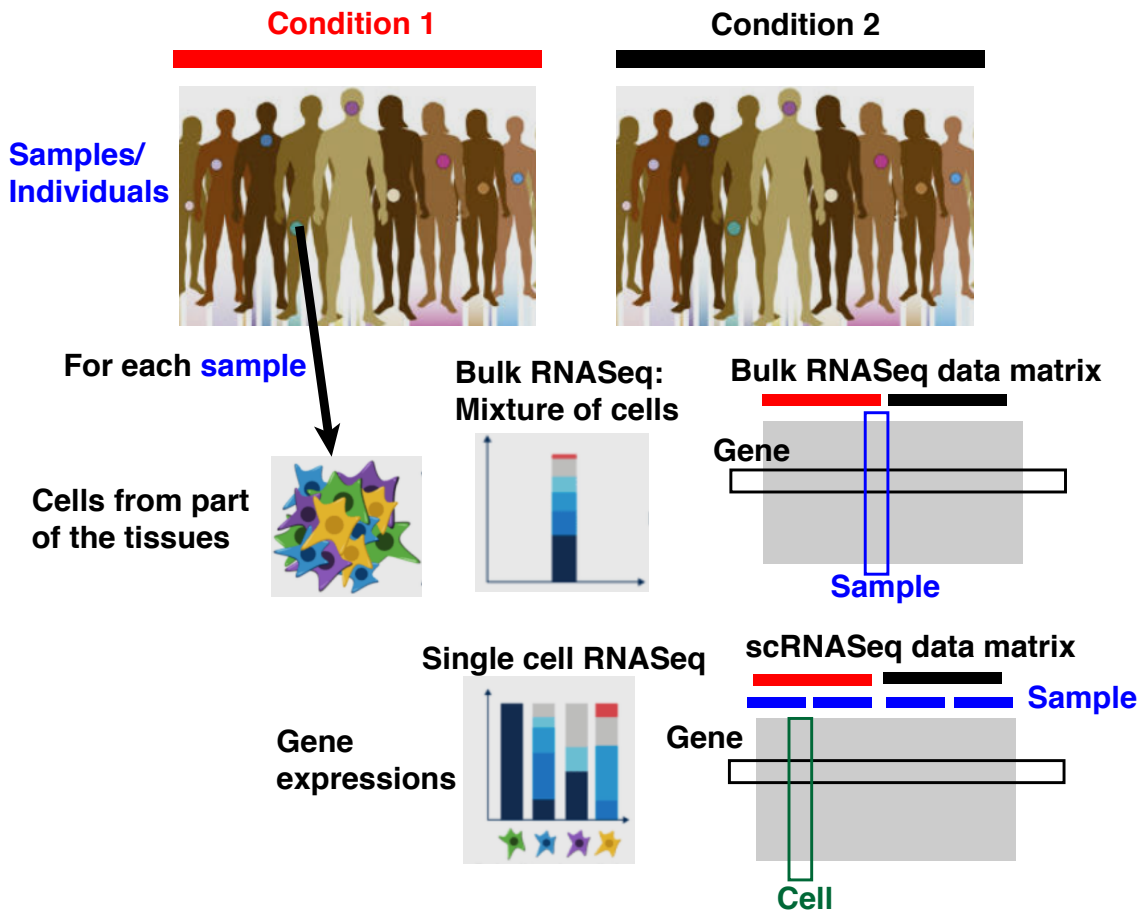
Multiple-Sample Differential Expression Analysis for Single-Cell RNA Sequencing Data

Songpeng Zu

October 22, 2020

Outline

Finding genes that are differentially expressed (shift of the mean) under different conditions.



- Bulk RNASeq:
  the gene i in the sample j
  $\log_2 CPM(X_{ij}) = \log_2(1 + X_{ij} \cdot S_j)$
  $S_j = \frac{1}{\sum_i X_{ij}} \cdot 10^6$
- UMI-based scRNAseq
  - Data: 70% zeros
  - Limited samples
  - Technical variations:
    1. Batch effect
    2. scRNA sequencing
  - Biological variations:
    1. Genetic background
    2. Cell heterogeneity

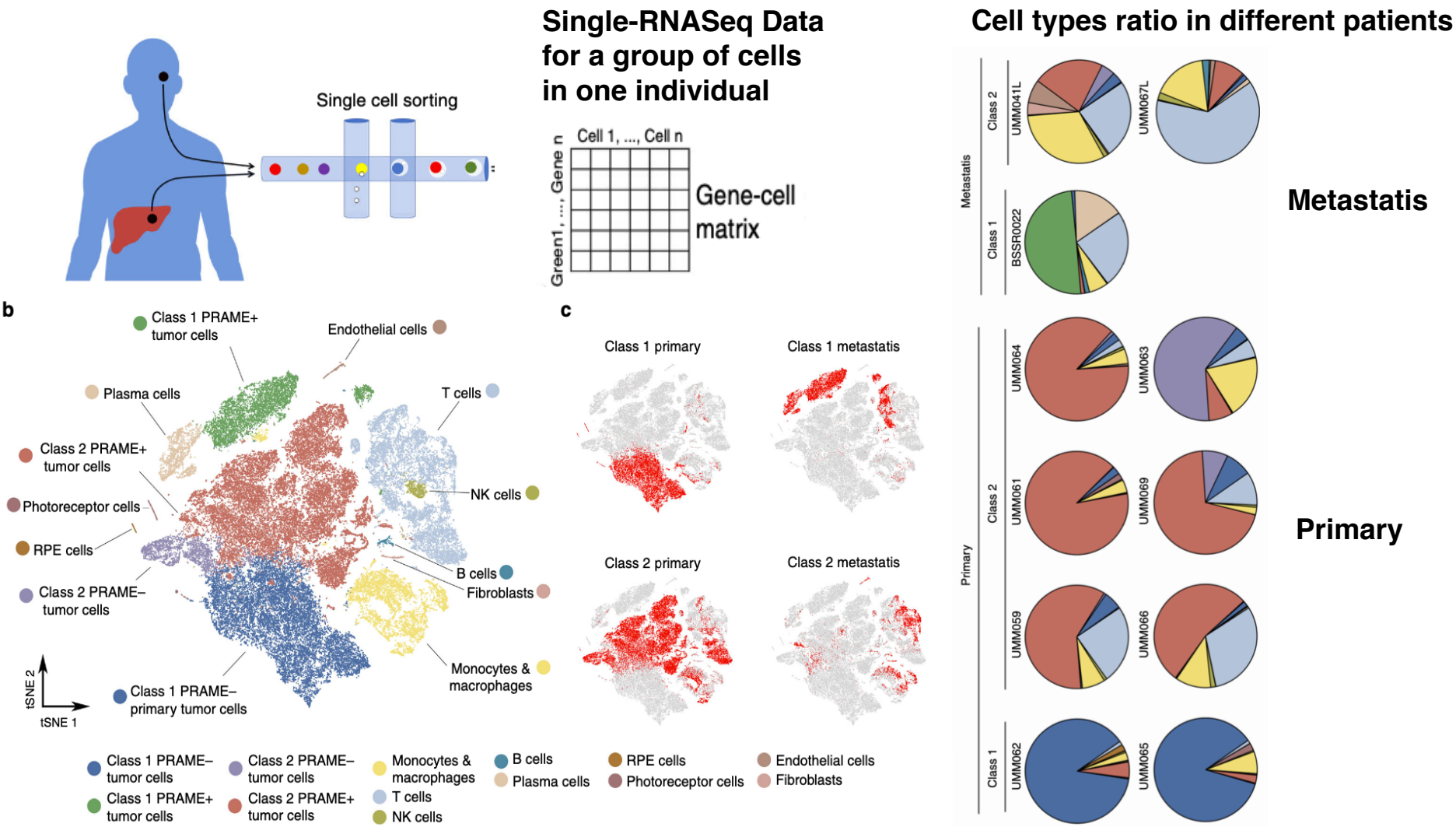


Figure 1: Multiple-sample scRNAseq Data [1].

