Analysis of single-cell RNA-seq data

Hao Wu
Department of Biostatistics and Bioinformatics
Emory University

CREiGS short course
October 2020

Outline

- **Background**
- **Data processing**
  - Preprocessing and data characteristics
  - Normalization
  - Batch effect correction
  - Imputation
- **Data analyses**
  - Cell clustering
  - Pseudo-time construction
  - Cell type identification
  - Differential expression
- **Data visualization**
  - TSNE and UMAP

Background

- Most of the biological experiments are performed on “bulk” samples, which contains a large number of cells (millions).
- The “bulk” data measure the average signals (gene expression, TF binding, methylation, etc.) of many cells.
- The bulk measurement ignores the inter-cellular heterogeneities:
  - Different cell types.
  - Variation among the same cell type.

Single cell biology

- The study of individual cells.
- The cells are isolated from multi-cellular organism.
- Experiment is performed for each cell individually.
- Provides more detailed, higher resolution information.
- High-throughput experiments on single cell is possible.

Single cell sequencing

- Different types of sequencing at the single-cell level:
  - DNA-seq
  - ATAC-seq, ChIP-seq
  - BS-seq
  - RNA-seq
- Very active research field in the past few years.

Basic experimental procedure

- Isolation of single cell. Techniques include
  - Laser-capture microdissection (LCM)
  - Fluorescence-activated cell sorting (FACS)
  - Microfluidics
- Open the cell and obtain DNA/mRNA/etc.
- PCR amplification to get enough materials.
- Perform sequencing.
- Note that single cell sequencing usually has higher error rates than bulk data.

Single cell RNA-seq (scRNA-seq)

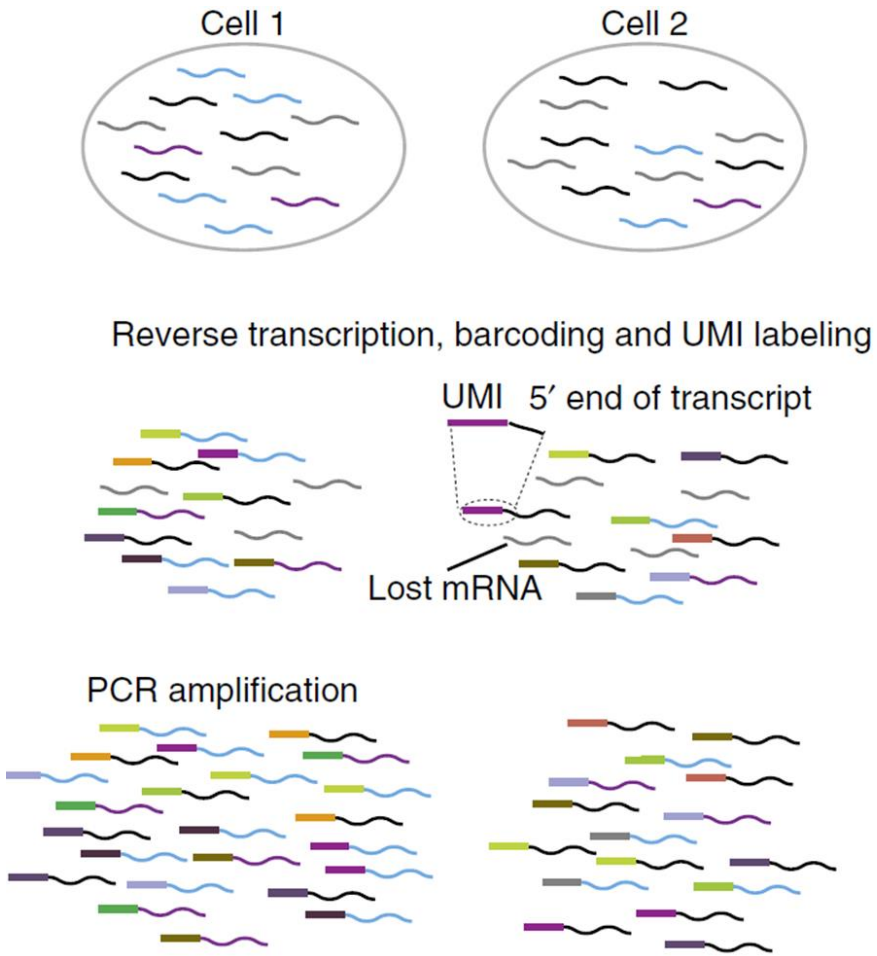
- The most active in the single cell field.
- Scientific goals:
  - Composition of different cell types in complex tissues.
  - New/rare cell type discovery.
  - Gene expression, alternative splicing, allele specific expression at the level of individual cells.
  - Transcriptional dynamics (pseudotime construction).
  - Above can be investigated and compared spatially, temporally, or under different biological condition.

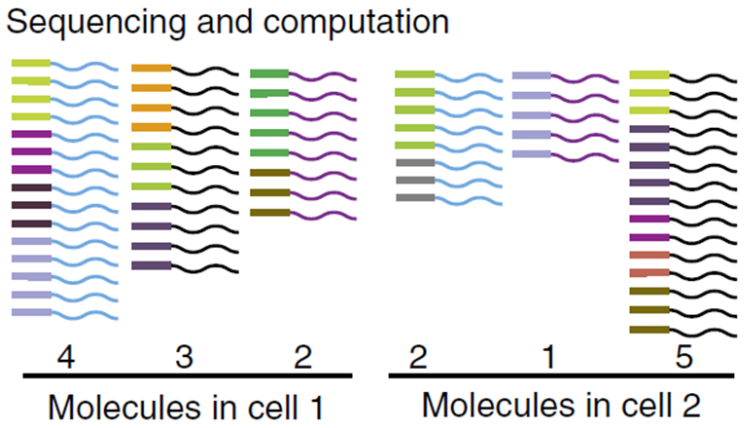
scRNA-seq technologies

- Full-length sequencing, such as Smart-Seq/Smart-Seq2
  - High sequencing depth
  - Better at detecting low expression genes
  - Good for isoform analysis, allele specific expression
- 3’ end sequencing: such as droplet-based (Drop-seq, inDrop, 10x genomics)
  - Many cells, low sequencing depth per cell
  - Good for identifying cell subpopulations

Universal molecular identifier (UMI)

- Short sequence tag added to the mRNA molecular before PCR, for reducing PCR bias.





Saiful Islam ··· Sten Linnarsson

Data processing

- Preprocessing
- Data characteristics
- Normalization
- Batch effect correction
- Imputation

scRNA-seq data preprocessing

- Sequence alignment and expression quantification
  - RNA-seq alignment software (Tophat, STAR, HISAT, etc.) can be used
  - Some commercial software, such as Cell Ranger for 10x genomics data.

Some data characteristics

- Data is very sparse (many zeros), especially for Drop-seq data.
- Number of transcripts detected is much lower compared to bulk RNA-seq under the same sequencing depth.

