advisable, as normalization procedures may not be able to adjust for QC and some samples simply have low quality.
- Normalization procedures based on QC measures (e.g., regression on first few PC of QC measures) should also be considered.

- Normalization.
  - After gene and sample filtering and before normalization, there are large differences in gene-level count distributions within and between batches (cf. RLE, housekeeping genes).
  - The counts are still zero-inflated.
  - There can be substantial association of counts and sample-level QC measures.

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 are truly due to the 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

- A variety of normalization procedures have been proposed, most being adaptations of methods for bulk RNA-Seq and microarrays, ranging from simple global-scaling methods to regression on known and unknown gene and sample-level covariates.
- However, widely-used bulk RNA-Seq methods are not well-suited for handling the increased technical noise and zero inflation of scRNA-Seq data (Vallejos et al., 2017).
- It is still common to only scale gene-level counts by total read count, e.g., counts per million (CPM), reads per million (RPM), and reads per kilobase of exon model per million mapped reads (RPKM). In many cases, however, one needs more aggressive normalization that accounts for more complex features of the genomic regions and experimental design.

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 for selecting a suitable normalization procedure – scone.

scone. A general framework for the normalization of scRNA-Seq data (and other data).

- 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.
- Numerical and graphical summaries of normalized read counts and metrics.
- Shiny app.
- Bioconductor R package scone: `www.bioconductor.org/packages/release/bioc/html/scone.h`

Normalization: scone

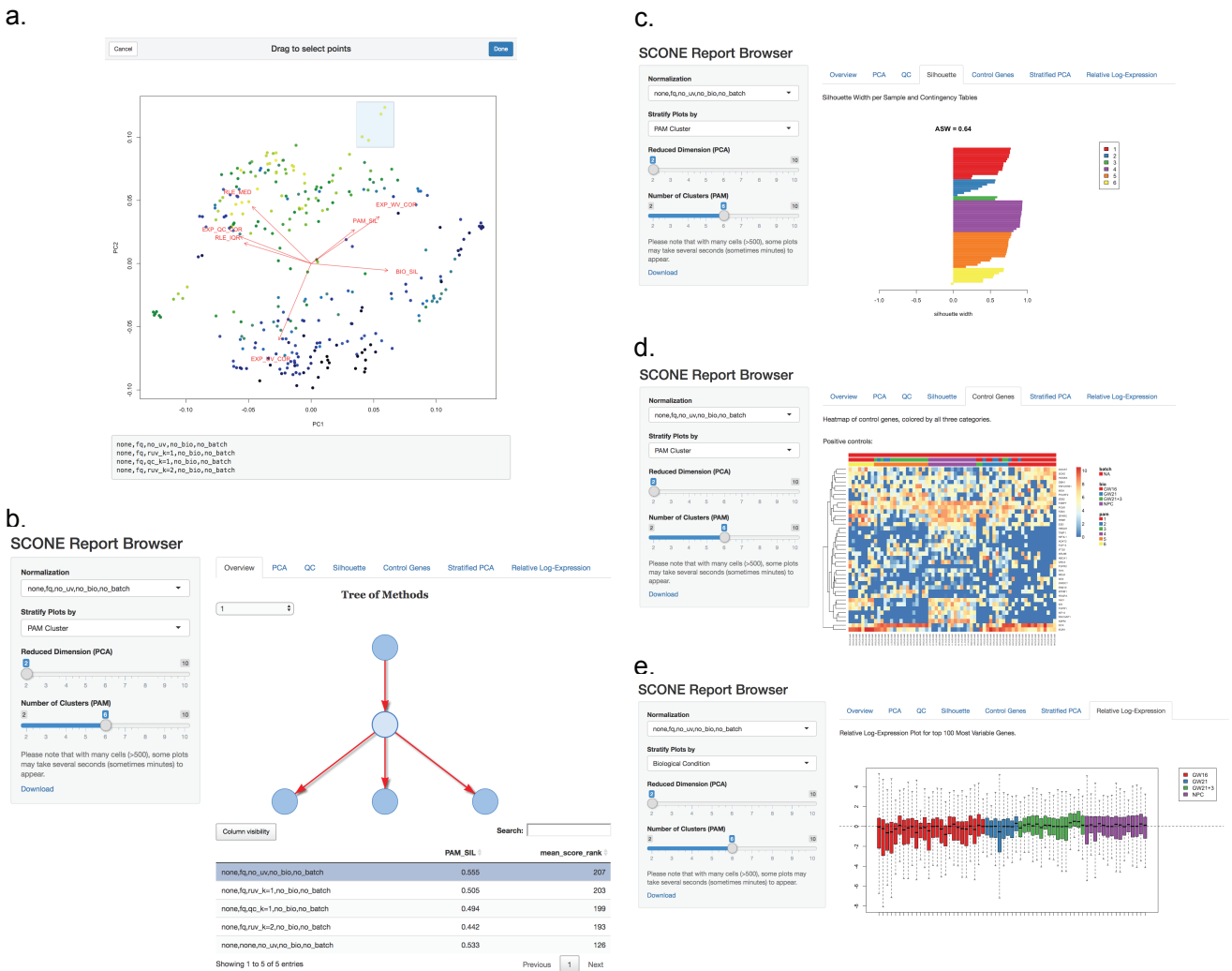


Figure 15: *scone*. Shiny app, report browser.

Normalization: scone

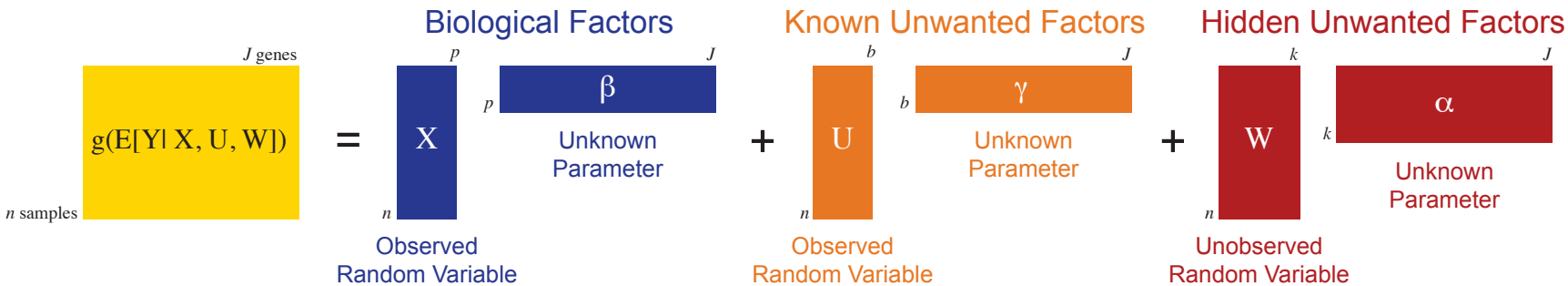


Figure 16: *scone*. Regression model for supervised and unsupervised normalization based on, respectively, known and unknown factors of unwanted variation.

Normalization performance metrics. Capture trade-offs between ability to remove unwanted variation and ability to preserve biological variation of interest.
Green: Good when high; Red: Good when low.

- Clustering of samples according to factors of wanted and unwanted variation.
  - BIO_SIL: Average silhouette width by biological condition of interest.
  - BATCH_SIL: Average silhouette width by batch.
  - PAM_SIL: Maximum average silhouette width for partitioning around medoids (PAM) clusterings, for a range of user-supplied numbers of clusters.
- Association of expression measures with factors of wanted and unwanted variation.

  - EXP_WV_COR: Based on coefficient of determination (R^2) for regression of count PCs on factors of wanted variation (derived from positive control genes).
  - EXP_UV_COR: Based on coefficient of determination (R^2) for regression of count PCs on factors of unwanted variation (derived from negative control genes, preferably distinct from those used in RUV).
  - EXP_QC_COR: Based on coefficient of determination (R^2) for regression of count PCs on QC measures.
- Between-sample distribution of expression measures.
  - RLE_MED: Mean squared median relative log expression (RLE).
  - RLE_IQR: Variance of inter-quartile range (IQR) of RLE.
- Select a normalization procedure based on (a function of) the performance metrics.

  E.g. Average metric or average metric rank.

- User can define and add other metrics, depending on application. We've adapted and applied scone to normalize adductomic and metabolomic data.

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`.