Patient-specific effects cause high false positives

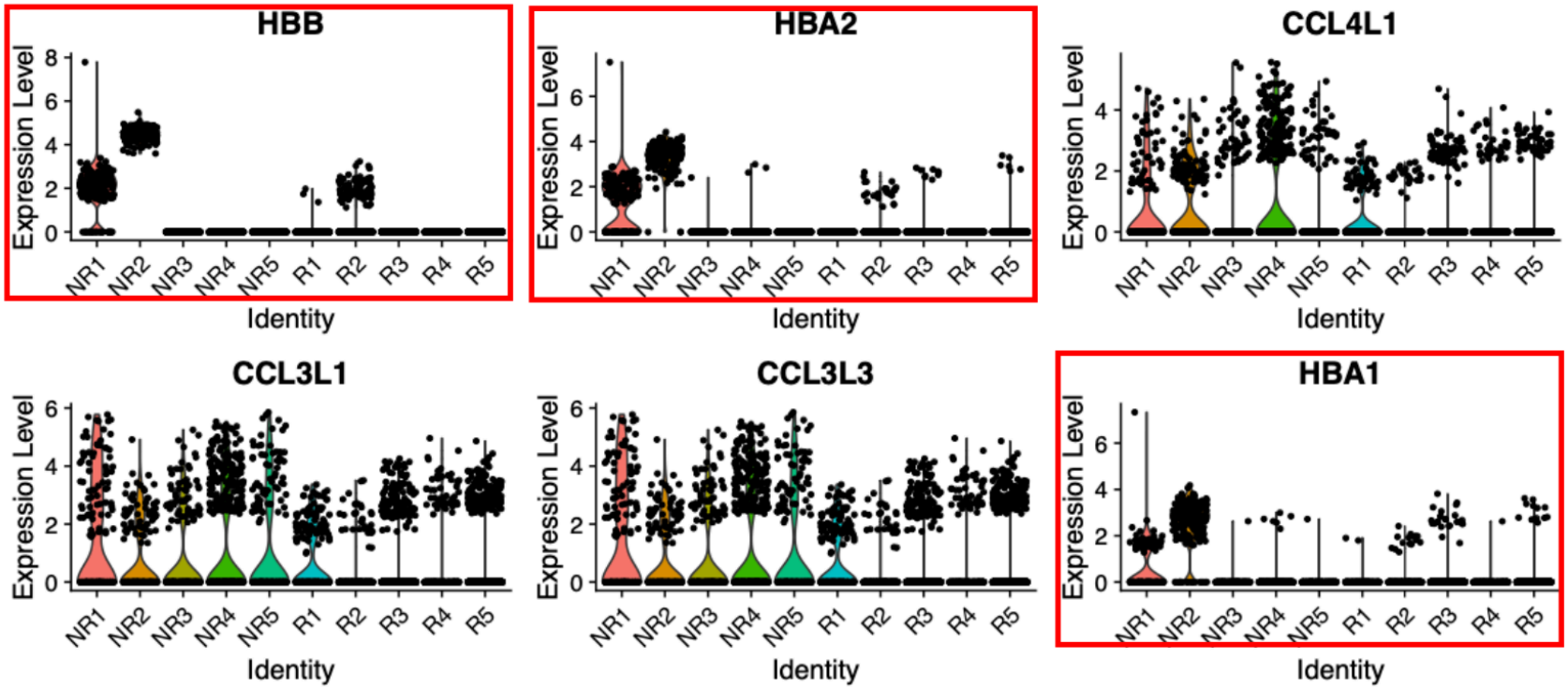
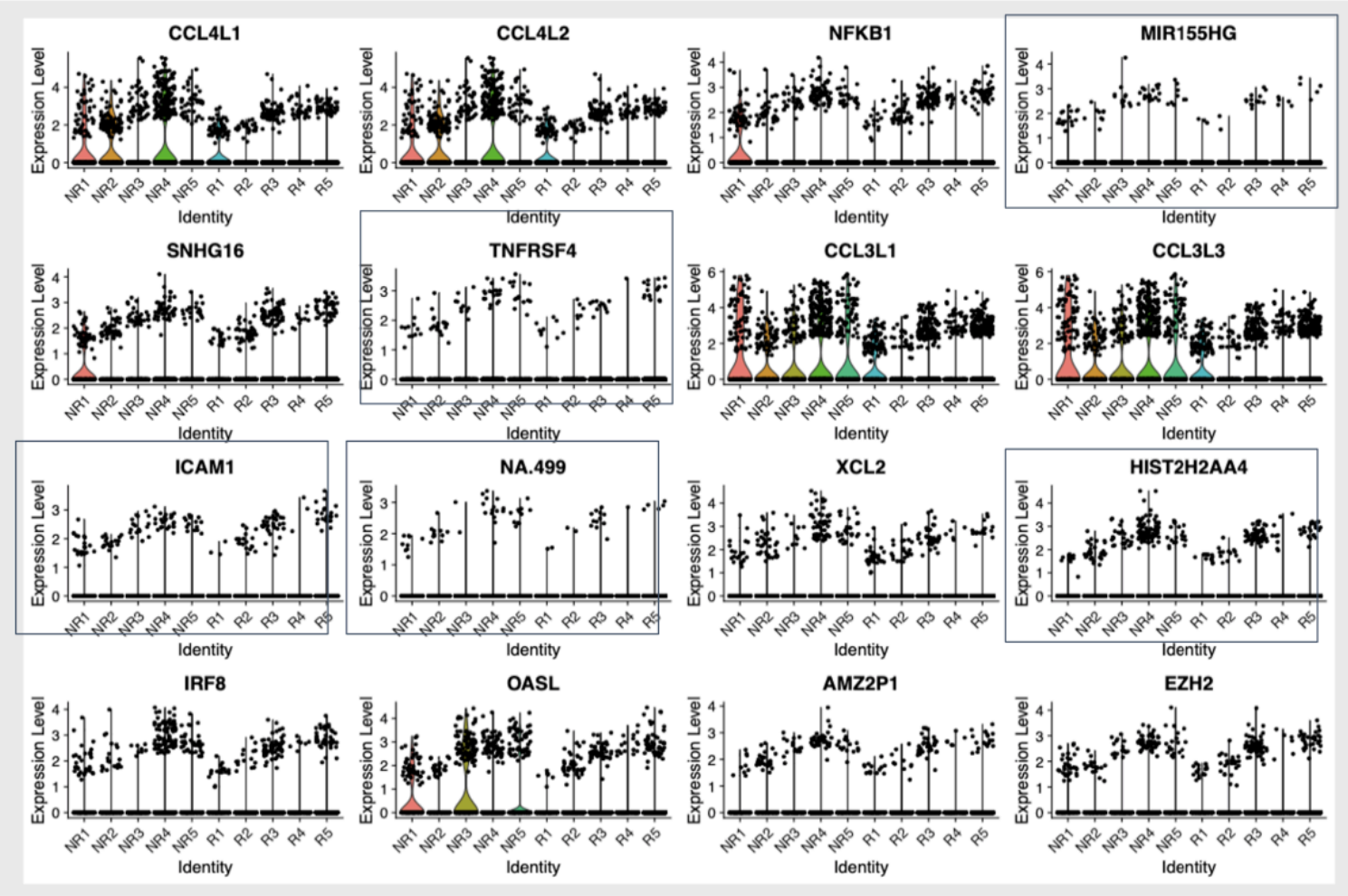


Figure 2: DE analysis on one cell sub-population inferred by Harmony [2]

Pseudo-bulk DEG is vulnerable to gene expressed in few cells



Any bulk RNA-seq might be also vulnerable to gene expressed in few cells!

Figure 3: Pseudo-bulk analysis [3] on PBMC dataset.

Outline

Modeling the mean

Let X_{ij} represent the UMI counts for the gene i in the cell j

$$X_{ij} \sim f(\mu_{ij}, \phi_i) \tag{1}$$

$$\mu_{ij} = S_j * \lambda_{ij} \tag{2}$$

$$g^{-1}(\lambda_{ij}) = \sum_r d_{jr}\beta_{ir} \tag{3}$$

$$\boldsymbol{\beta_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}) \tag{4}$$

A typical content in $\mathbf{d_j}$ is followed:

$$\mathbf{d_j} = \left(1, \overbrace{(0, 1, \ldots, 0, 0)}^{\text{individual one-hot vector}}, \underbrace{(1, 0)}_{\text{conditional one-hot vector}}\right)$$

Outline

UMI-Count vs TPM

- Data set: PBMC scRNAseq data [4], 10 individuals, 5 vs 5 in two conditions.
- Select genes:
  - DEGs: CCL4L1, CCL4L2, CCL3L1, CCL3L3
  - Genes with lots of zeros: MIR155HG, TNFRSF4, ICAM1, NA.499, HIST2H2AA4
  - Genes with strong individual effects: HBB, HBA2, HBA1

What is a violin plot?

Violin plots are similar to box plots, except that they also show the kernel probability density of the data at different values.