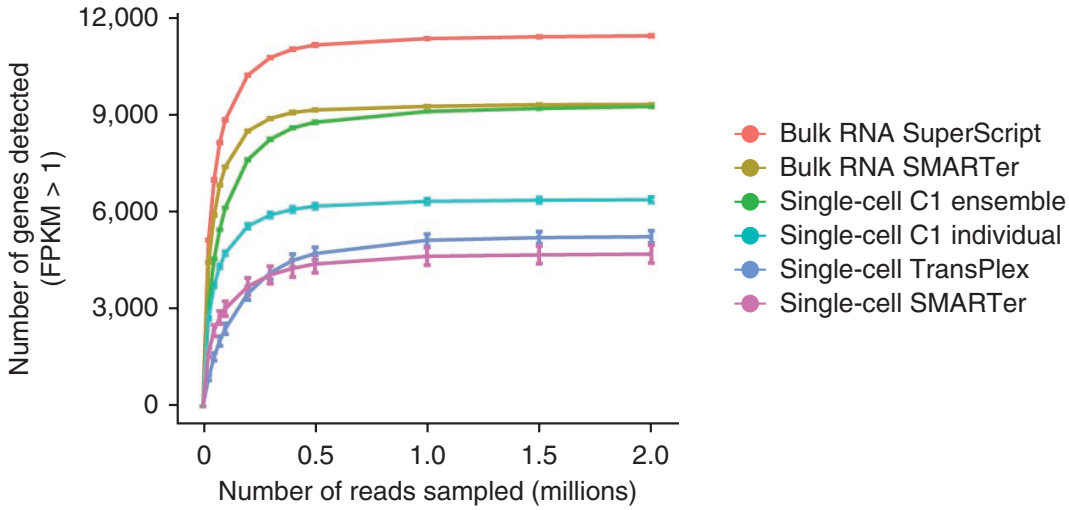
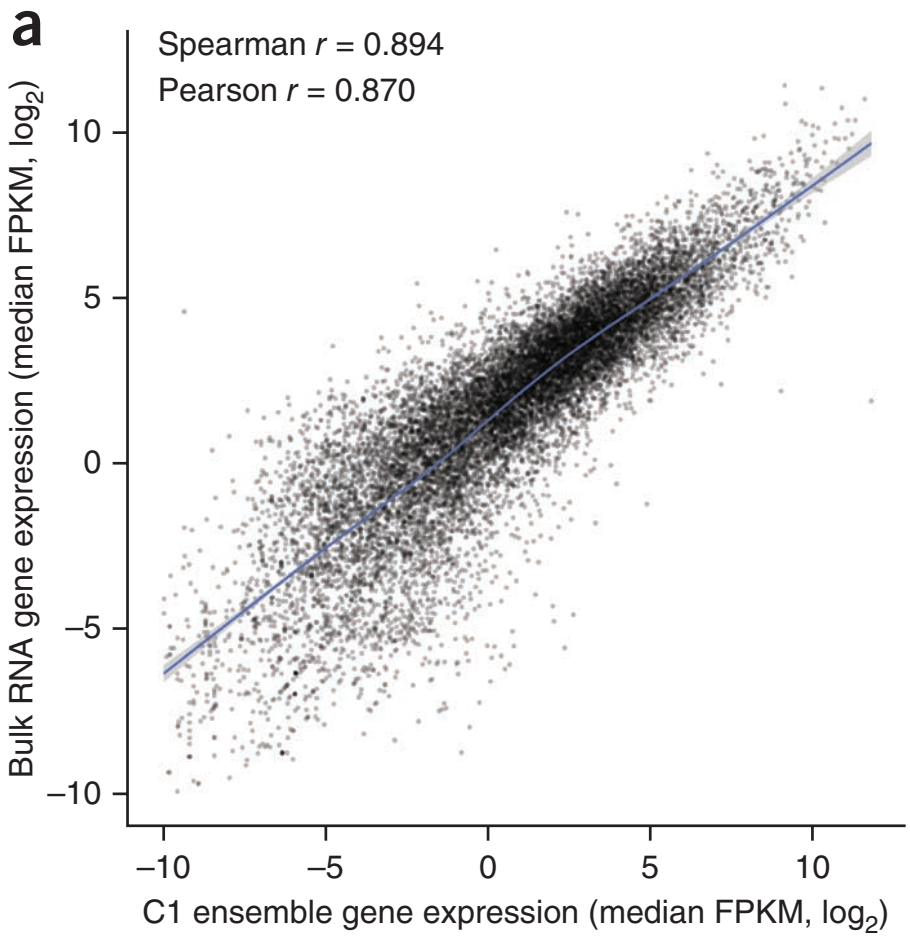


Figure 5 | Saturation curves for the different sample preparation methods. Each point on the curve was generated by randomly selecting a number of raw reads from each sample library and then using the same alignment pipeline to call genes with mean FPKM >1. Each point represents four replicate subsamplings. Error bars, standard error.

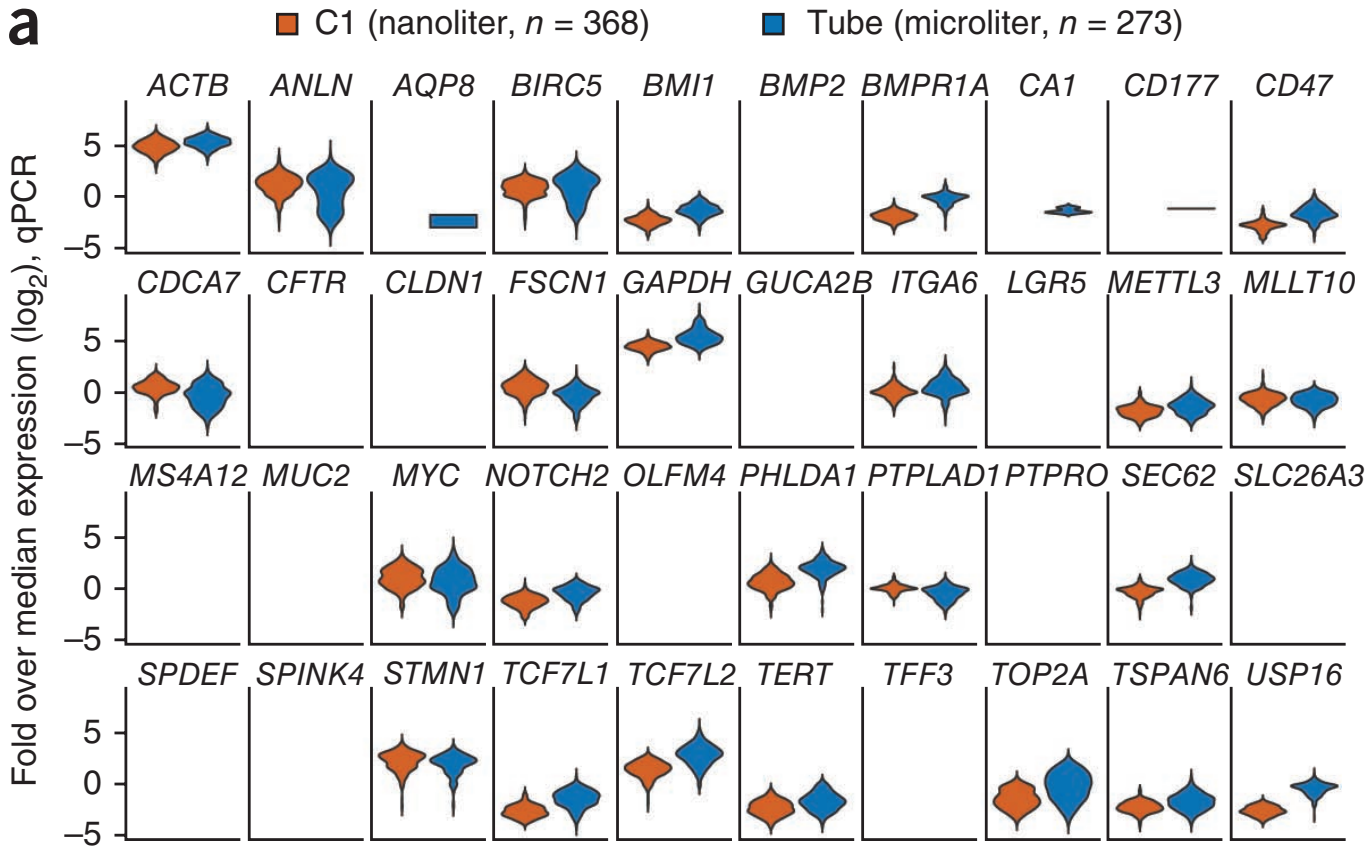
Wu et al. 2013 Nature Method

- Bulk and aggregated single cell expressions have good correlation.