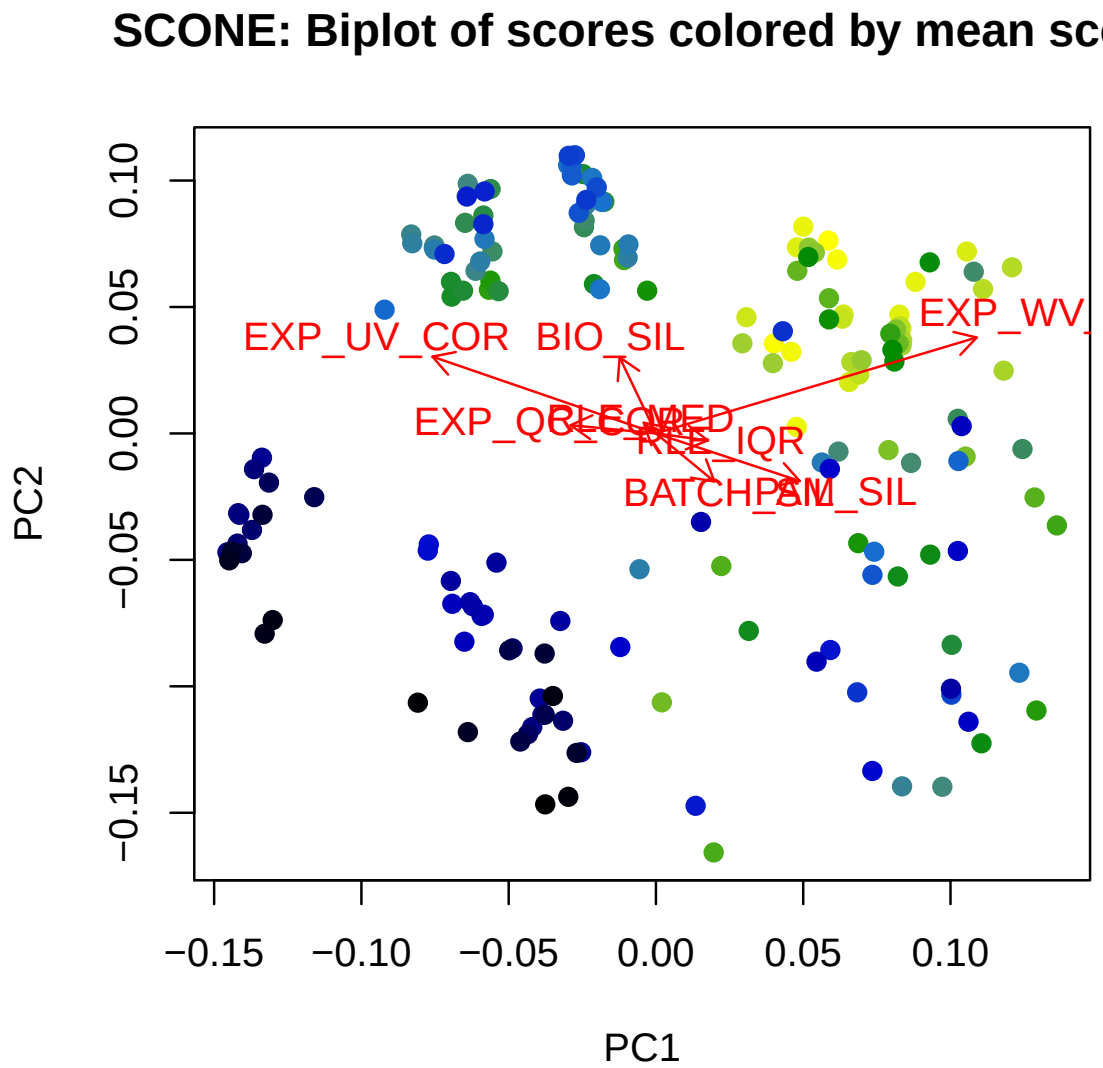


Figure 17: *scone*. Biplot of performance scores, colored by mean score (yellow high/good, blue low/bad).

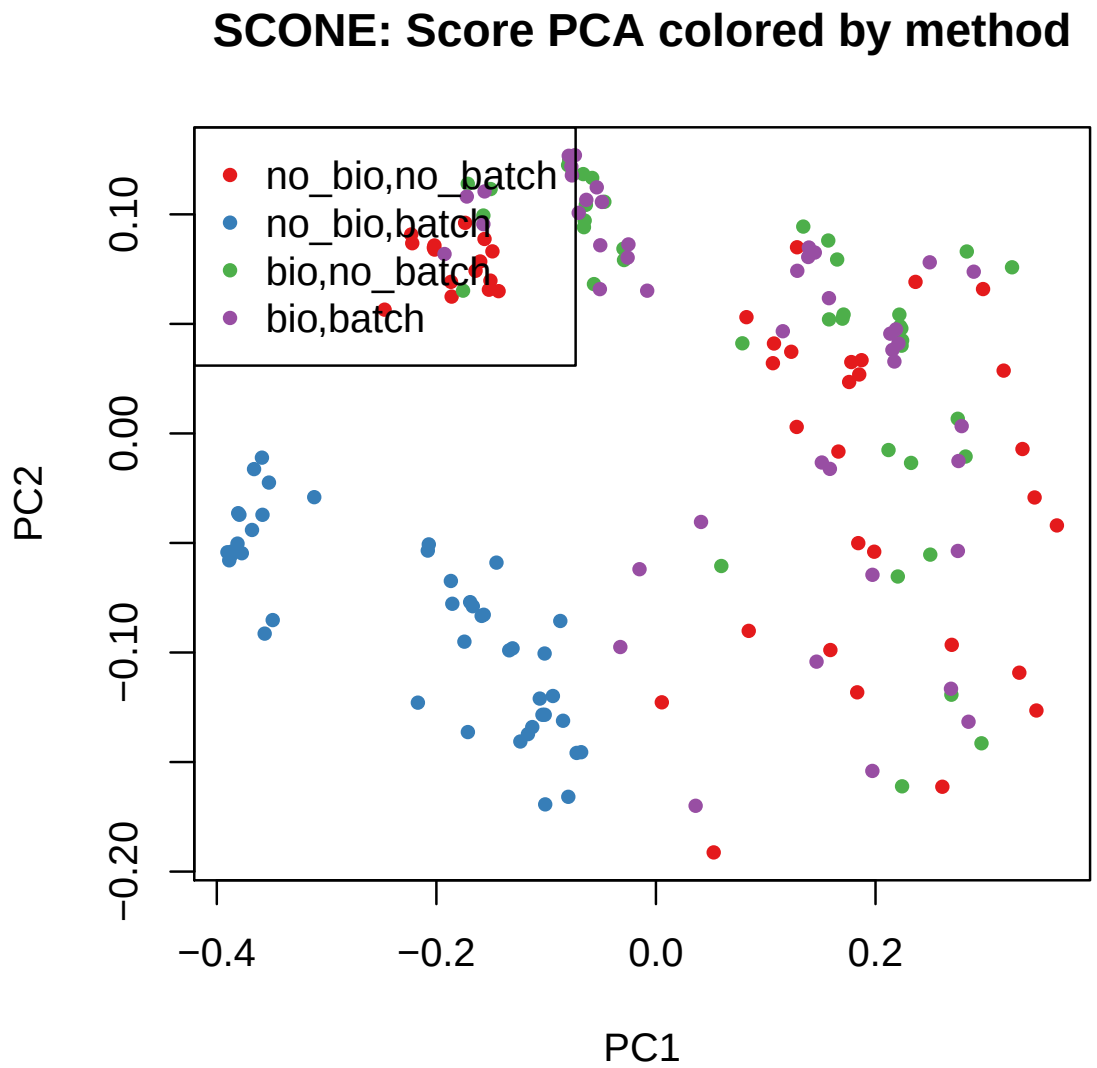
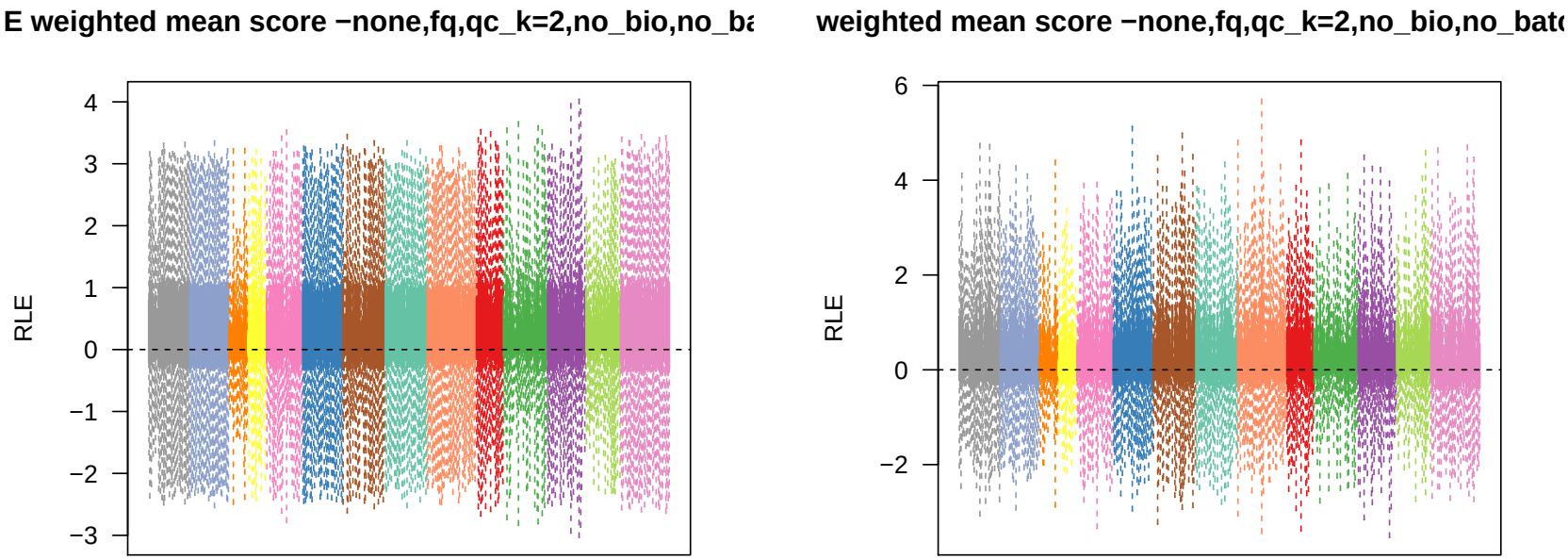


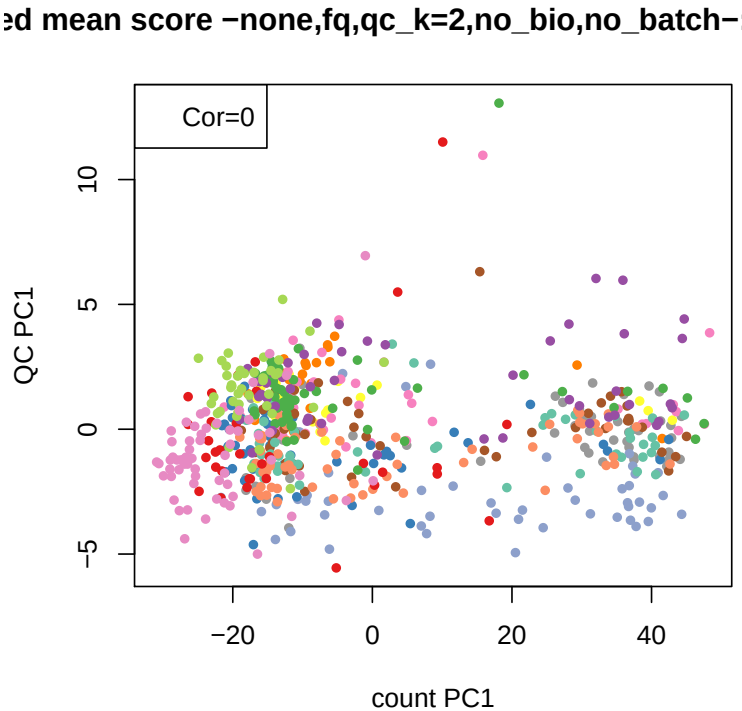
Figure 18: *scone.* PCA of performance scores, colored by method – `adjust_biology`, `adjust_batch`.