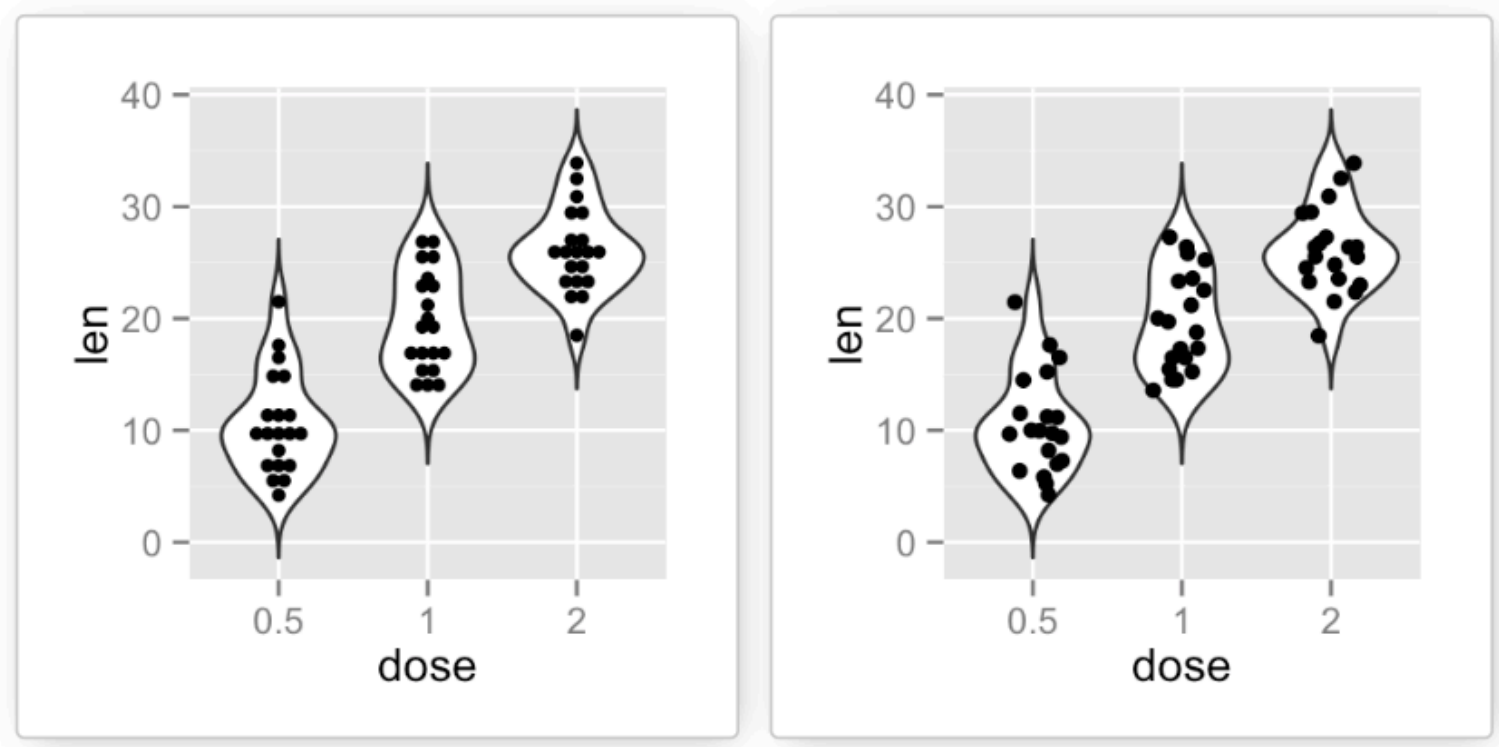
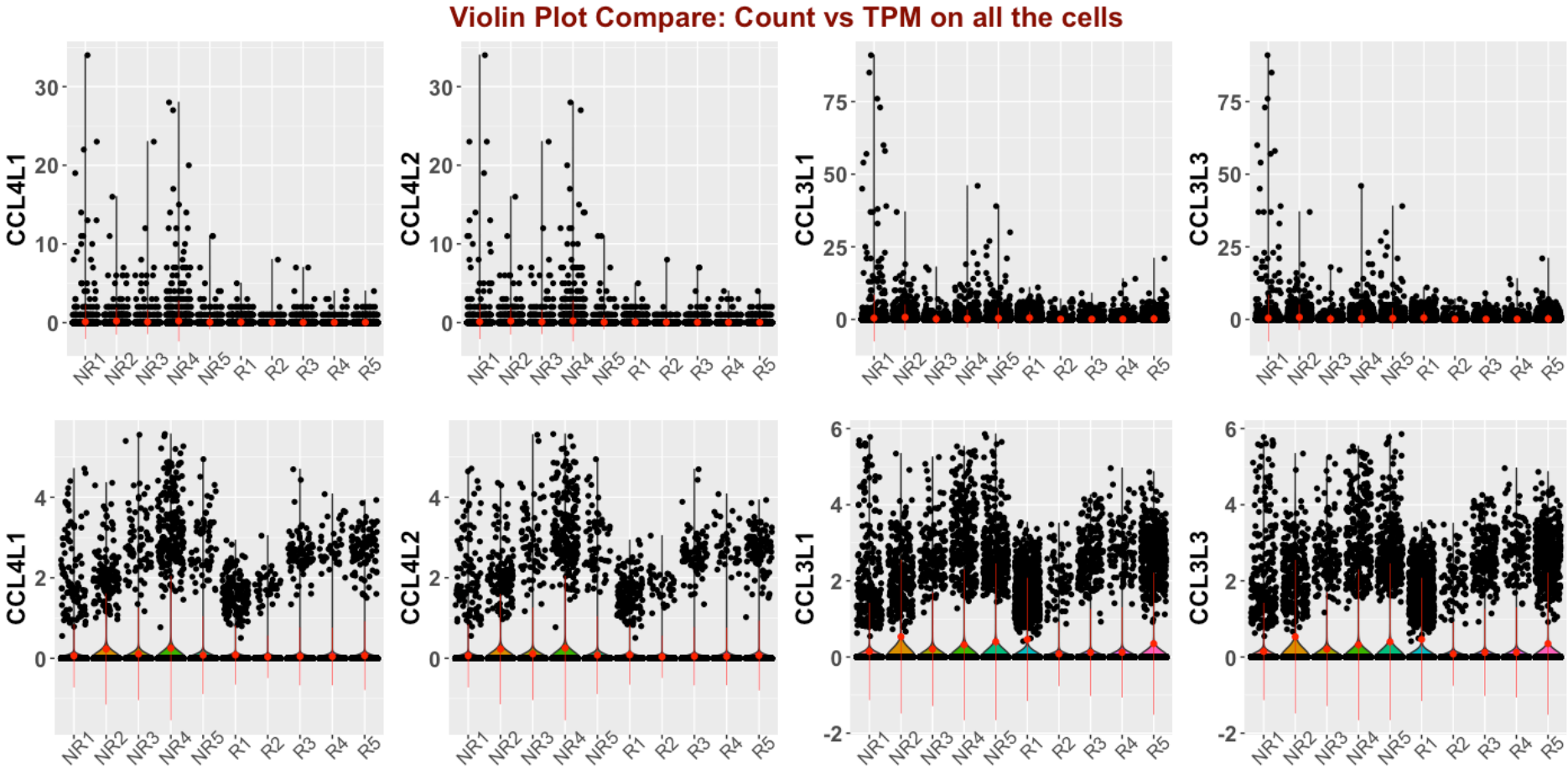
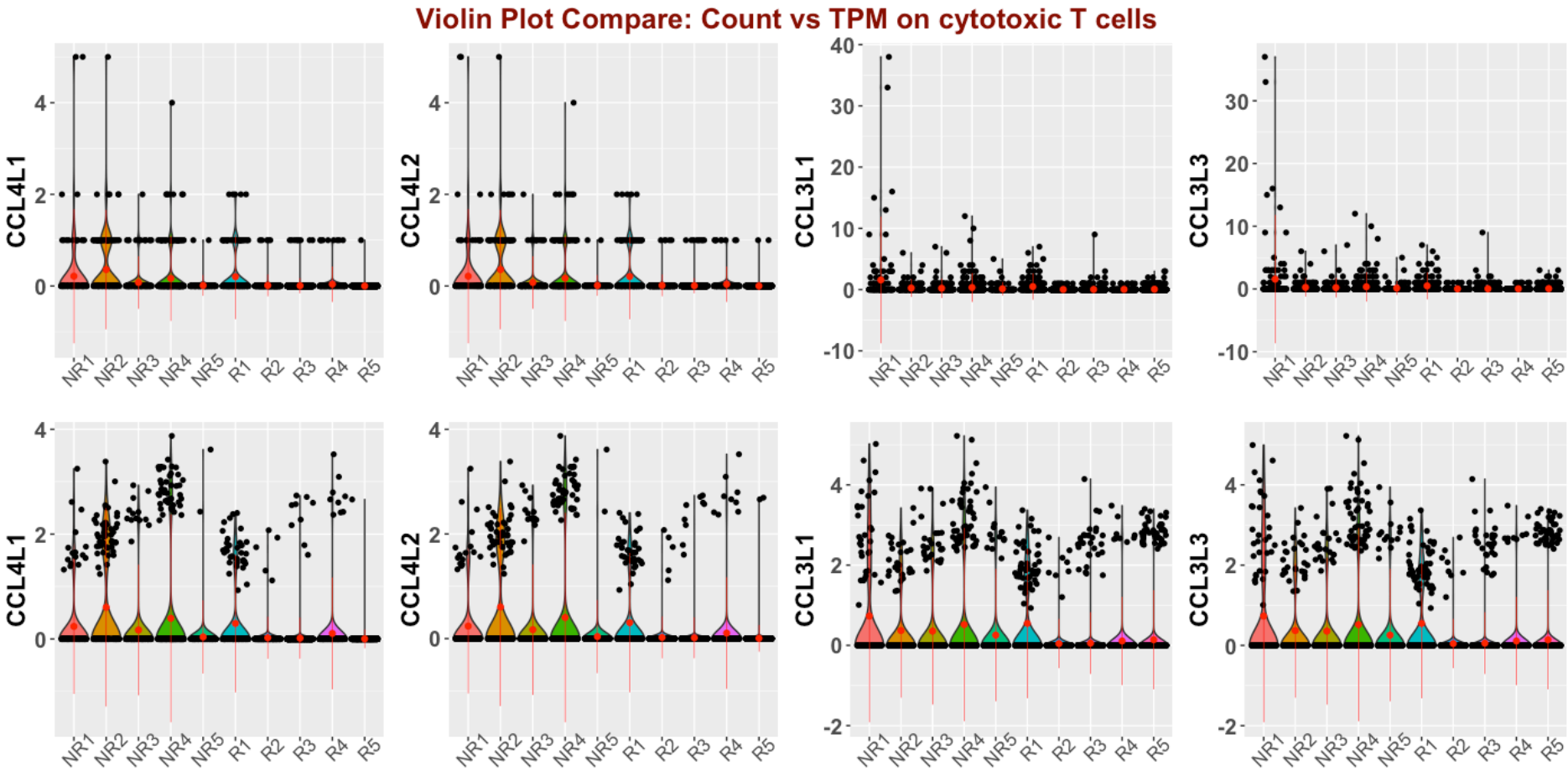


Figure 4: Violin plot. Left: Dotplot, where each dot corresponds to one sample; Right: Jitter plot, which adds small randomness to the data in order to reduce the overlap among the points.