Wu et al. 2013 Nature Method

- Expression levels for a gene in different cells sometimes show bimodal distribution.



Wu et al. 2013 Nature Method

Data normalization

- scRNA-seq is very noisy.
- Spike-in data is usually available.
  - Spike-ins from the external RNA Control Consortium (ERCC) panel contains 92 synthetic spikes based on bacterial genome with known expression level.
- UMI is helpful for removing amplification noise.
- A combination of spike-in and UMI can potentially be used for data normalization.
- Simple normalization (such as by sequencing depth) for bulk RNA-seq can be applied, e.g., TPM or FPKM.

Application Note

Normalization and noise reduction for single cell RNA-seq experiments

Bo Ding[1,#], Lina Zheng[1,#], Yun Zhu[1], Nan Li[1], Haiyang Jia[1,2], Rizi Ai[1], Andre Wildberg[1] and Wei Wang[1,3*]

[1]Department of Chemistry and Biochemistry, University of California, La Jolla, CA 92093, USA,

[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China.

[3]Department of Cellular and Molecular Medicine, University of California, La Jolla, CA 92093, USA,

[#]Equal contribution

Associate Editor: Dr. Ziv Bar-Joseph

- Log-transform FPKM values, denoted by x.
- Assume the expression value, y, follow Gamma distribution. The mean of Gamma is a polynomial function of x: $y = \mu(x)$.

$\mu(x) = \sum_{i=0}^{n} \beta_i x^i$. The model is the following:

$$y \sim Gamma(y; \mu(x), \varphi)$$

- Use MLE to estimate parameters based on ERCC data. Then the fitted model is applied to all genes to estimate concentration.

Lun *et al. Genome Biology* (2016) 17:75
DOI 10.1186/s13059-016-0947-7

Genome Biology

METHOD Open Access

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

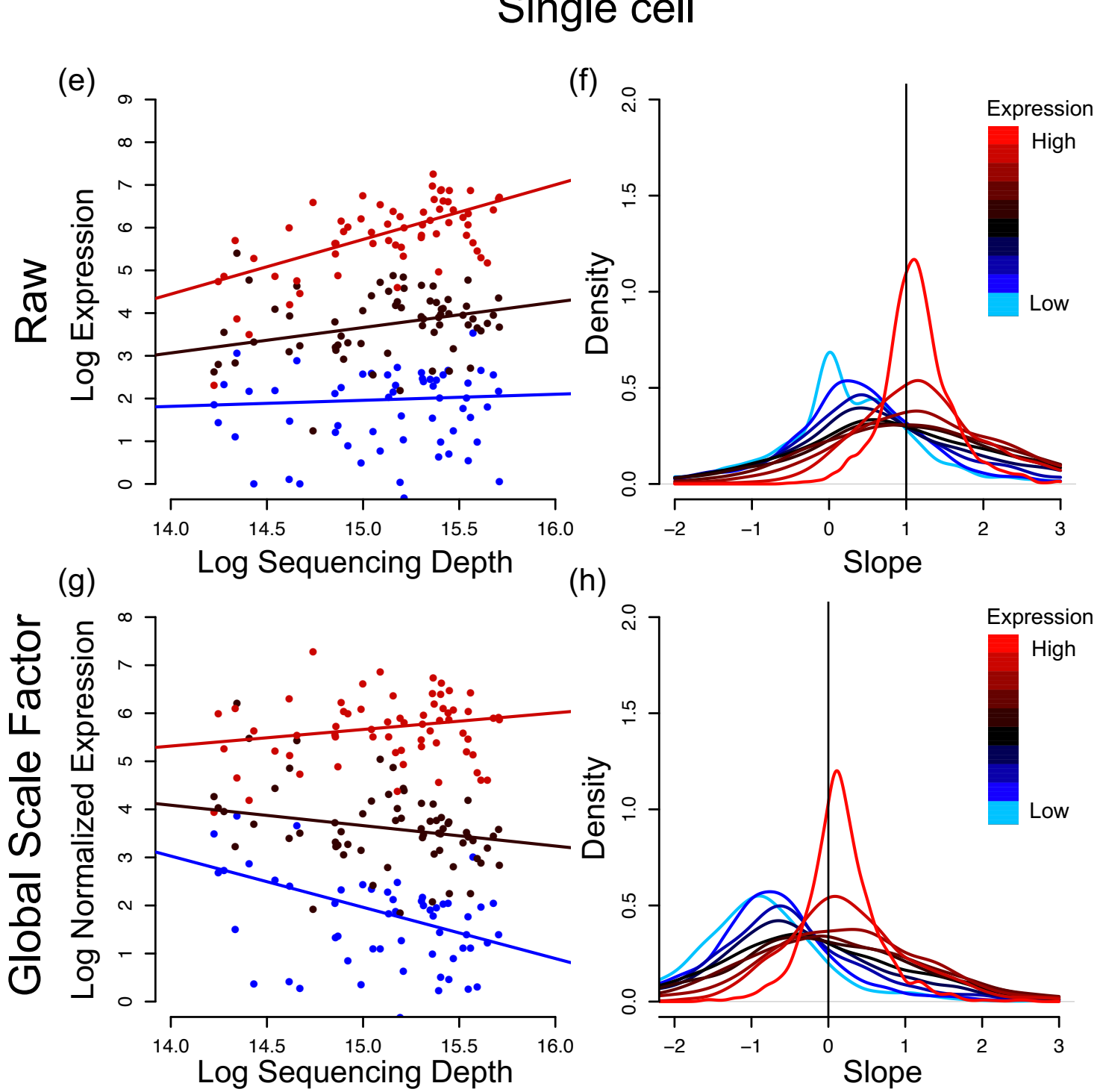
- Works for data without spike-in.
- The goal is to estimate a size factor for each cell.
- The idea is to normalize on summed expression values from pools of cells – it's more stable than using individual cell.
- Bioconductor package **scran.**

SCnorm: robust normalization of single-cell RNA-seq data

584 | VOL.14 NO.6 | JUNE 2017 | NATURE METHODS

Rhonda Bacher[1,5], Li-Fang Chu[2,5], Ning Leng[2], Audrey P Gasch[3], James A Thomson[2], Ron M Stewart[2], Michael Newton[1,4] & Christina Kendziorski[4]

- Basic idea: one normalization factor per cell doesn't fit all genes.
- Relationships of read counts and sequencing depths vary and depend on the expression levels.