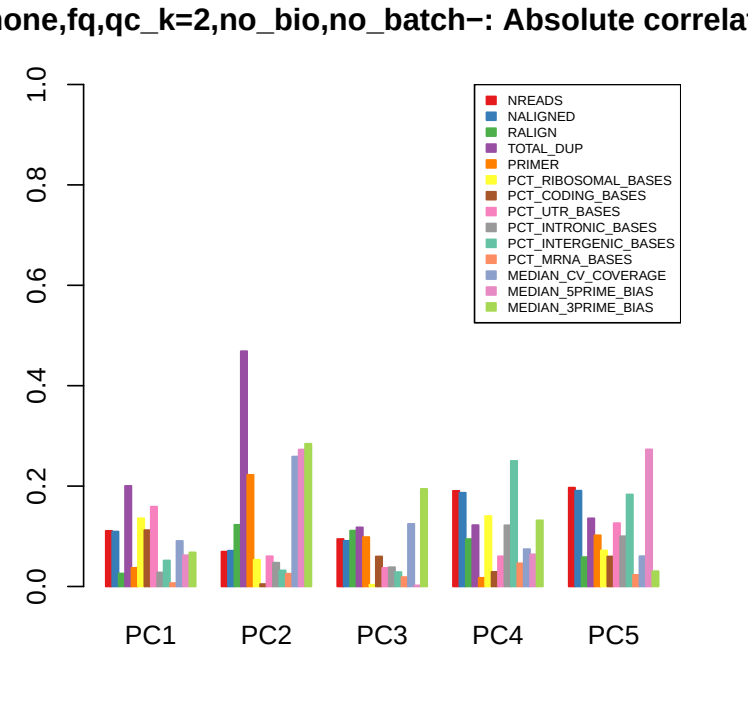
(a) All genes

(b) Housekeeping genes

Figure 19: *scone*. 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`.

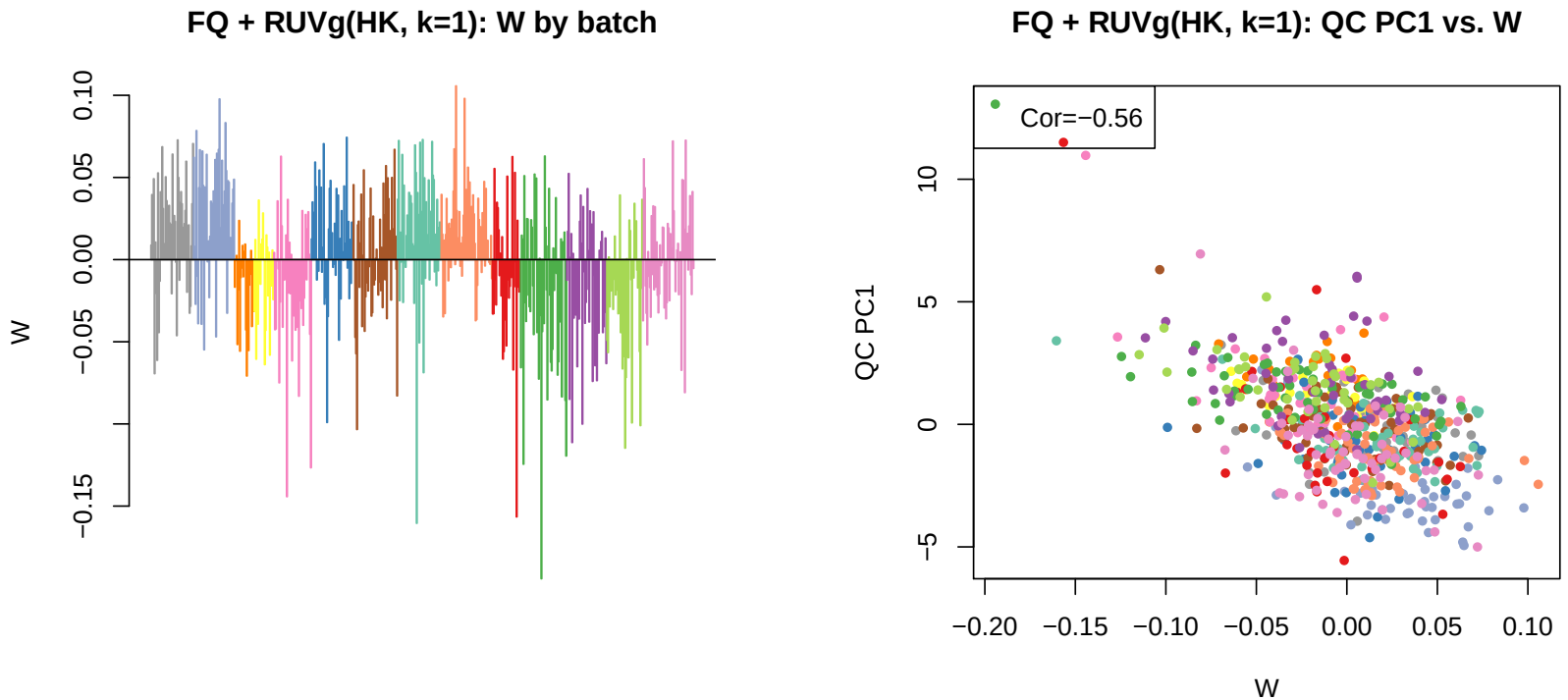


(a) QC PC1 vs. count PC1, color-coded by batch



(b) Correlation of count PC and QC measures

Figure 20: *scone*. Association of counts and sample-level QC measures, `none,fq,qc_k=2,no_bio,no_batch`.



(a) Unwanted factor *W*

(b) QC PC1 vs. *W*

Figure 21: *scone.* Association of RUVg unwanted factor *W* and QC measures, color-coded by batch, `none,fq,ruv_k=1,no_bio,no_batch`.

Normalization: Summary

- Unnormalized gene-level counts exhibit large differences in distributions within and between batches and association with sample-level QC measures.
- Different normalization methods vary in performance according to scone metrics and lead to different distributions of gene-level counts, hence downstream clustering and differential expression results.
- Global-scaling normalization. Not aggressive enough to handle potentially large batch effects and association of counts and QC measures. Biological effects are dominated by nuisance technical effects. Additionally, for DESeq, the scaling factors are computed based on only a handful of genes with non-zero counts in all cells (5/22,054).

Normalization: Summary

- Batch effect normalization. Adjusting for batch effects without properly accounting for the nesting of batch within biological effects (`no_bio,batch`) in the regression model is problematic, as this removes the biological effects of interest (e.g., empirical Bayes framework of ComBat).
- FQ followed by QC-based or RUVg normalization. Seems effective: Similar RLE distributions between samples, lower association of counts and QC measures. The first unwanted factor of RUVg is correlated with the first QC PC.
- Interpretation of performance metrics. Some metrics tend to favor certain methods over others, e.g., EXP_UV_COR (correlation between count PCs and factors of unwanted variation) naturally favors RUVg, especially when the same set of negative controls are used for normalization and evaluation. Hence, a careful, global interpretation of the metrics is recommended.

- Negative controls. The selection of proper, distinct sets of negative controls is important, as these are used for both normalization (RUVg) and assessment of normalization results (EXP_UV_COR).

Expression Quantitation: ZINB-WaVE

Risso et al. (2017). Zero-inflated negative binomial-based wanted variation extraction (ZINB-WaVE).

- General and flexible zero-inflated negative binomial (ZINB) regression framework to account for zero inflation, over-dispersion, and experimental design.
- With a suitable parameterization of the negative binomial (NB) mean and zero inflation (ZI) probability, one can perform normalization (for known and unknown factors of unwanted variation), dimensionality reduction, and differential expression (DE) analysis within the same framework.