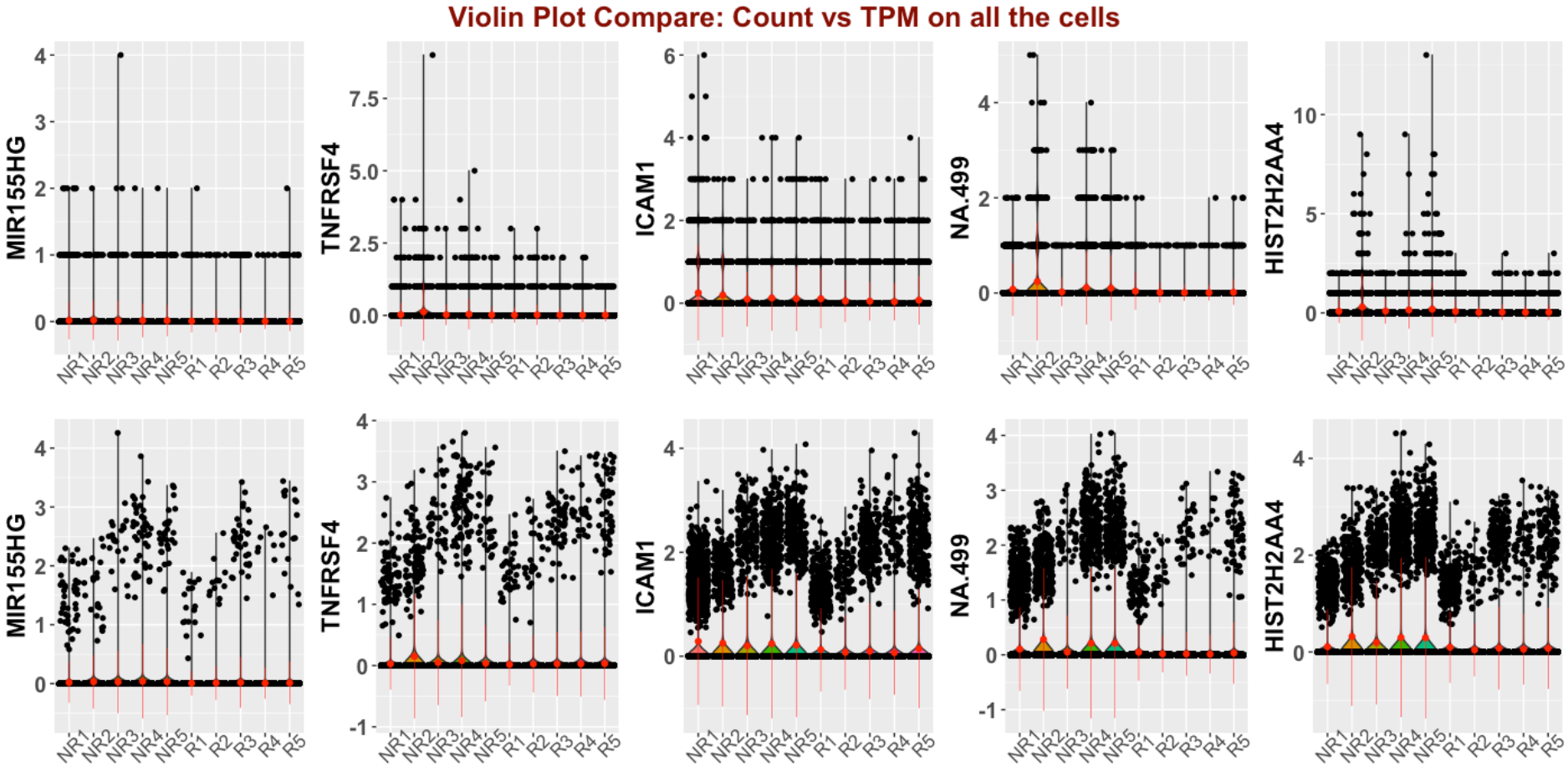
DEGs in all cells



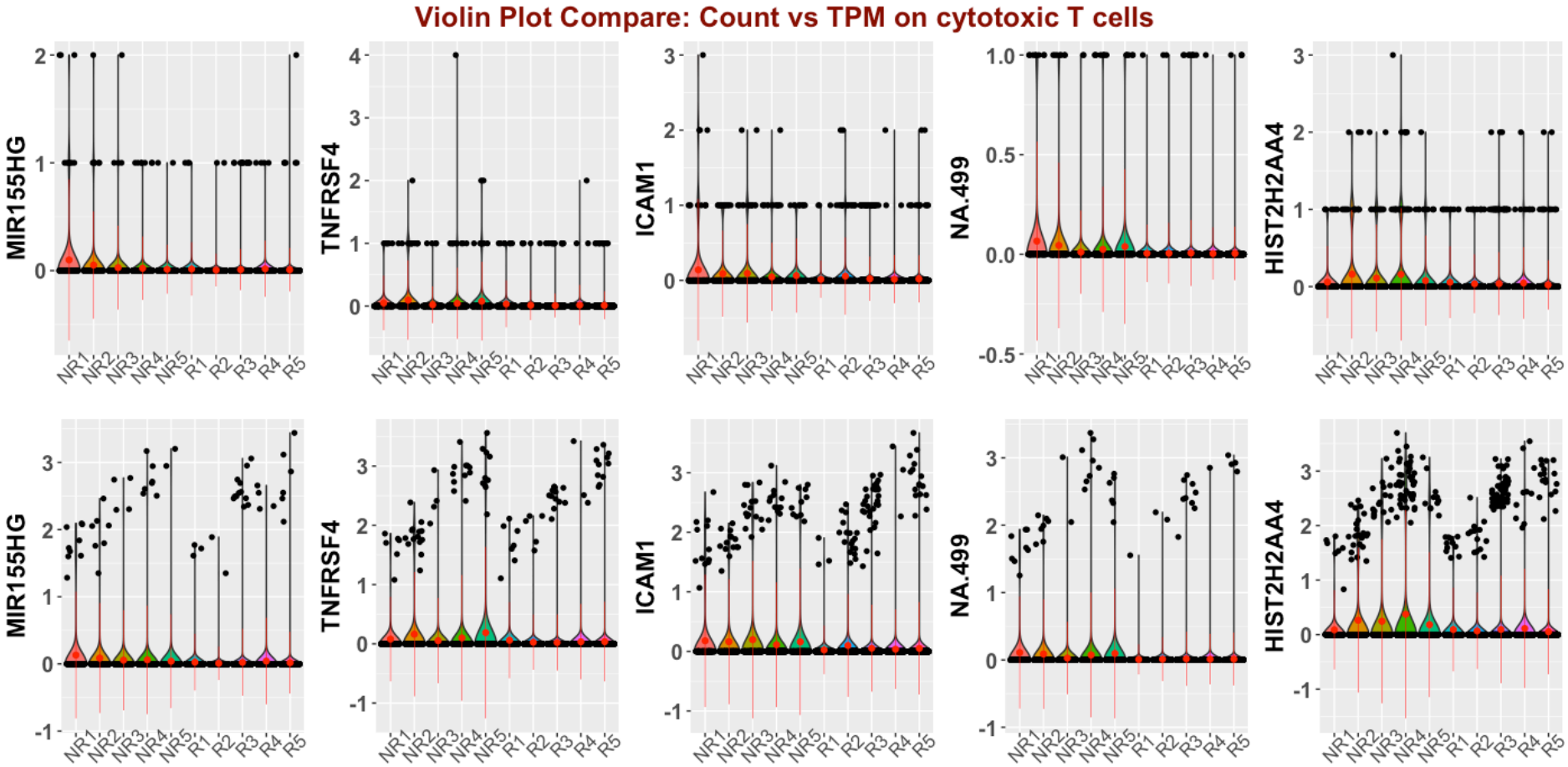
DEGs in cytotoxic T cells



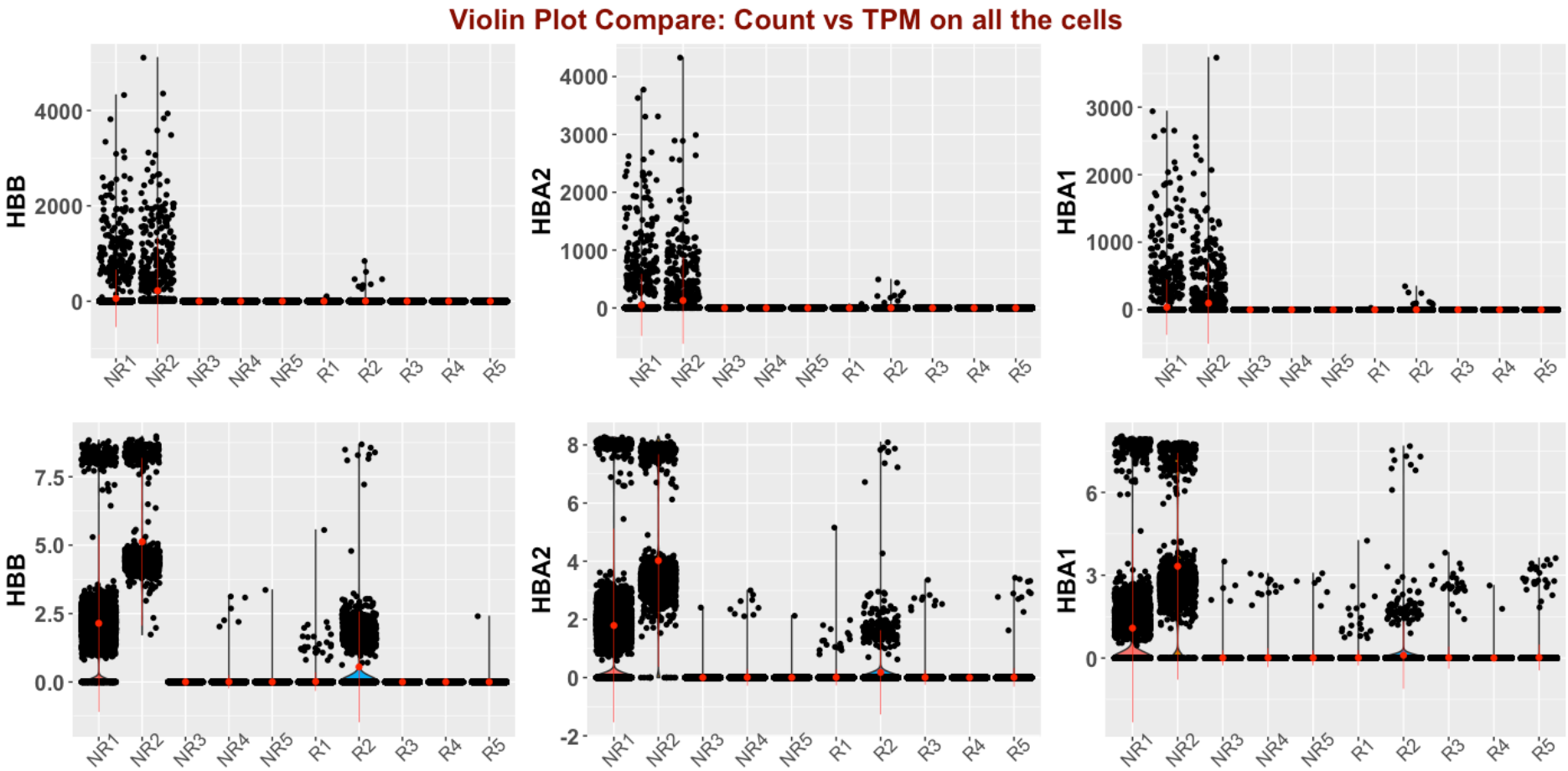
Genes with lots of zeros in all cells



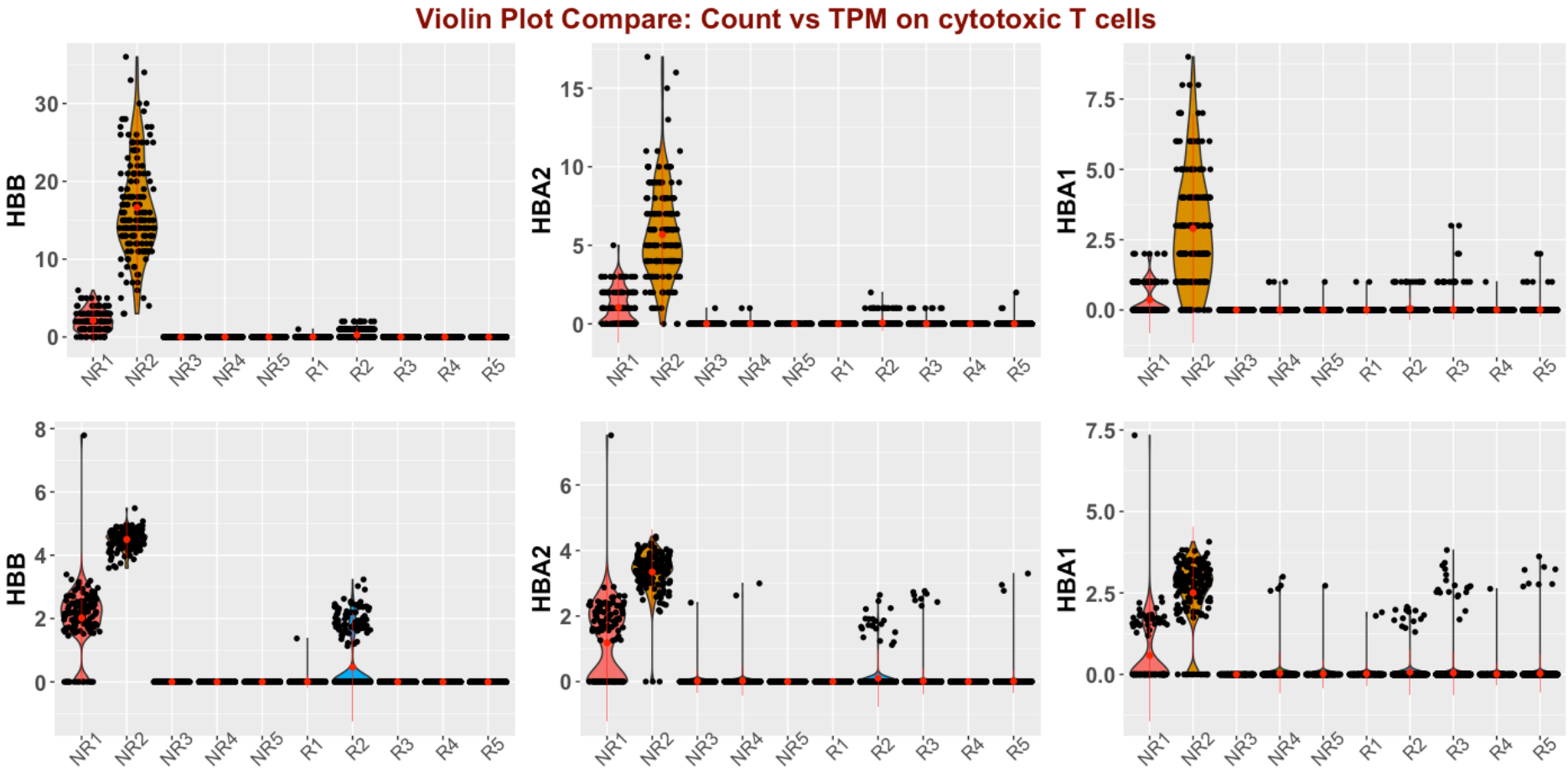
Genes with lots of zeros in cytotoxic T cells



Genes with strong individual effects in all cells

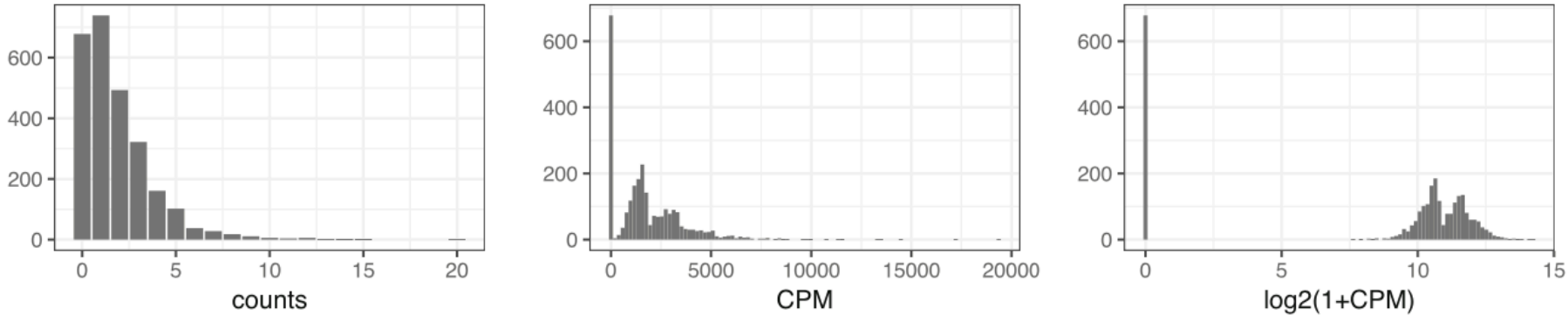


Genes with strong individual effects in cytotoxic T cells



Multinomial model on scRNASeq [5]

Under a multinomial distribution on the observed UMI counts, 74%-90% zeros, 22-30% ones, and less than 4% values above one. This will artificially enhance the gap between zero and nonzeros values on log-normalized data.