Single cell
(e)
(f)
Raw
Log Expression
Log Sequencing Depth
Density
Slope
Expression
High
Low
(g)
(h)
Global Scale Factor
Log Normalized Expression
Log Sequencing Depth
Density
Slope
Expression
High
Low

SCnorm Solution

- Uses quantile regression to estimate the dependence of read counts on sequencing depth for every gene.
- Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group.
- Bioconductor package **SCnorm**.

Batch effect correction

- Batch effect in scRNA-seq can be severe.
- It's difficult to randomize the design, i.e., batch is often confounded with individual, so it causes trouble for analyzing data from multiple individuals (more on this later).
- Bulk data methods such as Combat/SVA don't work well
- There are a number of methods specifically designed for scRNA-seq:
  - MNN (Haghverdi et al. 2018. Nat. Biotech.)
  - ZINB-WaVE (Risso et al. 2018 Nat. comm.)
  - LIGER (Welch et al. 2019. Cell)
  - Harmony (Korsunsky et al. 2019 Nat. Method)
  - BUSseq (Song et al. 2020. Nat. Comm.)

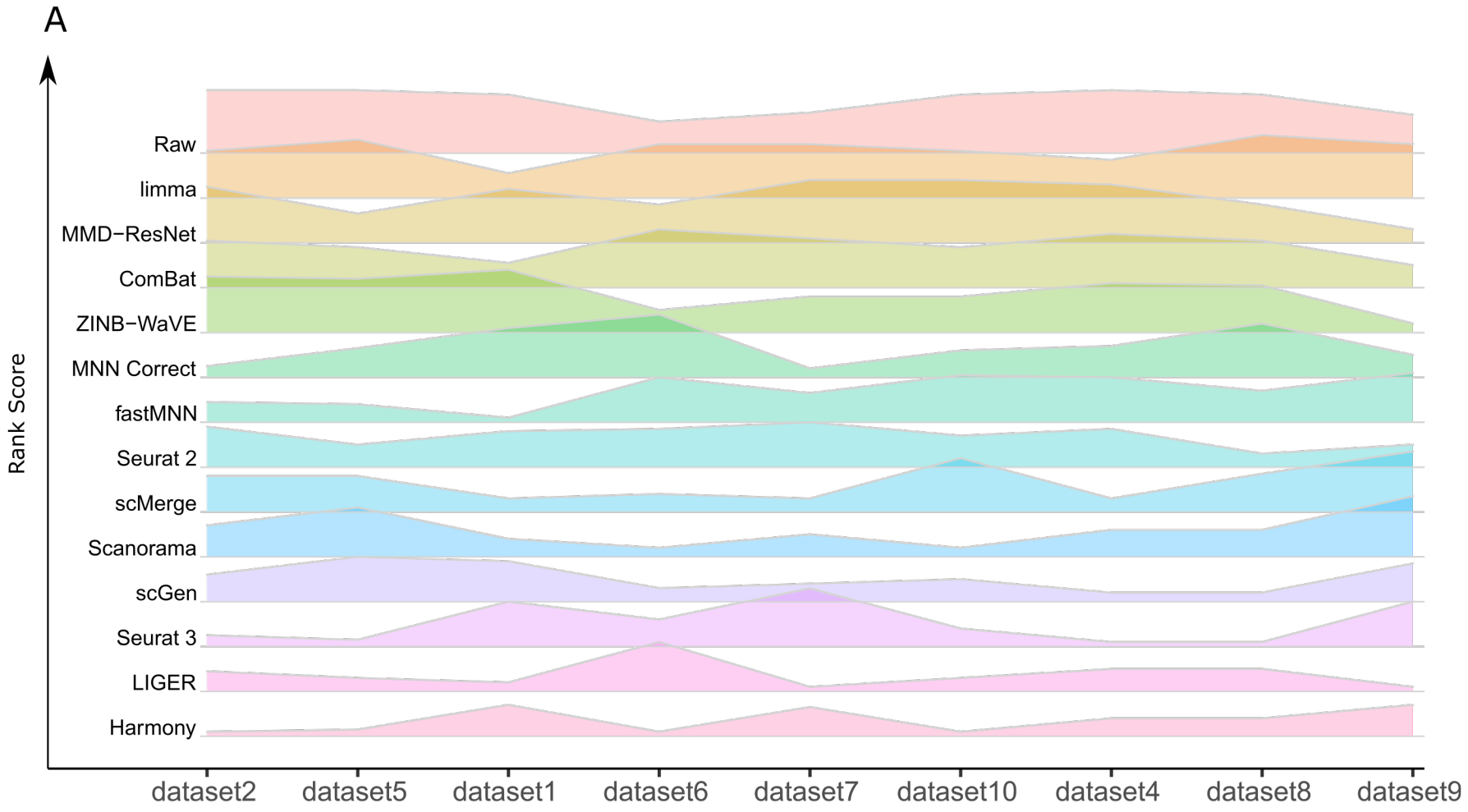
Tran *et al. Genome Biology* (2020) 21:12
https://doi.org/10.1186/s13059-019-1850-9

Genome Biology

RESEARCH Open Access

A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Hoa Thi Nhu Tran†, Kok Siong Ang†, Marion Chevrier†, Xiaomeng Zhang†, Nicole Yee Shin Lee, Michelle Goh and Jinmiao Chen*



Data imputation

- scRNA-seq has lots of missing data (dropout).
- Imputing the missing data help the downstream analyses.
- There are a number of methods:
  - SAVER (Huang et al. 2018 Nat. Methods)
  - ScImpute (Li et al. 2018 Nat. Comm.)
  - MAGIC (van Dijk et al. 2018 Cell)
  - SCRABBLE (Peng et al. 2019 GB)

General strategy for imputation

- The problem is similar to a “recommendation system”.
  - First compute the similarities among genes and cells.
  - To impute one element, borrow information from similar gene/cell.

Data analyses tasks

- Cell clustering
- Pseudotime construction
- Cell type identification
- Differential expression
- Rare cell type discovery
- Alternative splicing
- Allele specific expression
- RNA velocity

Cell clustering

- Perhaps the most active topic in scRNA-seq.
- The goals include:
  - Cluster cells into subgroups.
  - Model temporal transcriptomic dynamics: reconstruct “pseudo-time” for cells. This is useful for understanding development or disease progression.