Expression Quantitation: ZINB-WaVE

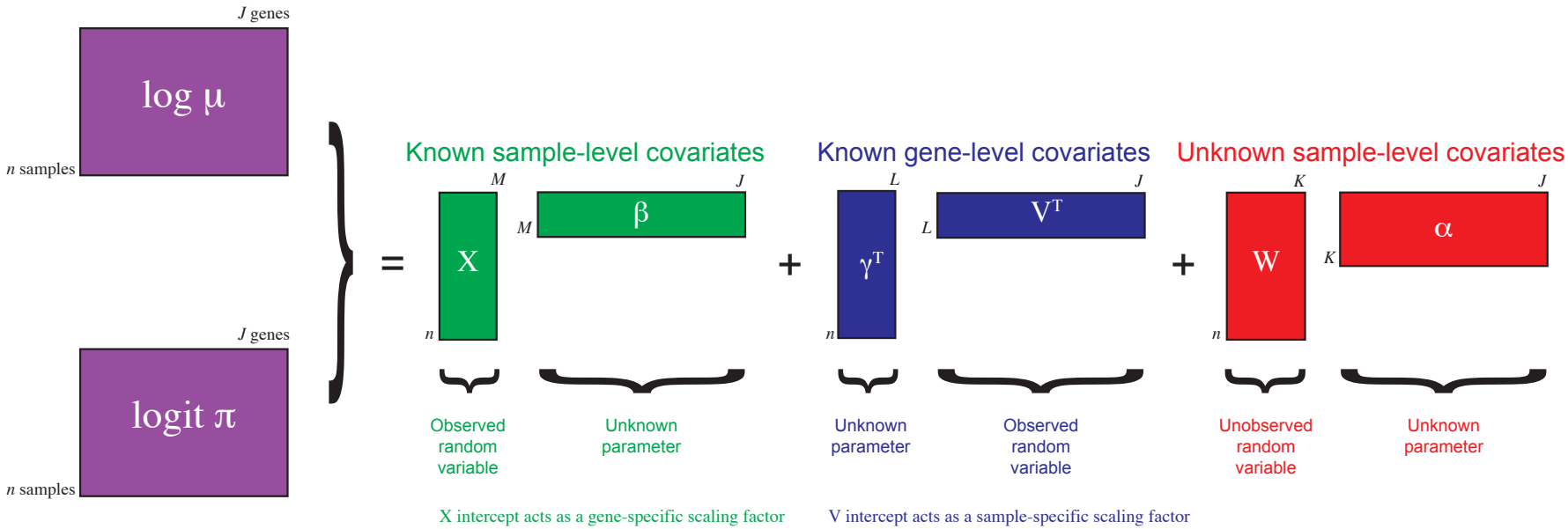


Figure 22: *ZINB-WaVE model.* Parameterization of NB mean and ZI probability. X: Variation of interest (e.g., cell type) or unwanted variation (e.g., batch) – Known. W: Variation of interest (cf. clustering) or unwanted variation (cf. RUV normalization) – Unknown.

ZINB-WaVE framework.

- Experimental design. Accommodate different experimental designs through X, e.g., use of contrasts for nesting of batch effects within biological effects.
- Normalization. Supervised and unsupervised normalization for gene and sample-specific unwanted effects through matrices X, V, and W (cf. RUV of Risso et al. (2014a,b)).
- Dimensionality reduction. Achieved through W. Useful for visualization and input to clustering procedures.
- Differential expression. (Very simple) Use weights based on ZINB-WaVE posterior probabilities to "unlock" standard bulk RNA-Seq methods.

Expression Quantitation: ZINB-WaVE

- Differential expression. (More complex) Likelihood-based tests of DE for both the negative binomial mean and the zero inflation probability, reflecting, respectively, a continuum in DE and on/off DE patterns.
- Bioconductor R package zinbwave: `www.bioconductor.org/packages/release/bioc/html/zinbwav`

Van den Berge et al. (2018). Using ZINB-WaVE weights to unlock bulk RNA-Seq DE methods.

- The posterior probability that a zero count arises from the NB count component is given by Bayes' rule

$$W_{ij} = \Pr(Z_{ij} = 0 | Y_{ij} = 0) = \frac{(1 - \pi_{ij}) f_{NB}(0; \mu_{ij}, \phi_j)}{f_{ZINB}(0; \mu_{ij}, \phi_j, \pi_{ij})}. \quad (1)$$

- We propose to use the ZINB-WaVE posterior probabilities as weights in standard bulk RNA-Seq DE analysis methods, such as those implemented in the Bioconductor R packages edgeR, DESeq2, and limma (limma-voom method with `voom` function).
- Standard bulk DE methods are based on the methodology of generalized linear models (GLM), which readily accommodates inference based on observation-level weights.

Expression Quantitation: ZINB-WaVE

- We have found, on real and simulated data, that using ZINB-WaVE weights in bulk RNA-Seq DE methods outperforms both state-of-the-art scRNA-Seq methods as well as unweighted bulk RNA-Seq methods.

Expression Quantitation: ZINB-WaVE

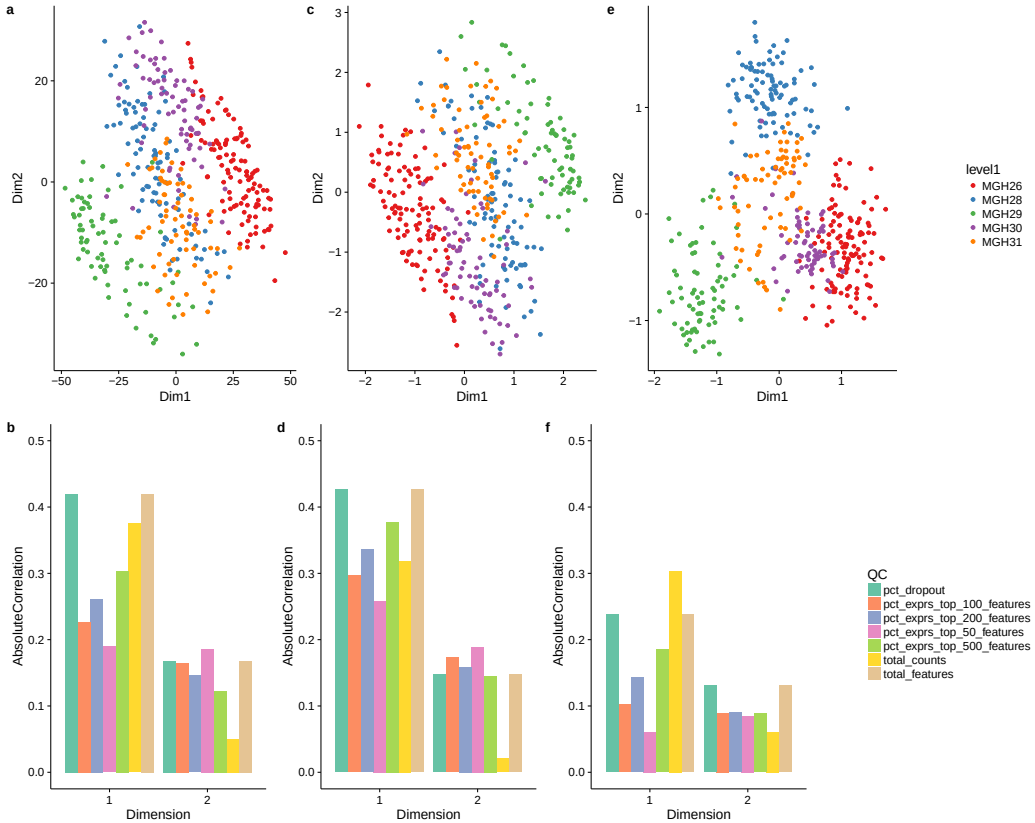


Figure 23: *zinbwave: Dimensionality reduction, Glioblastoma dataset.* **(a, b)** PCA (on TC-normalized counts). **(c, d)** ZIFA (on TC-normalized counts). **(e, f)** ZINB-WaVE W (no normalization needed). Cells color-coded by patient.

Expression Quantitation: ZINB-WaVE

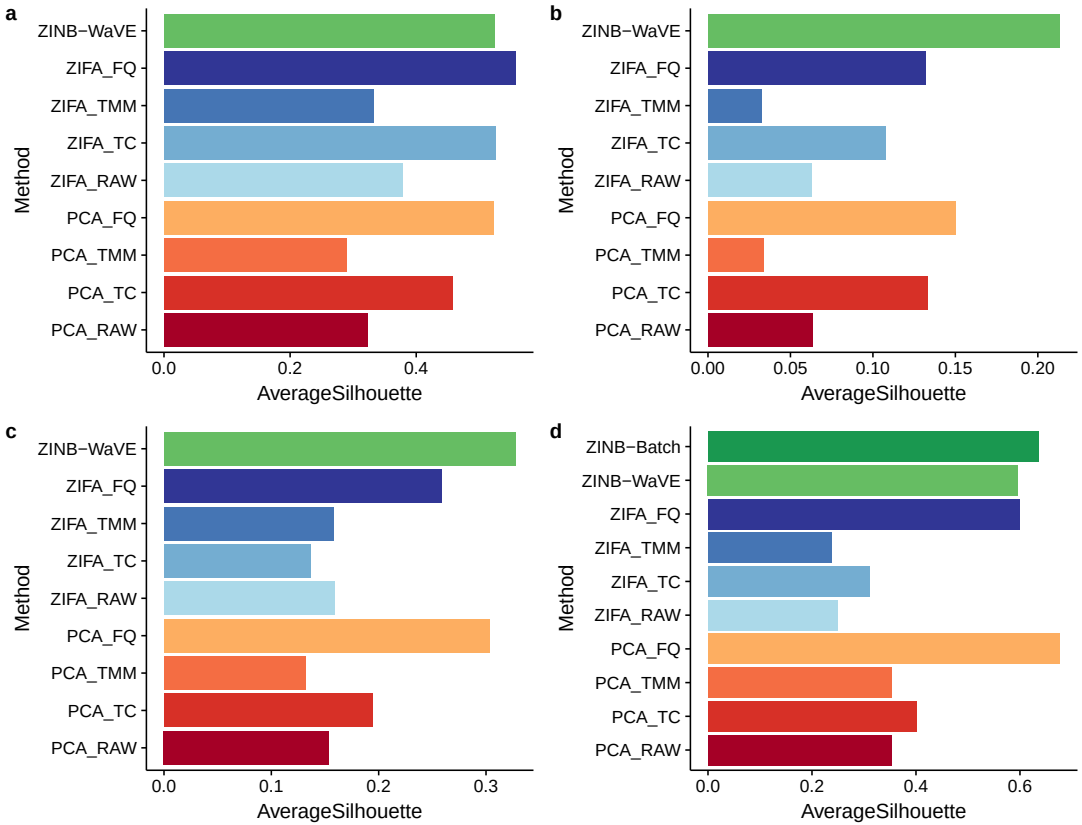


Figure 24: *zinbwave: Dimensionality reduction, real scRNA-Seq datasets.* Average silhouette width for biological groups. **(a)** V1 dataset. **(b)** S1/CA1 dataset. **(c)** Glioblastoma dataset. **(d)** mESC dataset.