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

Fig. 2 Example of how current approaches to normalization and transformation artificially distort differences between zero and nonzero counts. **a** UMI count distribution for gene ENSG00000114391 in the monocytes biological replicates negative control dataset. **b** Counts per million (CPM) distribution for the exact same count data. **c** Distribution of $log_2(1 + CPM)$ values for the exact same count data

Outline

For any given gene, the UMI count X_j in the cell j, S_j is the scale factor for the cell j, which reflect the sequencing depth in the cell j.

$$X_j \sim Poisson(\lambda) \tag{5}$$

$$X_j \sim Poisson(S_j \cdot \tilde{\lambda}) \tag{6}$$

$$X_j \sim PoissonlogNormal(\mu, \sigma^2)$$

$$= \int_\lambda Poisson(X|S_j \cdot \lambda) \cdot logNormal(\lambda|\mu, \sigma^2) \tag{7}$$

$$X_j \sim NegativeBinomial(S_j \cdot \mu, \phi) \tag{8}$$

- In NB distribution, we use the mean μ and dispersion ϕ (namely r) , and then the variance $\sigma^2 = \mu + \mu^2/\phi$, and the success probability $p = \frac{\mu}{\mu+\phi}$.
- NB can be treated as Gamma-Poisson mixture, i.e., $NB(\mu,\phi)=\int_\lambda Poisson(\lambda)Gamma(\phi,\phi/\mu)$, in which, ϕ is the shape, and ϕ/μ is the rate parameter in Gamma distribution.

One cell subpopulation from one individual

Gene	Obs	Poi	Pois	Poilognm	Poislognm	NB
CCL4L1	0.840	0.813	0.824	0.843	0.850	0.842
CCL4L2	0.837	0.810	0.822	0.840	0.847	0.839
CCL3L1	0.729	0.403	0.429	0.727	0.734	0.737
CCL3L3	0.729	0.404	0.430	0.727	0.734	0.736
MIR155HG	0.961	0.953	0.956	0.960	0.962	0.961
TNFRSF4	0.958	0.959	0.962	0.959	0.962	0.959
ICAM1	0.952	0.938	0.942	0.951	0.953	0.952
NA.499	0.971	0.971	0.973	0.971	0.973	0.971
HIST2H2AA4	0.942	0.944	0.948	0.944	0.948	0.944
HBB	0.661	0.445	0.470	0.621	0.624	0.644
HBA2	0.785	0.670	0.689	0.765	0.768	0.779
HBA1	0.875	0.863	0.871	0.872	0.878	0.874

Table 1: Zero ratio estimations for cells from Cluster 2, the individual R1

Two cell subpopulations from one individuals

Gene	Obs	Poi	Pois	Poilognm	Poislognm	NB
CCL4L1	0.961	0.953	0.957	0.960	0.963	0.961
CCL4L2	0.960	0.952	0.956	0.959	0.962	0.960
CCL3L1	0.802	0.689	0.711	0.798	0.806	0.803
CCL3L3	0.802	0.689	0.711	0.798	0.806	0.803
MIR155HG	0.990	0.989	0.990	0.990	0.990	0.990
TNFRSF4	0.989	0.989	0.990	0.989	0.990	0.989
ICAM1	0.803	0.783	0.800	0.802	0.812	0.803
NA.499	0.920	0.921	0.928	0.921	0.928	0.921
HIST2H2AA4	0.901	0.900	0.908	0.901	0.908	0.901
HBB	0.486	0.204	0.234	0.423	0.412	0.457
HBA2	0.535	0.333	0.365	0.496	0.492	0.520
HBA1	0.716	0.653	0.677	0.705	0.711	0.712

Table 2: Zero ratio estimations for cells from Cluster 1 and 2, the individual R1

One cell subpopulation from all the individuals

Gene	Obs	Poi	Pois	Poilognm	Poislognm	NB
CCL4L1	0.953	0.946	0.958	0.953	0.962	0.953
CCL4L2	0.952	0.945	0.957	0.952	0.961	0.953
CCL3L1	0.925	0.871	0.899	0.924	0.934	0.925
CCL3L3	0.925	0.871	0.899	0.924	0.934	0.925
MIR155HG	0.985	0.984	0.987	0.985	0.988	0.985
TNFRSF4	0.978	0.977	0.982	0.978	0.982	0.978
ICAM1	0.965	0.962	0.971	0.965	0.972	0.965
NA.499	0.987	0.987	0.990	0.987	0.990	0.987
HIST2H2AA4	0.948	0.945	0.957	0.948	0.958	0.948
HBB	0.911	0.469	0.557	0.892	0.907	0.911
HBA2	0.934	0.763	0.811	0.921	0.932	0.933
HBA1	0.941	0.867	0.895	0.936	0.945	0.941

Table 3: Zero ratio estimations for cells from Cluster2, 10 individuals.

Two cell subpopulations from all the individuals

Gene	Obs	Poi	Pois	Poilognm	Poislognm	NB
CCL4L1	0.969	0.965	0.974	0.969	0.975	0.969
CCL4L2	0.969	0.964	0.974	0.969	0.975	0.969
CCL3L1	0.857	0.761	0.819	0.856	0.873	0.858
CCL3L3	0.857	0.761	0.819	0.856	0.873	0.858
MIR155HG	0.991	0.990	0.993	0.990	0.992	NA
TNFRSF4	0.986	0.985	0.989	0.986	0.989	0.986
ICAM1	0.873	0.857	0.893	0.871	0.897	0.872
NA.499	0.926	0.916	0.938	0.926	0.941	0.926
HIST2H2AA4	0.901	0.887	0.916	0.901	0.921	0.901
HBB	0.823	0.289	0.403	0.812	0.819	0.823
HBA2	0.846	0.583	0.674	0.838	0.844	0.844
HBA1	0.890	0.775	0.830	0.887	0.898	0.890

Table 4: Zero ratio estimation for cells from Cluster 1 and 2, 10 individuals.

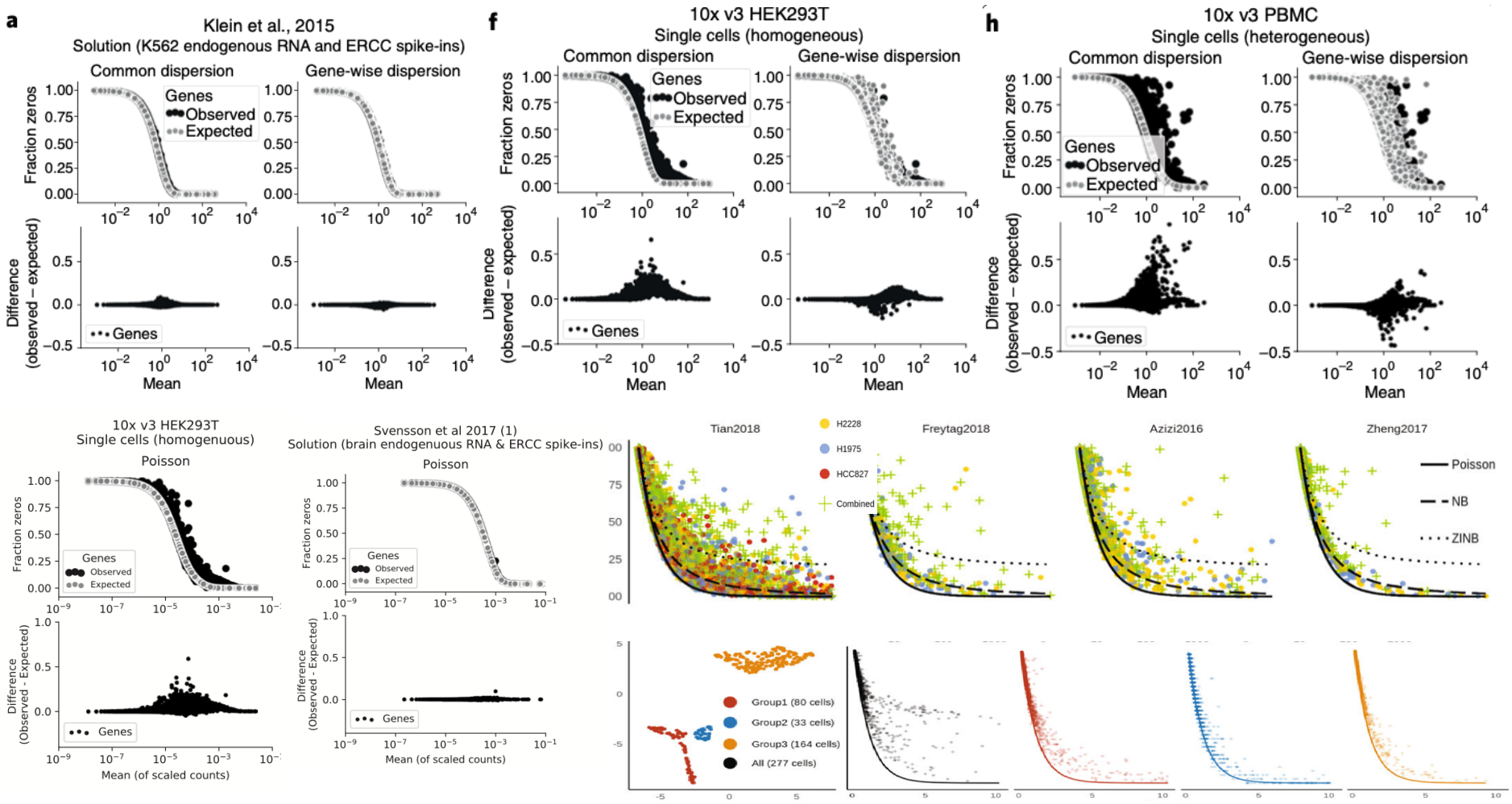


Figure 5: Observed and expected zeros with NB, Poisson or Poisson with scaled mean. Grey figures from [6]; color figures from [7].