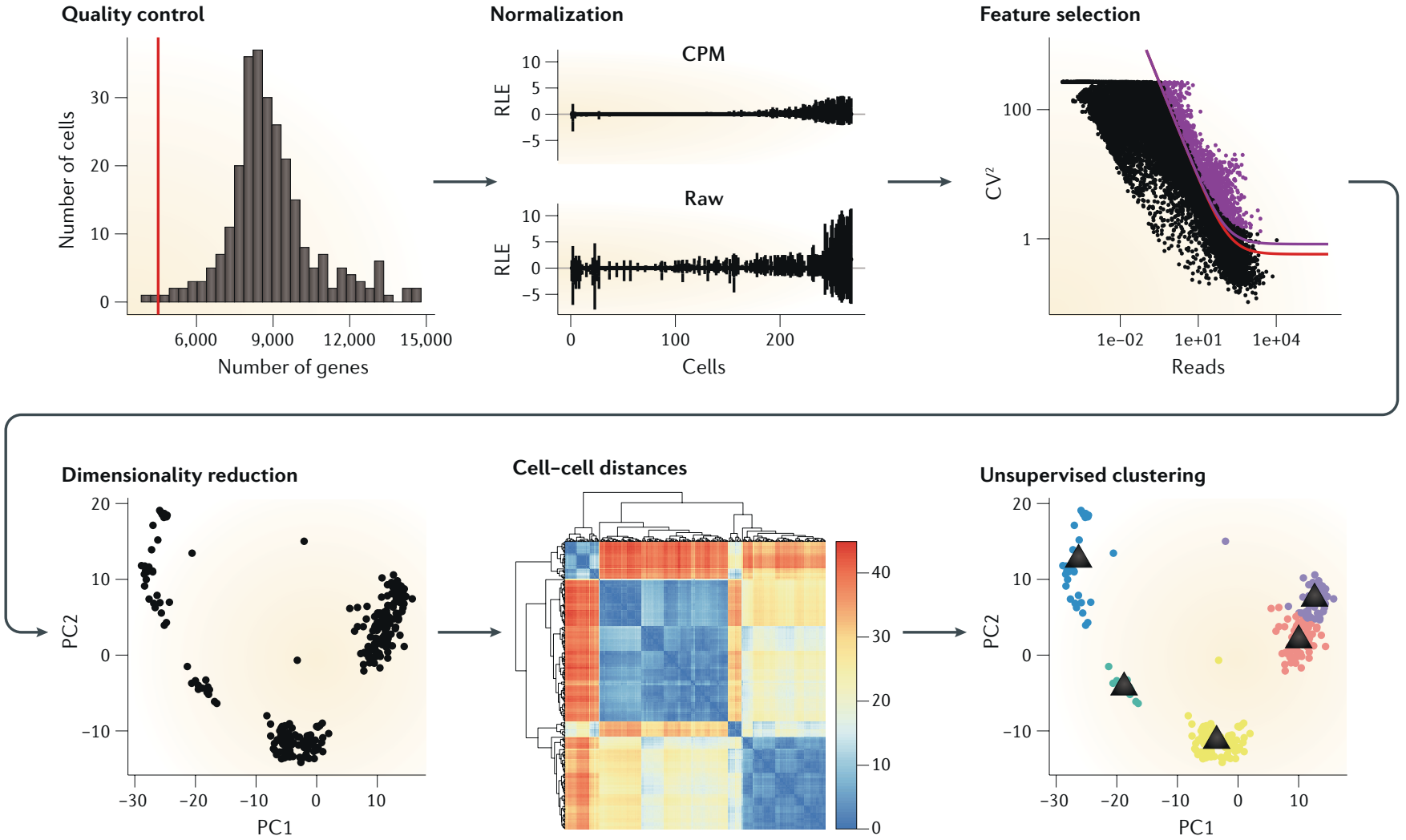
Cell clustering methods

- Many methods available
  - SC3, Seurat, TSCAN, Monocle, CIDR, …
  - Comprehensively compared in Duo et. al (2018) F1000 Research.
  - According to our experience: SC3 has the best performance, but is the slowest.

and robust [73]. Due to the heavy time consuming nature of consensus clustering, a rule of thumb for unsupervised single cell clustering is to use single-cell consensus clustering (SC3, integrated in Scater [52]) when the number of cells is < 5000 but use Seurat instead when there are more than 5000 cells.

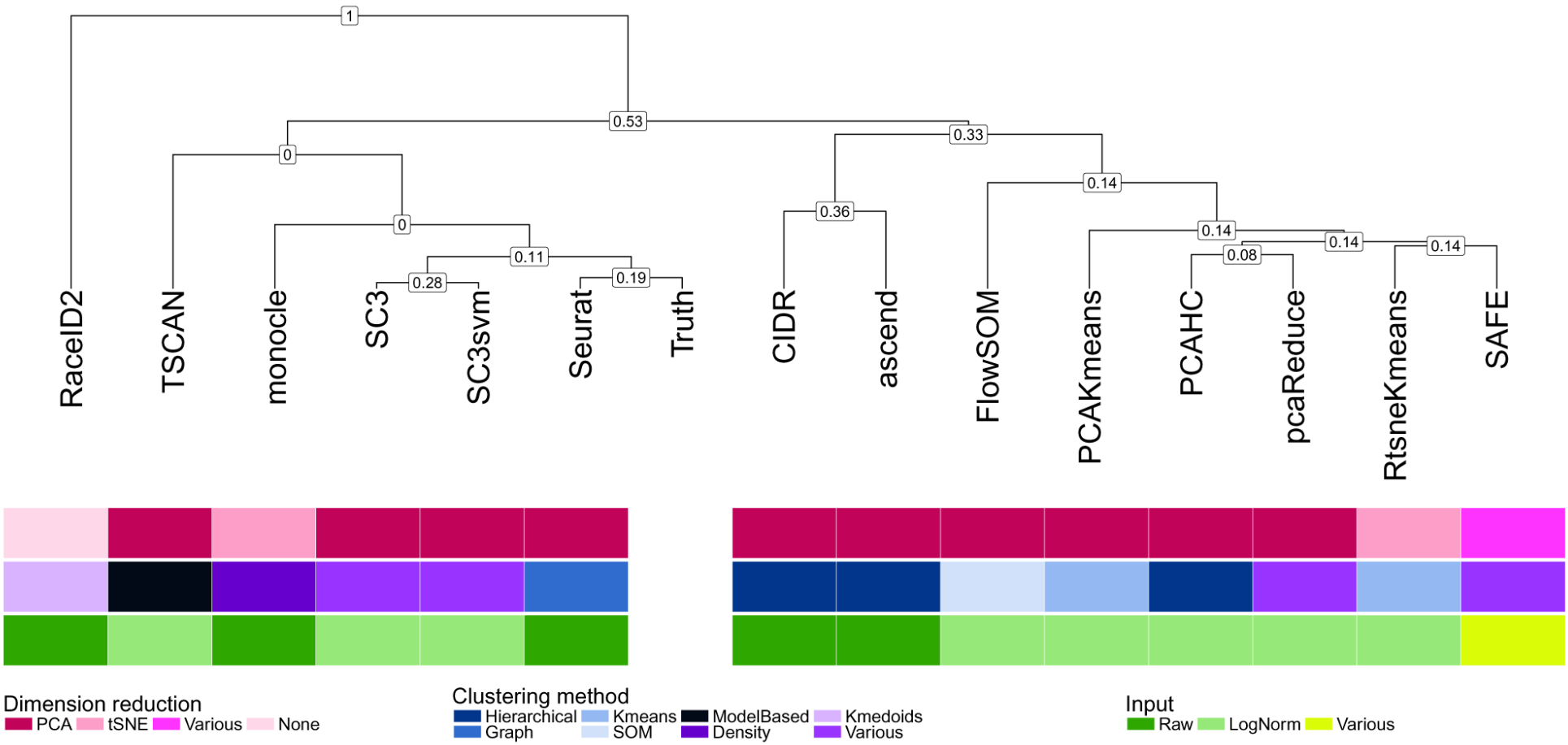
Mu et al. Genomics Proteomics Bioinformatics (2019)

Essence of the clustering methods



Kiselev et al. (2019) Nat. Rev. Genet.