Expression Quantitation: ZINB-WaVE

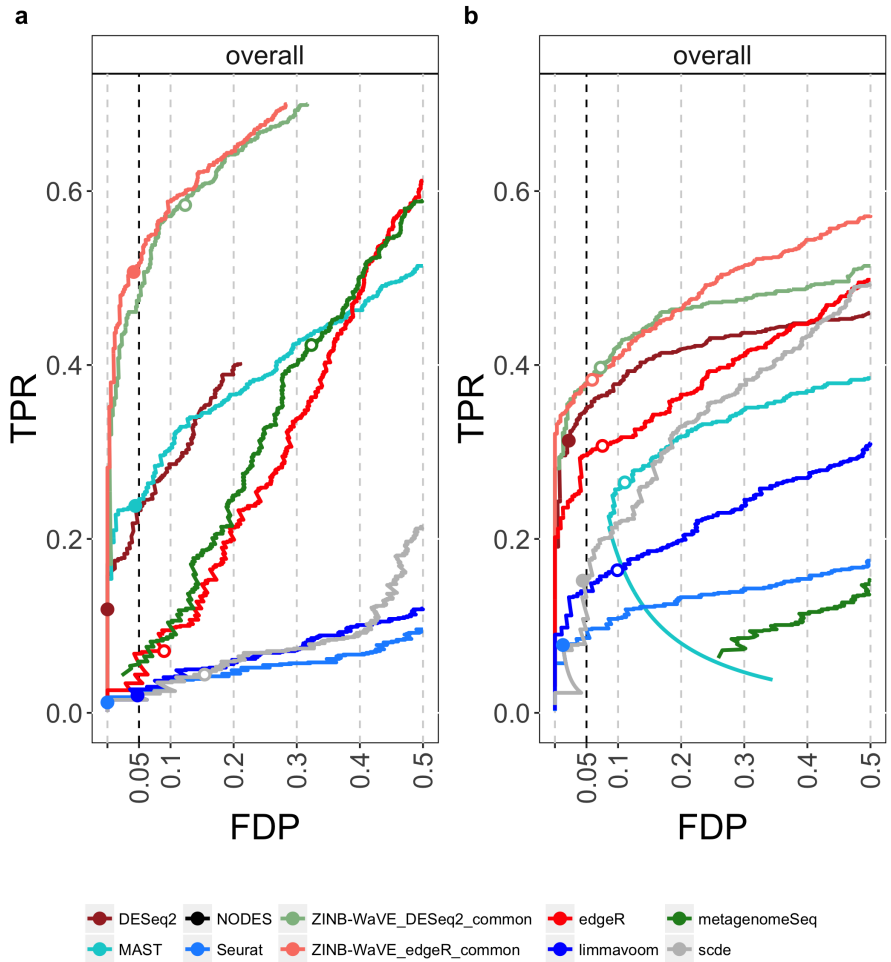


Figure 25: *zinbwave: Comparison of DE methods, simulated scRNA-Seq data.* Scatterplots of true positive rate (TPR) vs. false discovery proportion (FDP). Simulation based on **(a)** Islam et al. (2011) ($n = 90$) and **(b)** Trapnell et al. (2013) ($n = 150$) datasets.

- Robustness to choice of samples. Both hierarchical and partitioning methods tend to be sensitive to the choice of samples to be clustered. Outlying cells/clusters of cells (e.g., glia) are common in scRNA-Seq and mask interesting substructure in the data, often requiring the successive pruning out of dominating clusters to get to the finer structure.
- Robustness to pre-processing and clustering algorithm. Clustering results can be sensitive to pre-processing steps such as normalization and dimensionality reduction, as well as to the choice of clustering algorithm and associated tuning parameters (e.g., distance function, number of clusters).

- Not focusing on the number of clusters. A major tuning parameter of partitioning methods, such as partitioning around medoids (PAM) and k-means, is the number of clusters K. Methods for selecting K (e.g., silhouette width) can be sensitive to the choice of samples, normalization, and other tuning parameters. They tend to be conservative (low K) in scRNA-Seq applications, i.e., capture only the coarse clustering structure and mask interesting substructure in the data. Importantly, the number of clusters K is often not of primary interest.
  E.g. PAM and silhouette width criterion identify only $K = 2$ clusters for the OE p63 dataset.

Cluster Analysis: Motivation

- Not forcing samples into clusters. Some samples may be outliers, that do not really belong to any clusters. Leaving them out can improve the quality and interpretability of the clustering as well as downstream analyses (e.g., identification of cluster marker genes).
- Gene expression signatures for clusters. Common differential expression statistics are not well-suited for finding marker genes for the clusters, especially for the finer structure in a hierarchy.
- Goal. Provide a general and flexible framework that enables researchers to easily apply and compare a variety of different clustering algorithms and associated tuning parameters (e.g., k-means with a range of values for the number of clusters) and generate a stable consensus clustering from these many candidate clusterings.

- We have developed a resampling-based sequential ensemble clustering (RSEC) approach, with the aim of obtaining stable and tight clusters.
- Ensemble clustering, i.e., aggregating multiple clusterings obtained from different algorithms or applications of a given algorithm to resampled versions of the learning set, is a general approach for improving stability. This can be viewed as the unsupervised analog of ensemble methods in supervised learning, e.g., bagging, boosting, random forests.
- Bioconductor R package clusterExperiment: `www.bioconductor.org/packages/release/bioc/html/cluster`

Resampling-based sequential ensemble clustering.

- Given a base clustering algorithm (e.g., PAM, k-means) and associated tuning parameters (e.g., number of principal components, number of clusters, distance matrix), generate a single candidate clustering, optionally using
  - resampling-based clustering and/or
  - sequential clustering (Tseng and Wong, 2005).
- Generate a collection of candidate clusterings by repeating the above procedure for different base clustering algorithms and tuning parameters.
- Identify a single consensus clustering over the different candidate clusterings based on the co-clustering matrix.
- Merge non-differential clusters by creating a hierarchy of clusters, working up the tree, testing for differential expression between sister nodes, and collapsing nodes with insufficient DE.

- Find cluster signatures by testing for differential expression between selected subsets of clusters. E.g. Contrasts based on hierarchy of clusters.
- Visualization. Comparison of multiple clusterings of the same samples, pseudo-color images of co-clustering matrices, pseudo-color images with hierarchical clustering of genes and/or samples.

Gene expression signatures for clusters.

- Find cluster marker genes by testing for differential expression between selected subsets of clusters.
- Standard F-statistic. Tests for any difference between clusters. Sensitive to outlying samples/clusters. Non-specific, i.e., not useful for interpreting differences between clusters.
- Clustering-specific contrasts. Standard solution is to consider contrasts between groups of clusters. By relying on the (generalized) linear model machinery, we use all of the samples in testing these contrasts, rather than just those samples involved in the corresponding clusters.
  - All pairwise. All pairwise comparisons between clusters.
  - One against all. Compare each cluster to union of remaining clusters.

  - Dendrogram. Create a hierarchy of clusters, work up the tree, test for DE between sister nodes (as in approach used for merging clusters).

- For each contrast, one can then test for DE using a variety of methods.
  E.g. Negative binomial GLM in edgeR and DESeq (with ZINB-WaVE weights), limma linear model with `voom` weights to account for over-dispersion.