Modeling scRNAseq Using Negative Binomial

- We need the dispersion ϕ for each gene to model the variance of gene expression due to heterogeneity. In reality, we cannot get a group of homogeneous cells[1]. Meanwhile, we can estimate ϕ in each gene by sharing the same prior since we understand it will reflect the character of a given cell sub-population.
- Like DESeq2 [8] for bulkRNAseq DE analysis, we model the mean μ across different conditions and different samples.
- We use Bayesian modeling since it could be more robust than usual MLE estimation. The posterior distribution about the parameters can tell us lots of information.

[1] Cell annotation itself is now a really hot topic in scRNAseq data analysis.

Outline

Negative Binomial Generalized Linear Model

Let X_{ij} represent the UMI counts for the gene i in the cell j

$$X_{ij} \sim NB(\mu_{ij}, \phi_i) \tag{9}$$

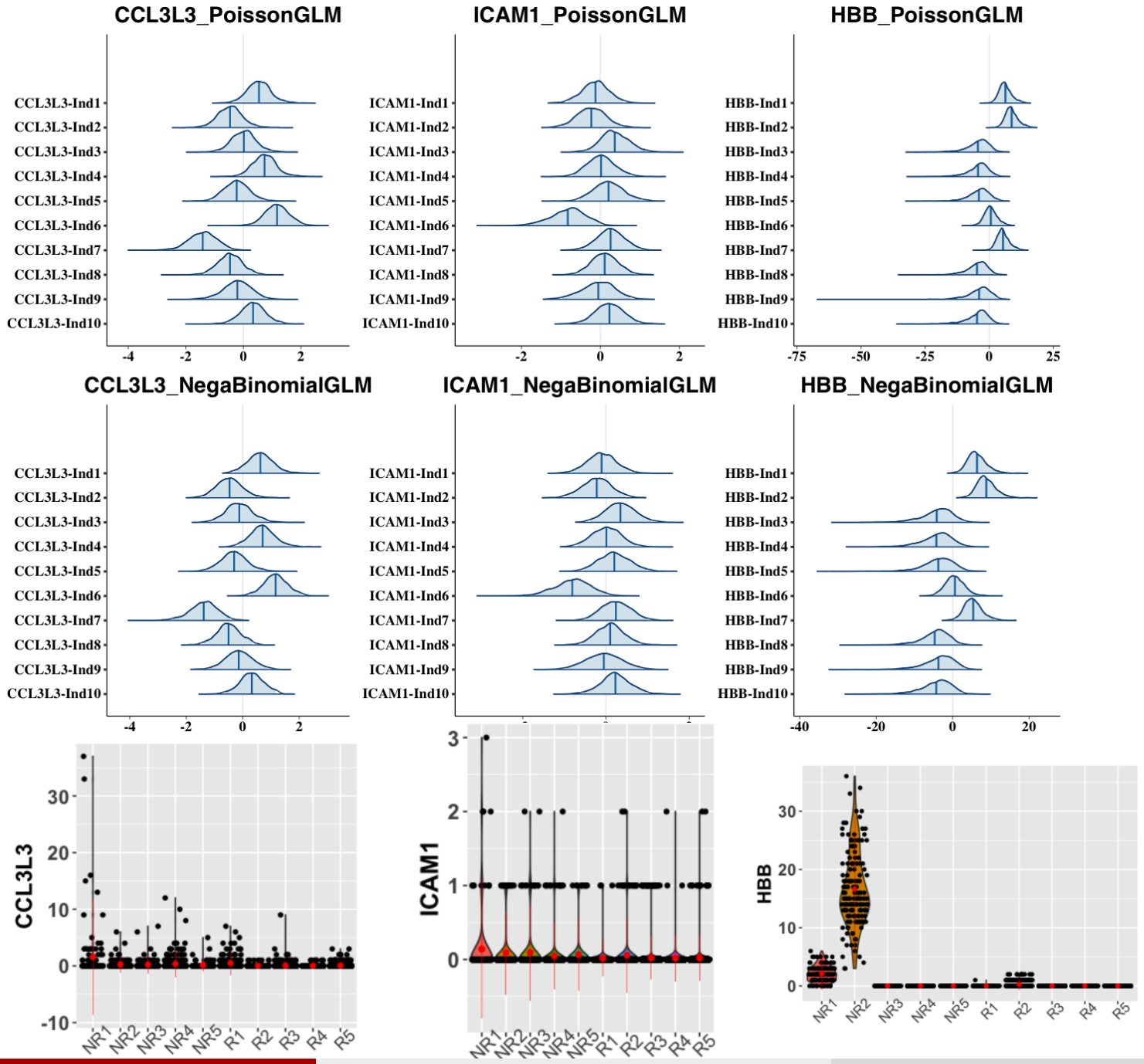
$$\mu_{ij} = S_j * \lambda_{ij} \tag{10}$$

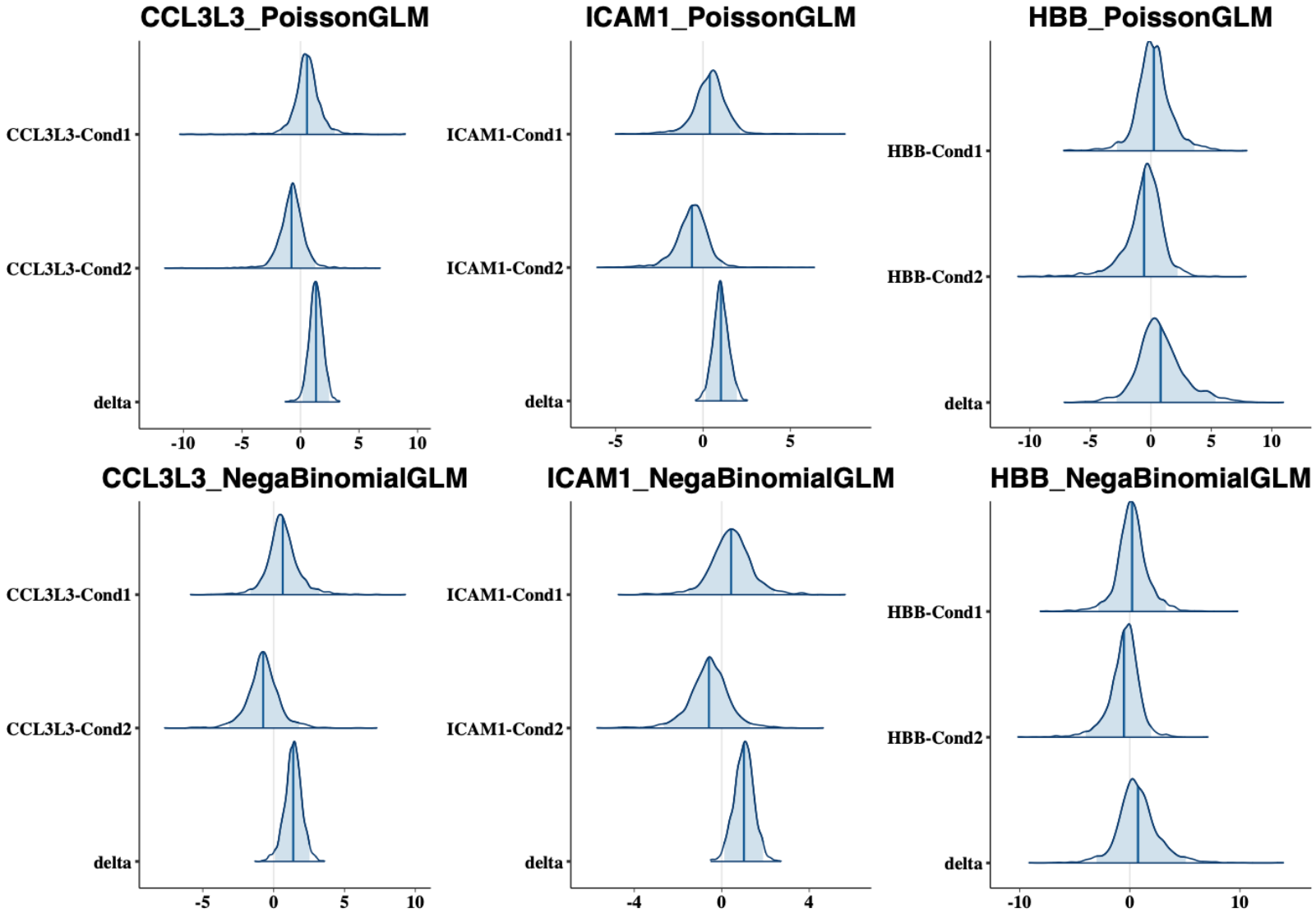
$$\log(\lambda_{ij}) = \sum_r d_{jr}\beta_{ir} \tag{11}$$

$$\phi_i \sim Gamma(\alpha_0, \beta_0) \tag{12}$$

$$\boldsymbol{\beta_r} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_0^2 & & & & & \\ & \sigma_{ind}^2 & & & & \\ & & \ddots & & & \\ & & & \sigma_{ind}^2 & & \\ & & & & \sigma_{cond}^2 & \\ & & & & & \sigma_{cond}^2 \end{bmatrix}\right)$$

$$\mathbf{d_j} = \left(1, \overbrace{(0, 1, \ldots, 0, 0)}, \underbrace{(1, 0)}\right)$$


CCL3L3_PoissonGLM
ICAM1_PoissonGLM
HBB_PoissonGLM
CCL3L3_NegaBinomialGLM
ICAM1_NegaBinomialGLM
HBB_NegaBinomialGLM
CCL3L3-Ind1
CCL3L3-Ind2
CCL3L3-Ind3
CCL3L3-Ind4
CCL3L3-Ind5
CCL3L3-Ind6
CCL3L3-Ind7
CCL3L3-Ind8
CCL3L3-Ind9
CCL3L3-Ind10
ICAM1-Ind1
ICAM1-Ind2
ICAM1-Ind3
ICAM1-Ind4
ICAM1-Ind5
ICAM1-Ind6
ICAM1-Ind7
ICAM1-Ind8
ICAM1-Ind9
ICAM1-Ind10
HBB-Ind1
HBB-Ind2
HBB-Ind3
HBB-Ind4
HBB-Ind5
HBB-Ind6
HBB-Ind7
HBB-Ind8
HBB-Ind9
HBB-Ind10
CCL3L3
ICAM1
HBB
NR1
NR2
NR3
NR4
NR5
R1
R2
R3
R4
R5



t statistic: (Pois, NB)

2.256, 2.268 **2.307, 2.249** **0.412, 0.386**

Pseudo bulk with DESeq2: log2fc (p-value)