Cell clustering methods



Duo et. al (2018) F1000 Research

Example codes for SC3

```
sce = SingleCellExperiment(
    assays = list(
        counts = as.matrix(counts),
        logcounts = log2(as.matrix(counts) + 1)
    )
)
sce = sc3_prepare(sce)
if( missing(K) ) { ## estimate number of clusters
    sce = sc3_estimate_k(sce)
    K = metadata(sce)$sc3$k_estimation
}

sce = sc3_calc_dists(sce)
sce = sc3_calc_transfs(sce)
sce = sc3_kmeans(sce, ks = K)
sce = sc3_calc_consens(sce)
result = colData(sce)[,1]
```

Example code for Seurat

```
seuset = CreateSeuratObject( counts )
seuset = NormalizeData(object = seuset)
seuset = FindVariableFeatures(object = seuset)
seuset = ScaleData(object = seuset)
seuset = RunPCA(object = seuset)
seuset = FindNeighbors(object = seuset)
seuset = FindClusters(object = seuset)
Result = seuset@active.ident
```

Pseudotime construction

- This belongs to the “clustering” category.
- Instead of putting cells into independent, exchangeable groups, it orders the cells by underlying temporal stage (estimated).
- Methods/tools:
  - Monocle/monocle2: Trapnell et al. (2014) Nat. Biotechnol; Qiu et al. (2017) Nat. Methods.
  - Waterfall: Shin et al. (2015) Cell Stem Cell
  - Wanderlust: Bendall et al. (2014) Cell
  - TSCAN: Ji et al. (2016) NAR

Pseudotime construction method

General steps:

1. Select informative genes.
2. Dimension reduction of GE.
3. Cluster the cells based on reduced data. Often want to over-cluster them to have many groups.
4. Construct a MST (miminum spanning tree) from the clustering results.
5. Map cells to the MST.

Cell clustering for multiple samples

- When scRNA-seq data are from multiple samples, batch effects could have significant impact on the results.
- Cells from the same sample, instead of the same cell type form different sample, can cluster together.
- Possible solution:
  - Remove batch effect then cluster: MNN + SC3
  - Jointly model cell type and sample effect: BAMM-SC (Sun et al. 2019, Nat. Comm)
- Still an open problem.

Cell type annotation

- Another paradigm to identify cell type.
- Cell clustering:
  - Cluster cells to multiple clusters (unsupervised). then assign cell type for each cluster.
- Cell type assignment:
  - Directly assign each cell to a cell type.
  - Requires some reference, or training data (supervised).
  - Potentially work better for data from multiple samples.
  - Can incorporate the hierarchy in cell types.
  - Cannot identify new cell types (restricted to the known cell types in the reference).

Cell annotation methods

- Pre-train a classifier using training set first, predict labels by kNN/correlation/RF etc.
  - scmap (Kiselev et al. 2018 Nat. Methods)
  - CaSTLe (Lieberman et al. 2018 Plos One)
  - Garnett (Pliner et al. 2019 Nat. Methods)
  - CHETAH (Kanter et al. 2019 Nucleic Acids Research)
- Marker-based classifier
  - CellAssign (Zhang et al. 2019 Nat. Methods)
- Other generic machine learning methods: SVM, LDA, RF, kNN, RF
- Comprehensively compared in Abdelaal et al. Genome Biology 2019
- Annotation performance is a trade-off between accuracy and un-assigned rate

scmap: projection of scRNA-seq data across datasets

- Correlation based assignment
- User can specify a threshold. Cells below the threshold are "unassigned"

```
sce <- SingleCellExperiment(assays =
    list(normcounts = as.matrix(trainmat)),
    colData = DataFrame(cell_type1 = trainlabel))
logcounts(sce) <- log2(normcounts(sce) + 1)
rowData(sce)$feature_symbol <- rownames(sce)
sce <- selectFeatures(sce, suppress_plot = TRUE)

sce_test <- SingleCellExperiment(assays =
    list(normcounts = as.matrix(testmat)),
    colData = DataFrame(cell_type1 = testlabel))
logcounts(sce_test) <- log2(normcounts(sce_test) + 1)
rowData(sce_test)$feature_symbol <- rownames(sce_test)

sce <- indexCluster(sce)
scmapCluster_results <- scmapCluster(projection = sce_test,
        index_list = list(metadata(sce)$scmap_cluster_index))
```

CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing

- Adopt a hierarchical structure when assign the cells
- Allow intermediate or unassigned categories
- Especially good when cells of unknown type are encountered, e.g. tumor

```
sce_train <- SingleCellExperiment(assays =
    list(counts = as.matrix(trainmat)),
        colData = DataFrame(celltypes=trainlabel))

sce_test <- SingleCellExperiment(assays =
    list(counts = as.matrix(testmat)),
         colData = DataFrame(celltypes = testlabel))

#run classifier
test <- CHETAHclassifier(input = sce_test, ref_cells = sce_train)
test$celltype_CHETAH
```

Differential expression (DE)

- DE analysis is the most important task for bulk expression data (microarray or RNA-seq).
- DE in scRNA-seq is a little different:
  - Traditional methods test mean changes, while the consideration and modeling of “drop-out” event (non-expressed) is important in sc data.
  - Considering cell types: can compare cross cell types or compare the same cell type cross biological conditions.

DE methods

- SCDE (Kharchenko et al. 2014 Nat. Methods)
- MAST (Finik et al. 2015 GB)
- SC2P (Wu et al. 2018 Bioinformatics)
- Seurat and monocle also provides DE functions.
- Bulk methods (DESeq, edgeR) are sometimes used.
- A comparison paper: Soneson and Robinson (2018) Nat. Methods

Finak *et al. Genome Biology* (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

- MAST: "Model-based Analysis of Single- cell Transcriptomics."
- Bioconductor package **MAST**.

MAST for DE

- Main ideas:
  - Use log2(TPM+1) as input data
  - Both dropout probability and expression level depends on experimental conditions.