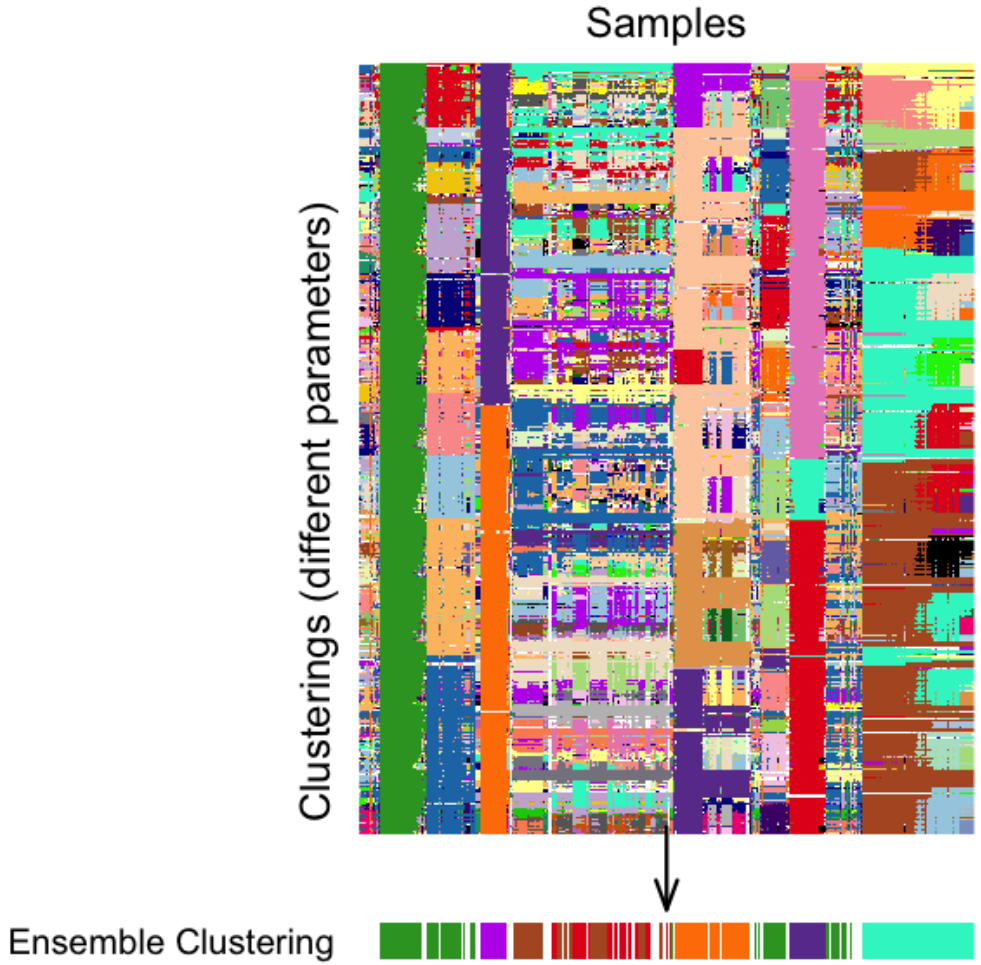


Figure 26: *clusterExperiment*. Comparison of `clusterMany` and `combineMany` clusterings (`plotClusters`).

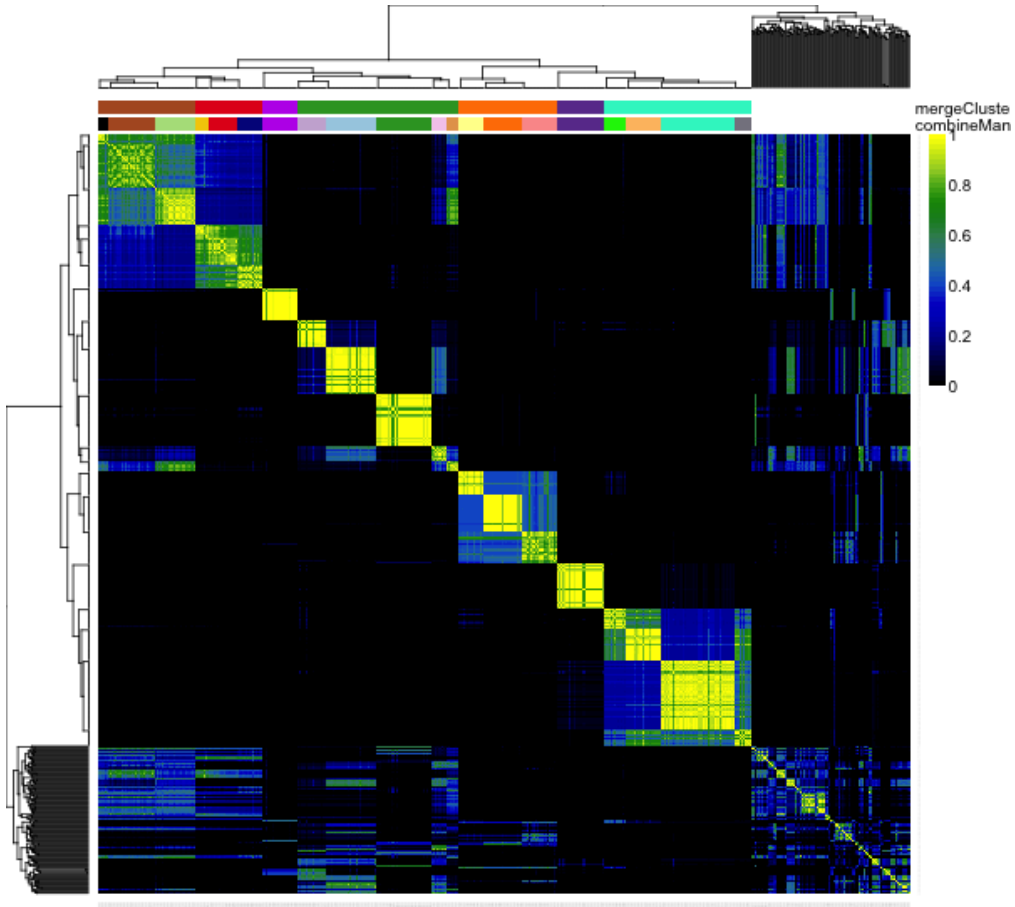


Figure 27: *clusterExperiment*. Pseudo-color image of co-clustering matrix for `clusterMany` clusterings, used to create `combineMany` clustering (`plotHeatmap`).

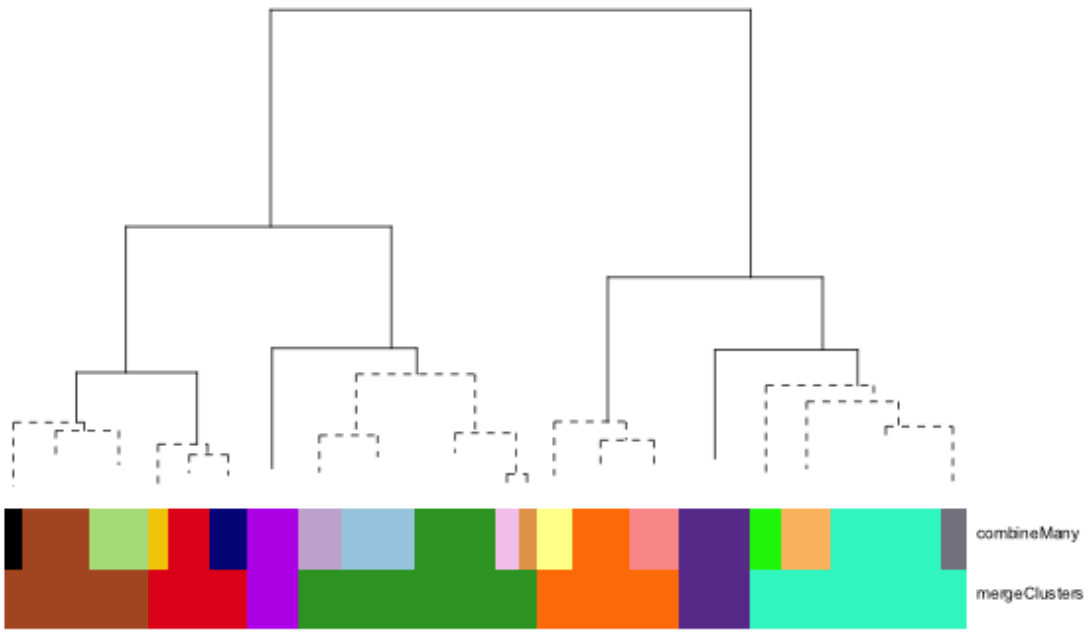


Figure 28: *clusterExperiment*. Dendrogram of `combineMany` clusters (`makeDengrogram`). Similar clusters in sister nodes are merged using `mergeClusters` (dashed lines).

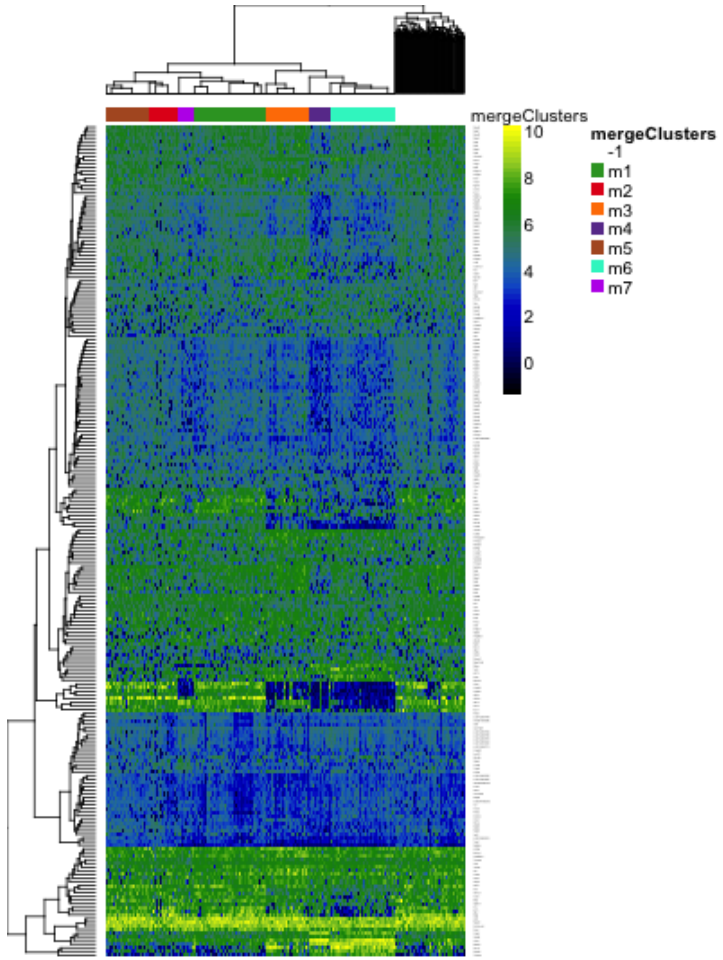


Figure 29: *clusterExperiment*. Pseudo-color image of log-counts for DE genes based on F-statistics (top 472; `getBestFeatures`).