2.113 (5.8e-4) **1.636 (3.4e-5)** **6.708 (2.03e-2)**

Outline

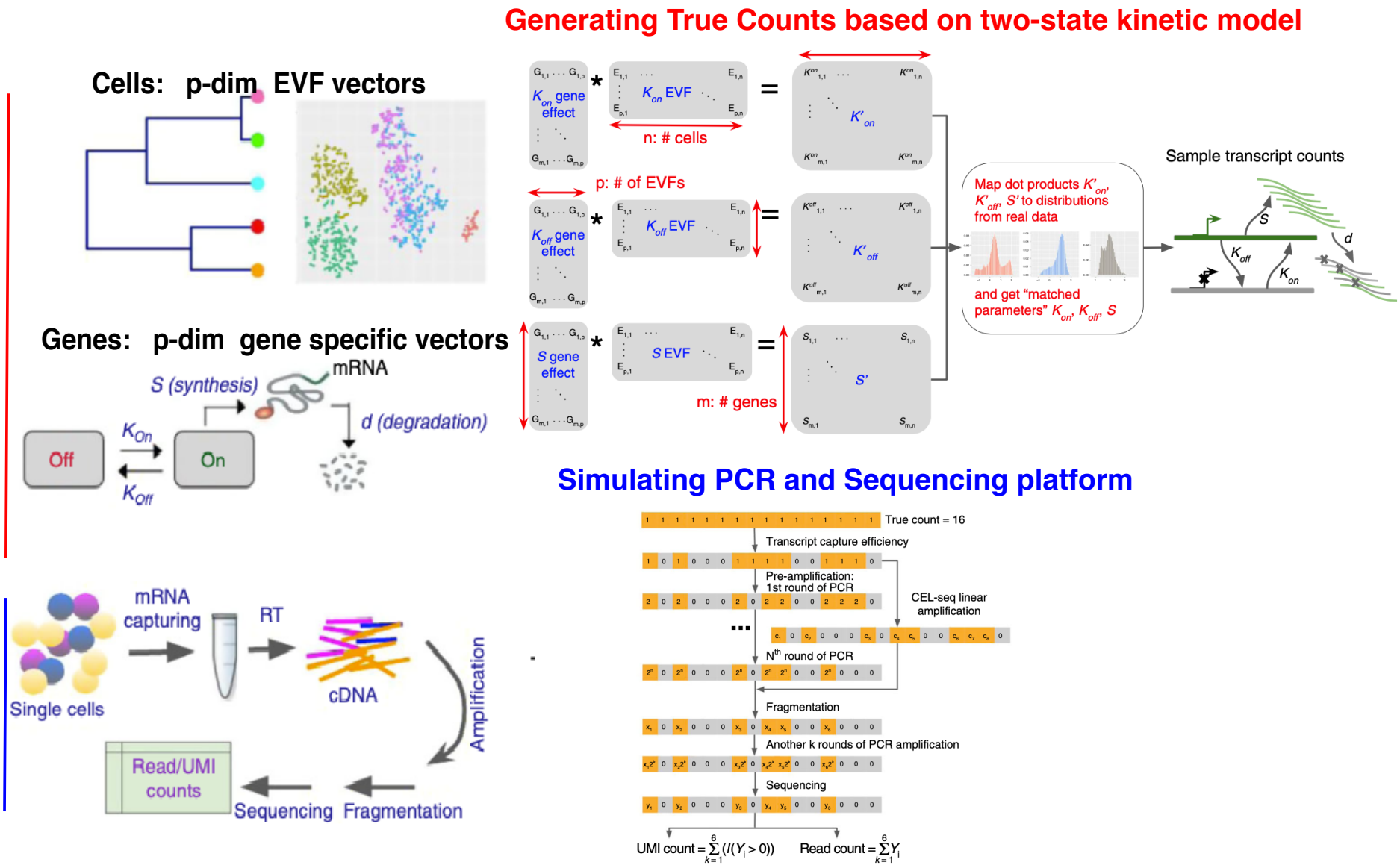


Figure 6: SymSim Overview ([9])

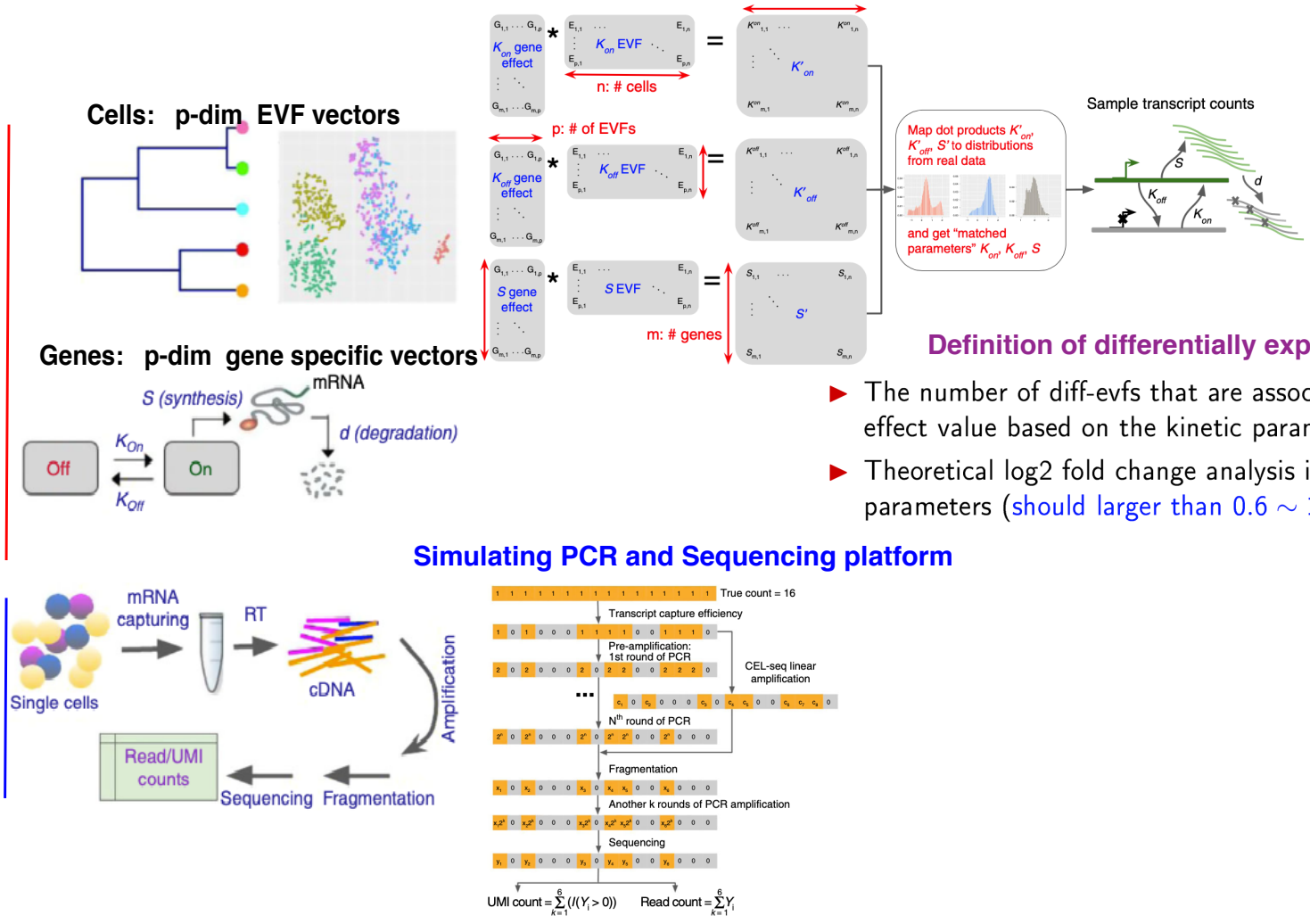
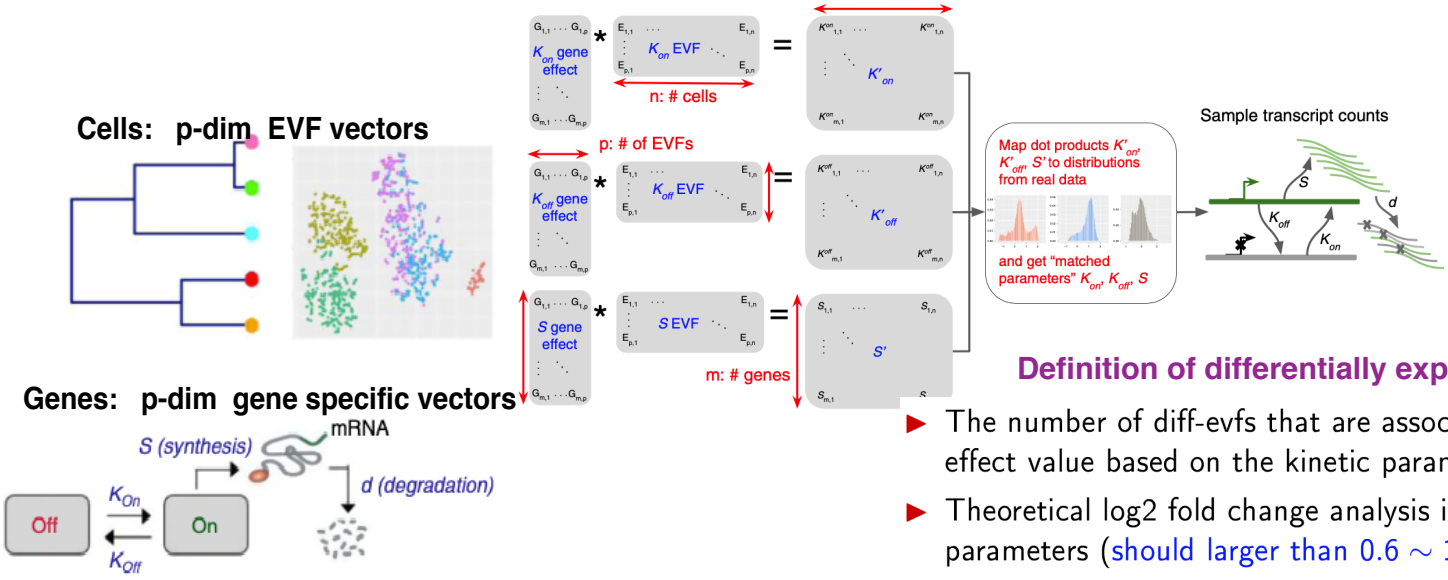


Figure 7: DE analysis in SymSim

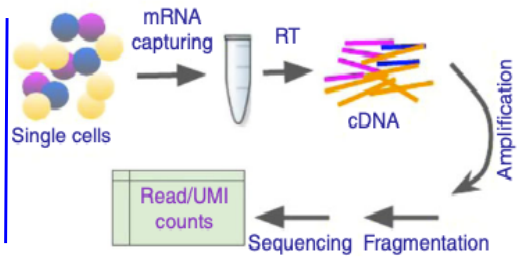
Generating True Counts based on two-state kinetic model

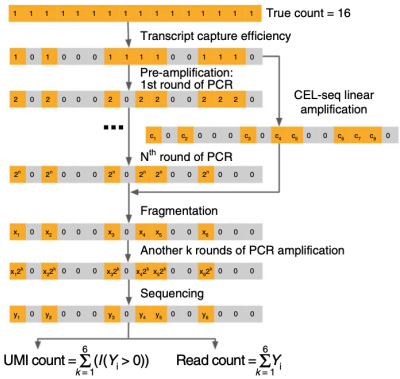


Definition of differentially expressed (DE) genes.

- The number of diff-evfs that are associated with a non-zero gene effect value based on the kinetic parameters (should larger than zero)
- Theoretical log2 fold change analysis is based on the kinetic parameters (should larger than 0.6 $\sim$ 1).