$$logit\left(Pr(Z_{ig}=1)\right)=\mathrm{X_i}\,\beta_{\mathrm{g}}^{\mathrm{D}}$$

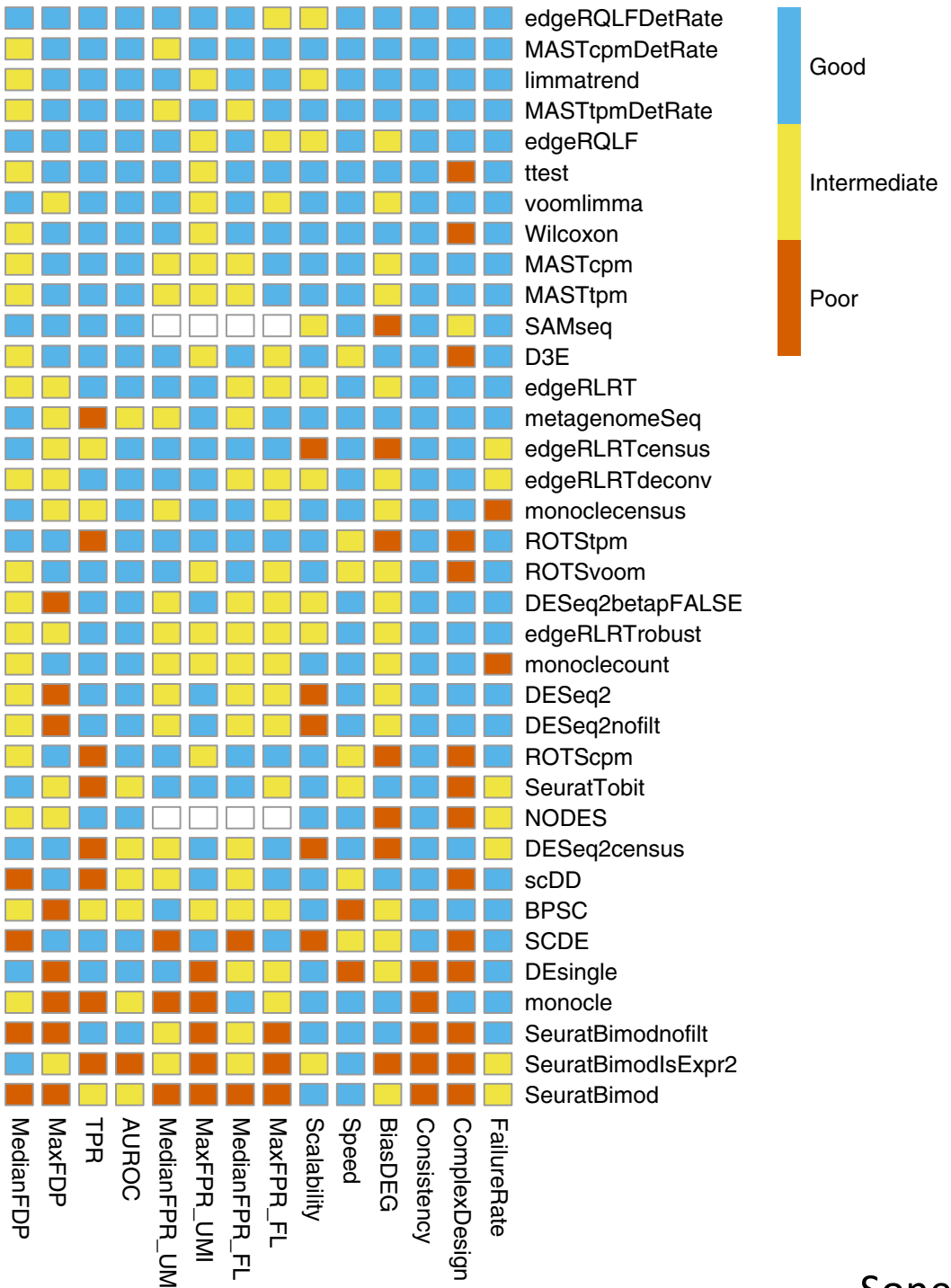
$$\Pr\left(Y_{ig}=y|Z_{ig}=\ 1\right)=\ \mathrm{N}\left(\mathrm{X_i}\beta_{\mathrm{g}}^{\mathrm{C}},\ \sigma_g^2\right)$$

  - Model fitting with some regularization.
  - DE is based on chi-square or Wald test.

Example codes for MAST

- Start from log TPM and biological condition

```
sca <- FromMatrix(ltpm,
       cData=data.frame(celltype))
cdr2 <- colSums(assay(sca)>0)
colData(sca)$cngeneson <- scale(cdr2)
thres <- thresholdSCRNACountMatrix(assay(sca),
       nbins=200, min_per_bin=30)
assays(sca) <- list(thresh=thres$counts_threshold,
              tpm=assay(sca))
## fit model and perform test
fit <- zlm(~celltype, sca)
lrt <- lrTest(fit, "celltype")
```



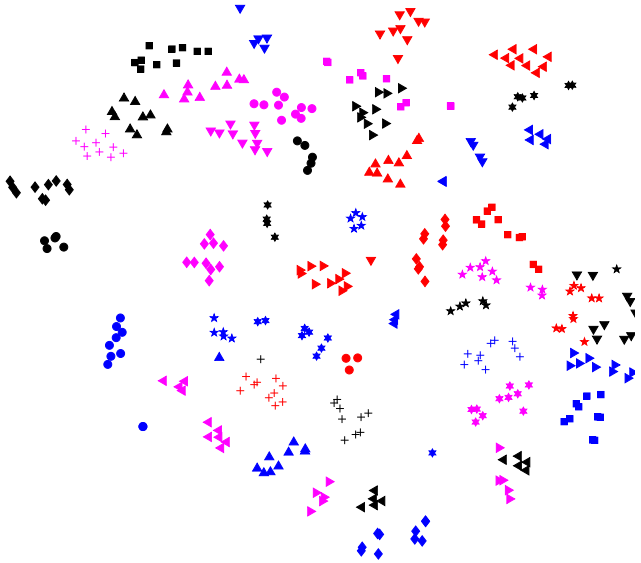
Soneson and Robinson (2018) Nat. Methods

Visualization

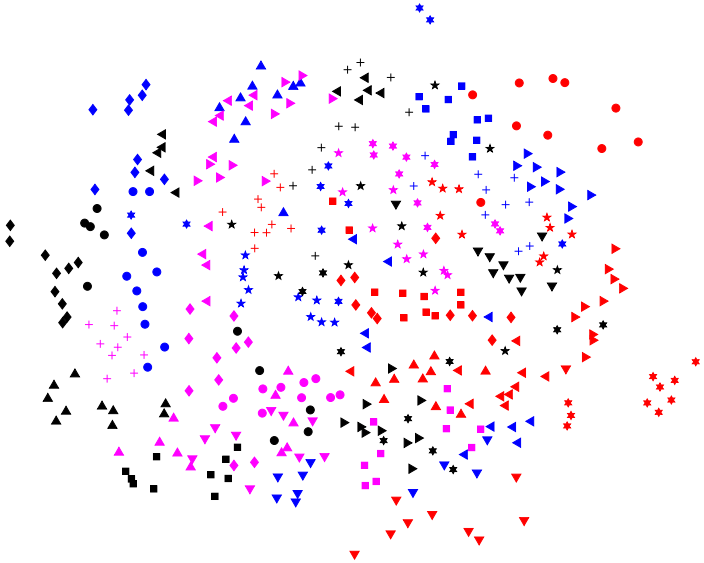
- TSNE
- UMAP

t-SNE: a useful visualization tool

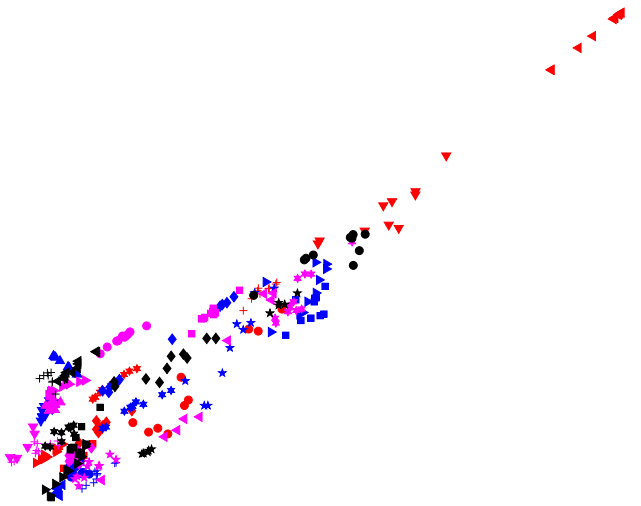
- t-SNE (t-distributed stochastic neighbor embedding): visualize high-dimensional data on 2-/3-D map.
- When project high-dimensional data into lower dimensional space, preserve the distances among data points.
  - This alleviate the problem that many clusters overlap on low dimensional space.
- Try to make the pairwise distances of points similar in high and low dimension.
- This is used in almost all scRNA-seq data visualization.
- Has “Rtsne” package on CRAN.