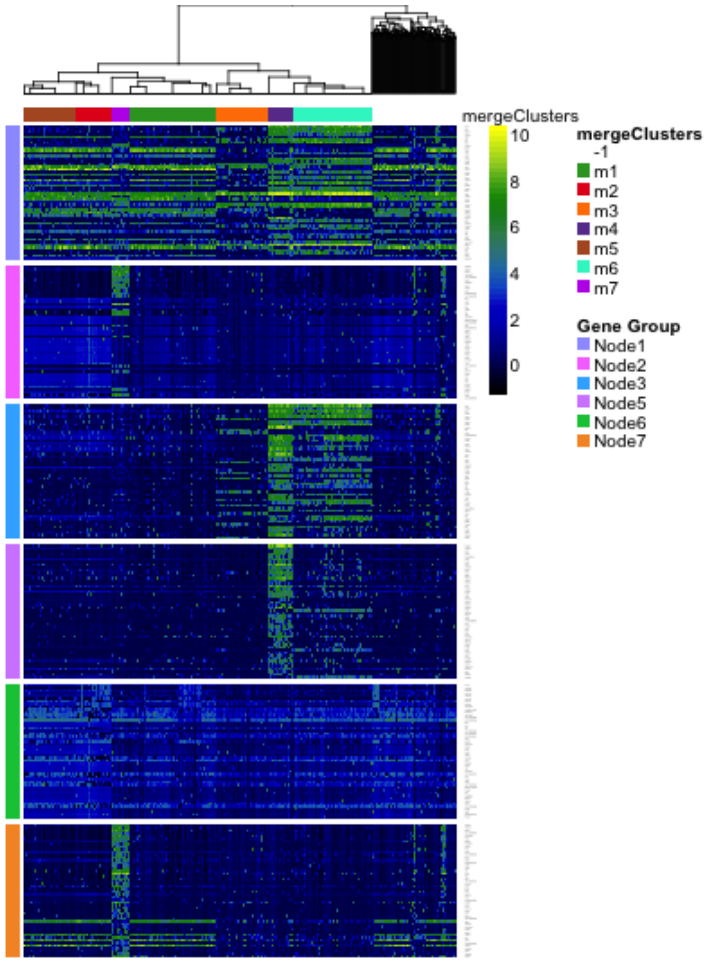


Figure 30: *clusterExperiment*. Pseudo-color image of log-counts for DE genes using dendrogram contrasts (472 genes, top 50 genes for each node; `getBestFeatures`).

Application to OE p63 dataset. The clusters identified by RSEC correspond to a variety of states in the differentiation process of stem cells into neurons 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

- Mapping transcriptional progression from stem cells to specialized cell types is essential for properly understanding the mechanisms regulating cell and tissue differentiation.
- Often, there is no clear distinction between states, but rather a smooth transition, with individual cells undergoing gradual transcriptional changes and existing on a continuum between states.
- The relationship between states can be represented as a continuous lineage dependent upon an underlying spatial or temporal variable. This representation, referred to as pseudotemporal ordering, can help us understand how cells differentiate and how cell fate decisions are made.
- We have developed slingshot as a flexible and robust framework for inferring cell lineages and pseudotimes.

Inference of Cell Lineages and Pseudotimes: Slingshot

- Slingshot allows the identification of any number of novel lineages, with the option of incorporating subject-matter knowledge to supervise parts of the inference process (e.g., known terminal states).
- R package slingshot: `github.com/kstreet13/slingshot`.

Inference of Cell Lineages and Pseudotimes: Slingshot

- Input/Output.
  - Input. Reduced-dimensional representation of normalized gene expression measures; cell clustering.
  - Output. Cell lineages, i.e., ordered subsets of cell clusters. Cell pseudotimes (and weights) for each lineage.
- Semi-supervised inference of global lineage structure.
  - Minimum spanning tree (MST) over cell clusters, with between-cluster distance based on covariance-scaled Euclidean distance between cluster means.
  - Root and leaf nodes. Either pre-specified or automatically selected.
    Root node. If not pre-specified, selected based on parsimony (i.e., set of lineages with maximal number of clusters shared between them).
    Leaf nodes. If pre-specified, constrained MST.

Inference of Cell Lineages and Pseudotimes: Slingshot

  - Outlying clusters. Identified using granularity parameter ω that limits maximum edge weight in the tree. Specifically, build MST using an artificial cluster Ω, with distance ω from other clusters (a fraction of maximum pairwise distance between clusters), and then remove Ω.
  - A lineage is then defined as any unique path coming out of the root node and ending in a leaf node.
  - Constructing the MST on clusters (Ji and Ji, 2016; Shin et al., 2015, TSCAN,Waterfall) vs. cells (Trapnell et al., 2014, Monocle) offers greater stability and computational efficiency, less complex lineages, and easier determination of directionality and branching.
- Inferring cell pseudotimes along each lineage.
  - Simultaneous principal curves. Iterative procedure based on principal curves algorithm of Hastie and Stuetzle (1989) (`principal.curve` function in `princurve` package).

Inference of Cell Lineages and Pseudotimes: Slingshot

  - In the case of branching lineages, a shrinkage step is included at each iteration, shifting the curves to a consensus path in areas where lineages share many common cells, but allowing the curves to separate as lineages share fewer and fewer cells. This ensures smooth bifurcations of the paths.
  - Pseudotimes are derived by orthogonal projection onto the curves.
  - Cells belonging to clusters that are included in multiple lineages have multiple, similar pseudotimes.
  - We find that principal curves provide discerning power not found in piecewise linear trajectories, while also adding stability over a range of dimensionality reduction and clustering methods.
- Slingshot is robust to pre-processing steps implicit in virtually all lineage inference methods: normalization, dimensionality reduction, clustering.

Inference of Cell Lineages and Pseudotimes: Slingshot

- Differential gene expression within and between lineages. Regression of gene expression measures on pseudotime, e.g., generalized additive models (GAM) (Ji and Ji, 2016, TSCAN); dynamic time warping (DTW).
- Visualization. Two- and three-dimensional plots of cell lineages and pseudotimes, gene-level trajectories, pseudo-color images for DE genes.

Inference of Cell Lineages and Pseudotimes: Slingshot

Application to OE p63 dataset.

- Cell clusters. We use the RSEC clustering to define intermediate states in the differentiation of HBC to neuronal and sustentacular cells.
- Leaf-node supervision. Known terminal states: Mature sustentacular cells (mSus), microvillous cells (MV), and mature olfactory sensory neurons (mOSN) (only the first had an effect).
- Slingshot identifies three lineages: HBC–mSus, HBC–GBC–MV, HBC–GBC–mOSN.
  In the first lineage, sustentacular cells are produced via direct conversion of HCB (without cell division). By contrast, in the second and third lineages, microvillous and neuronal cells are produced via an intermediate, proliferative state (GBC).

Inference of Cell Lineages and Pseudotimes: Slingshot

- Without leaf-node supervision, we draw the (known) false conclusion that sustentacular cells may develop into GBC.
- Slingshot and subsequent DE analysis along lineages reveal 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 Cell Lineages and Pseudotimes: Slingshot

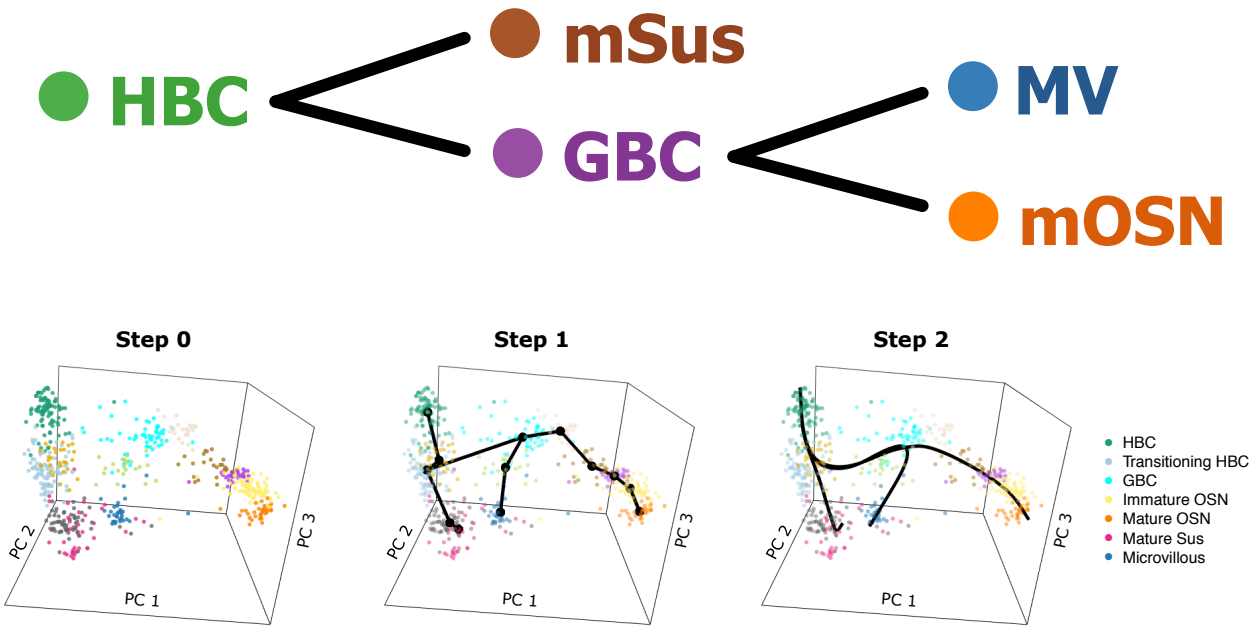


Figure 31: *slingshot*. MST on cell clusters and simultaneous principal curves. Slingshot identifies three lineages: HBC–mSus, HBC–GBC–MV, HBC–GBC–mOSN.

Inference of Cell Lineages and Pseudotimes: Slingshot

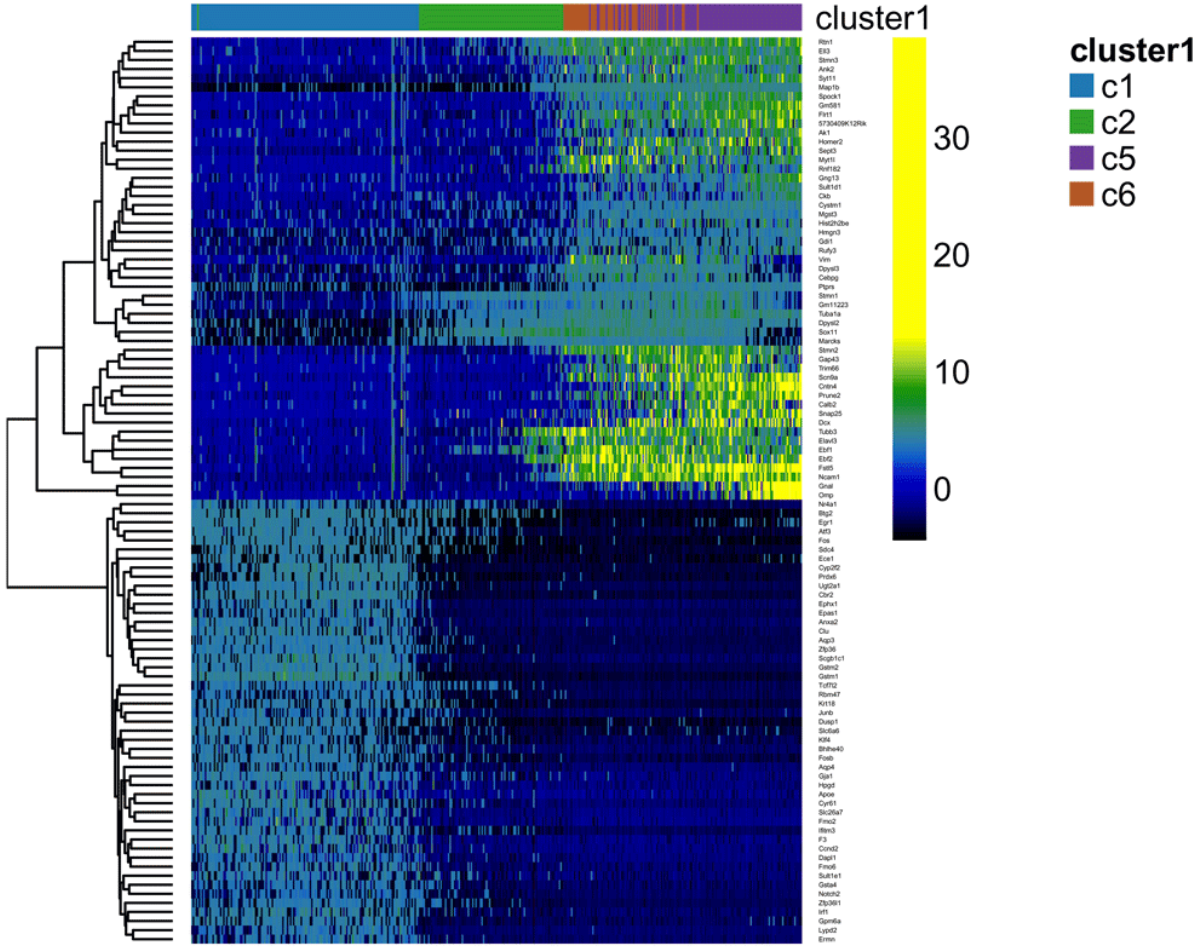


Figure 32: *slingshot*. Pseudo-color image of top 100 DE genes for neuronal lineage (GAM).

Discussion

- Experimental design. Applying basic design principles could easily avoid glaring problems such as confounding of biological effects of interest with nuisance technical effects.
- Validation. Need control genes and samples.
- Computational reproducibility. Still far from being the norm. And confusion with biological reproducibility.
- Notes to self.
  - Look at data: Garbage in/garbage out.
  - Don't reinvent the wheel.
  - "Why do it the easy way when you can do it the hard way?"
  - Beware of hammers looking for nails: Start from the biological question and data.
  - Beware of models and assumptions: Results should be driven by data, not models.
  - Interpretation of probabilities: Source of randomness? Sampling units?

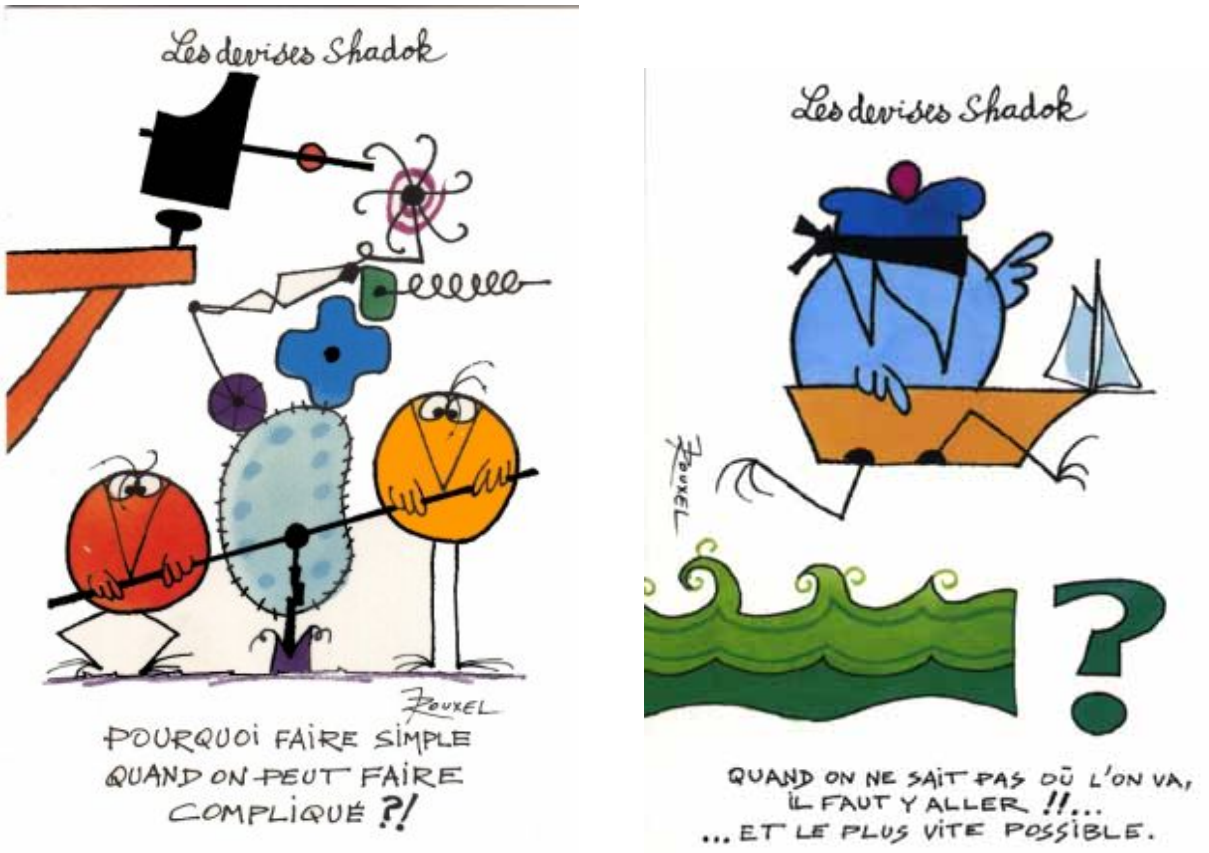


Figure 33: *Les devises Shadok.*

- 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`
- 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.`
- scone: Normalization procedures and performance assessment.
  `www.bioconductor.org/packages/release/bioc/html/scone.h`
- slingshot: Cell lineage and pseudotime inference.
  `github.com/kstreet13/slingshot`.

Software

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

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

References

S. Anders and W. Huber. Differential expression analysis for sequence count data. Genome Biology, 11(10):R106, 2010.

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).

R. A. Fisher. Statistical methods in genetics. The Bateson Lecture, 1951. Heredity, 6:1–12, 1952.

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: 10.1016/j.stem.2017.04.003.

T. Hastie and W. Stuetzle. Principal curves. Journal of the American Statistical Association, 84(406):502–516, 1989.

References

S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J. B. Fan, P. Lönnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research, 21(7):1160–1167, 2011.

Z. Ji and H. Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Research, 2016.

G. Mendel. Versuche über pflanzen-hybriden (experiments in plant hybridization). In Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn 4, pages 3–47, 1866. URL `www.mendelweb.org`.

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: 10.12688/f1000research.12122.1.

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`.

References

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.

D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert. ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. bioRxiv, April 2017. doi: 10.1101/125112. URL http://biorxiv.org/content/early/2017/04/06/125112.

J. Shin, D. A. Berg, Y. Zhu, J. Y. Shin, J. Song, M. A. Bonaguidi, G. Enikolopov, D. W. Nauen, K. M. Christian, G. Ming, and H. Song. Single-cell RNA-Seq with Waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell, 17(3):360–372, 2015.

C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology, 31(1):46–53, 2013.

References

C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 4(32):381–391, 2014.

G. C. Tseng and W. H. Wong. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. Biometrics, 61(1):10–16, 2005.

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: 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, 2018. Accepted.