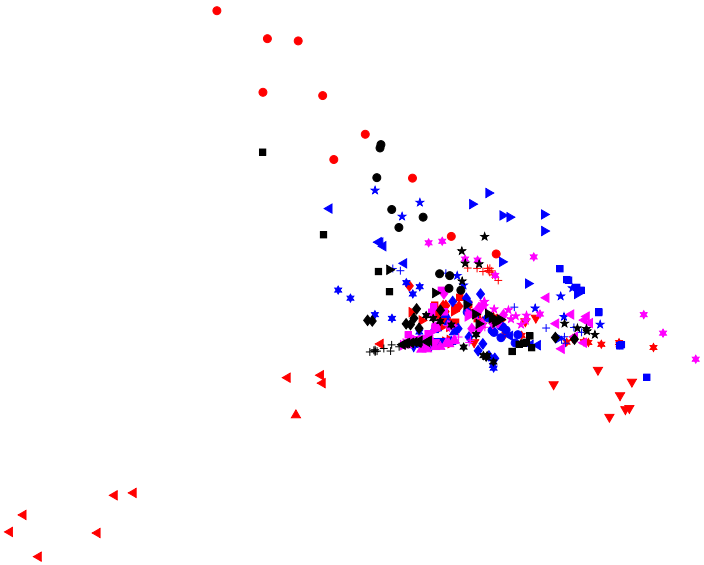
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



(c) Visualization by Isomap.



(d) Visualization by LLE.

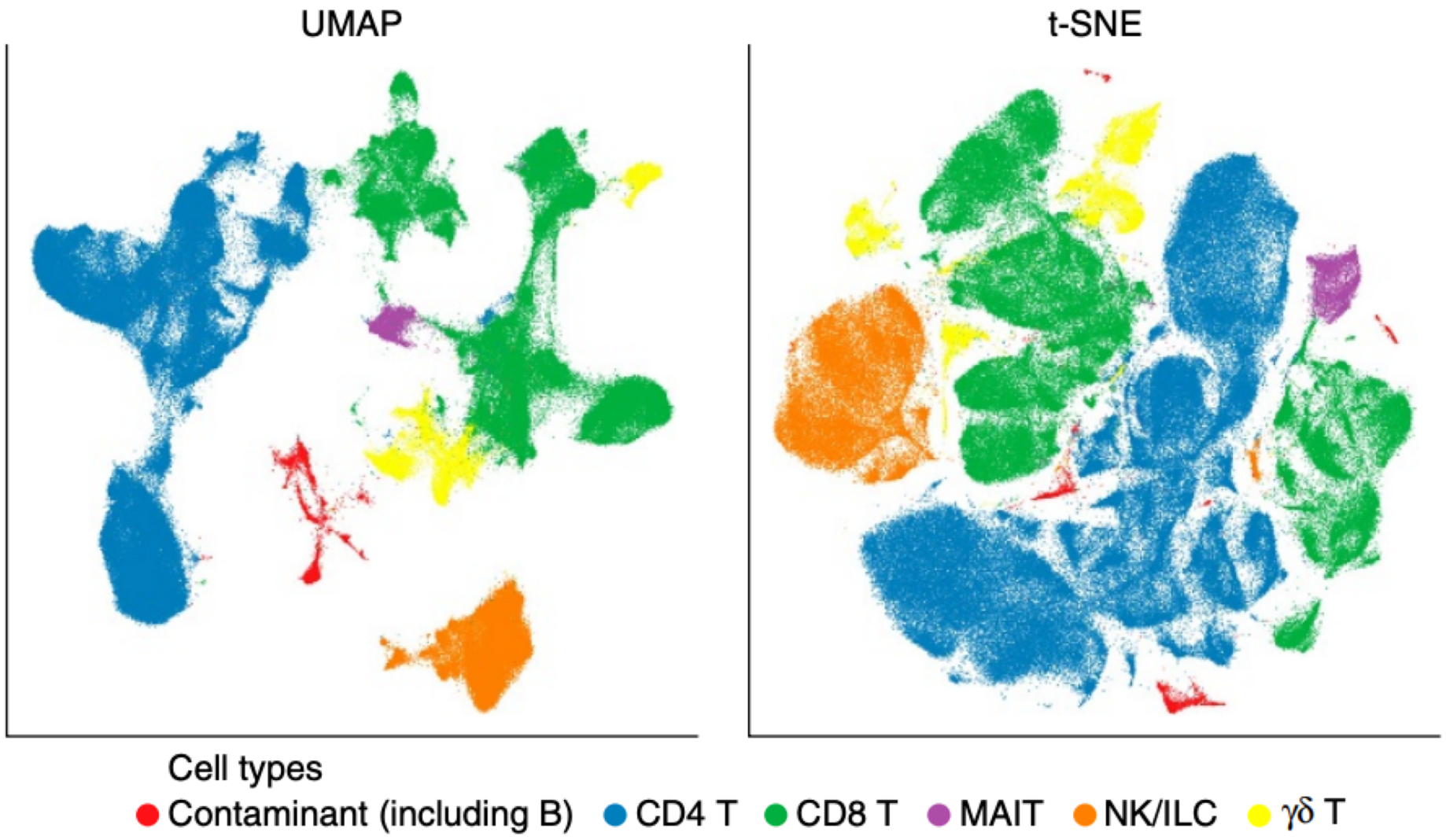
Example code for t-SNE

```
library(Rtsne)
tsne_model_1 = Rtsne(datamatrix,  check_duplicates=FALSE, pca=TRUE,
                perplexity=30, theta=0.5, dims=3)
tsne_out = as.data.frame(tsne_model_1$Y)

pdf(”your_figure_name.pdf", width = 5, height = 5)
par(mar = c(2.4, 2.4, 0.5, 0.5), mgp = c(1.2, 0.4, 0))
plot(tsne_out$V1, tsne_out$V2,  pch = 19, cex = 0.4, col = mycolor)
legend("bottomleft", col = mycolor,  legend = uniqCT, pch = 19,
       cex = 0.5, bty = "n")
dev.off()
```

UMAP: a newer (and better?) visualization tool

- UMAP (uniform manifold approximation and projection): a recently developed dimension reduction tool
- "*Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters.* " ---- Betcht et al. 2018 Nat Biotech
- "*UMAP, which is based on theories in Riemannian geometry and algebraic topology, has been developed, and soon demonstrated arguably better performance than t-SNE due to its higher efficiency and better preservation of continuum.*" ---- Mu et al. 2018 GBP
- Has "umap" package on CRAN.



Betcht et al. 2018 Nat Biotech

Example code for UMAP

```
library(umap)
sim_umap <- umap(datamatrix)
sim_umap2 <- sim_umap$layout
colnames(sim_umap2) <- c("UMAP1", "UMAP2")

pdf(”your_figure_name.pdf", width = 5, height = 5)
par(mar = c(2.4, 2.4, 0.5, 0.5), mgp = c(1.2, 0.4, 0))
plot(sim_umap2[,1], sim_umap2[,2], pch = 19, cex = 0.4, col = mycolor)
legend("bottomleft", col = mycolor, legend = uniqCT, pch = 19,
            cex = 0.5, bty = "n")
dev.off()
```

Summary

- The main interests are inter-cellular heterogeneity, expression dynamics, cell type discovery, etc.
- Many statistical methods and computational tools for different biological questions.
  - Data pre-processing: normalization, batch effect, imputation
  - Cell clustering and cell type annotation
  - Differential expression