Simulating PCR and Sequencing platform





Batch Effect Added on Counts

- Each cell c is assigned to the batch b randomly as b_c
- Given the user-defined batch effect size (be, default as 1), SymSim will firstly samples $m_{gb} \sim \mathcal{N}(0, 1)$; then update it by sampling $\hat{m}_{gb} \sim \mathcal{U}(m_{gb} - be, m_{gb} + be)$.
- Then for cell c and gene g, let $m_{gc} \sim \mathcal{N}(m_{gb_c}, 0.01)$
- Lastly, the observed count $\hat{o}_{gc} = o_{gc} * 2^{m_{gc}}$ (i.e., added in the log_2 level).

Figure 8: Batch effect in SymSim

Simulation based on SymSim

- 2000 cells, 300 genes, 10 batch (individuals)
  - Each individual (ind): about 200 cells
  - Condition 1: Ind 1 to 5 *v.s.* Condition 2: Ind 6 to 10
  - Use the two-leaf phylo-tree, which lets SymSim to generate two cell populations as one cell type in two conditions in MSSC.
  - 1000 cells in each condition
- Batch effect added in two sequential steps:
  1. The default SymSim strategy to add both gene and individual-specific batch effect.
  2. Sample some non-differential expressed genes, add strong batch effect sizes for 1, 2 and 6 to simulate the biased response. *For instance, in our simulation, 71 DE genes, and 34 strictly non-DE genes. We sample 20 genes from the strictly non-DE genes in the second step.*

DE genes when no batch effect

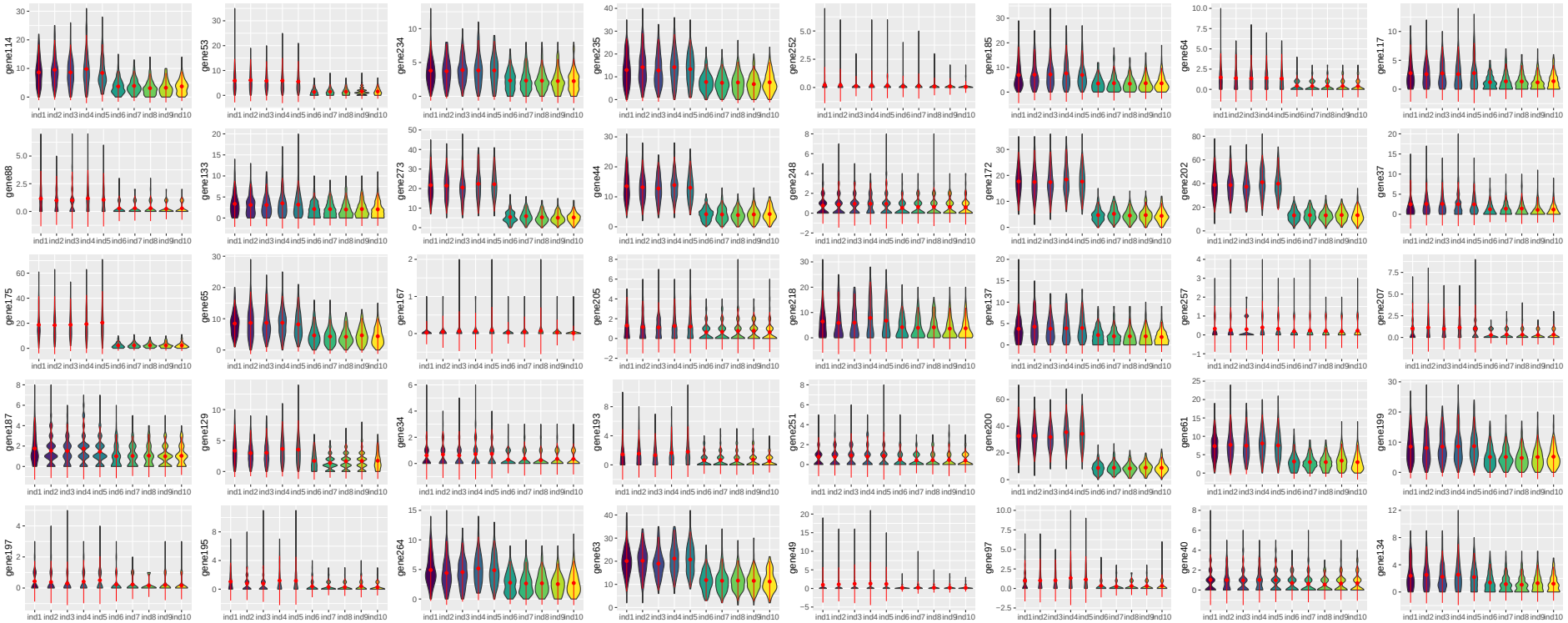


Figure 9: Before batch effect: DE genes

non-DE genes when no batch effect

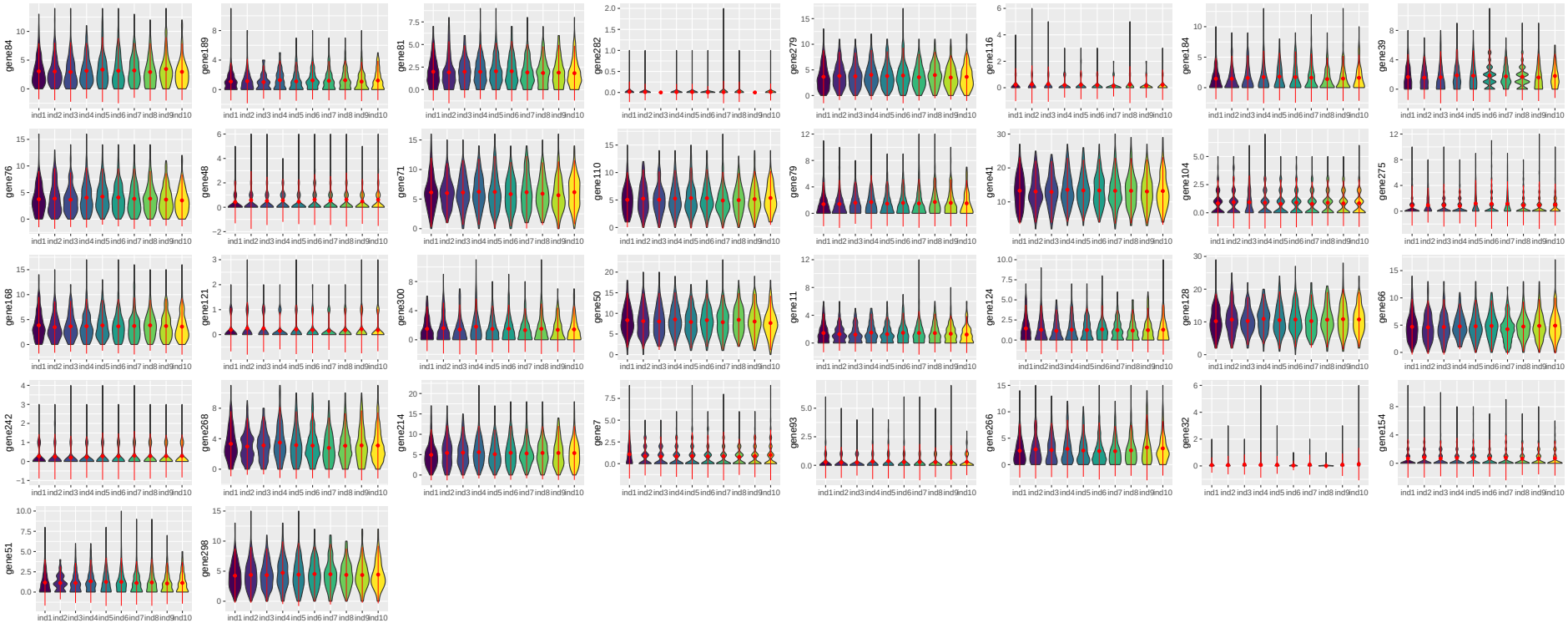


Figure 10: Before batch effect: non-DE genes

DE genes after the first batch effect

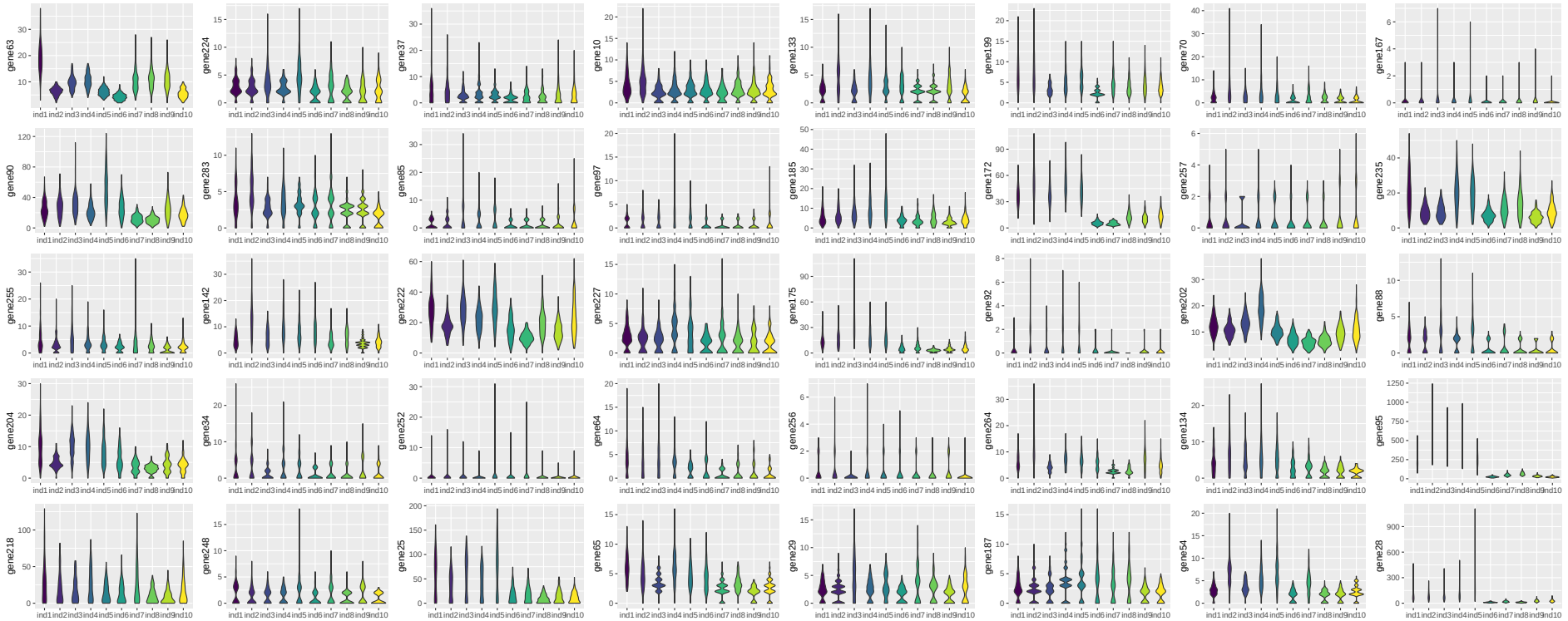


Figure 11: First batch effect: DE genes

Non-DE genes after the first batch effect

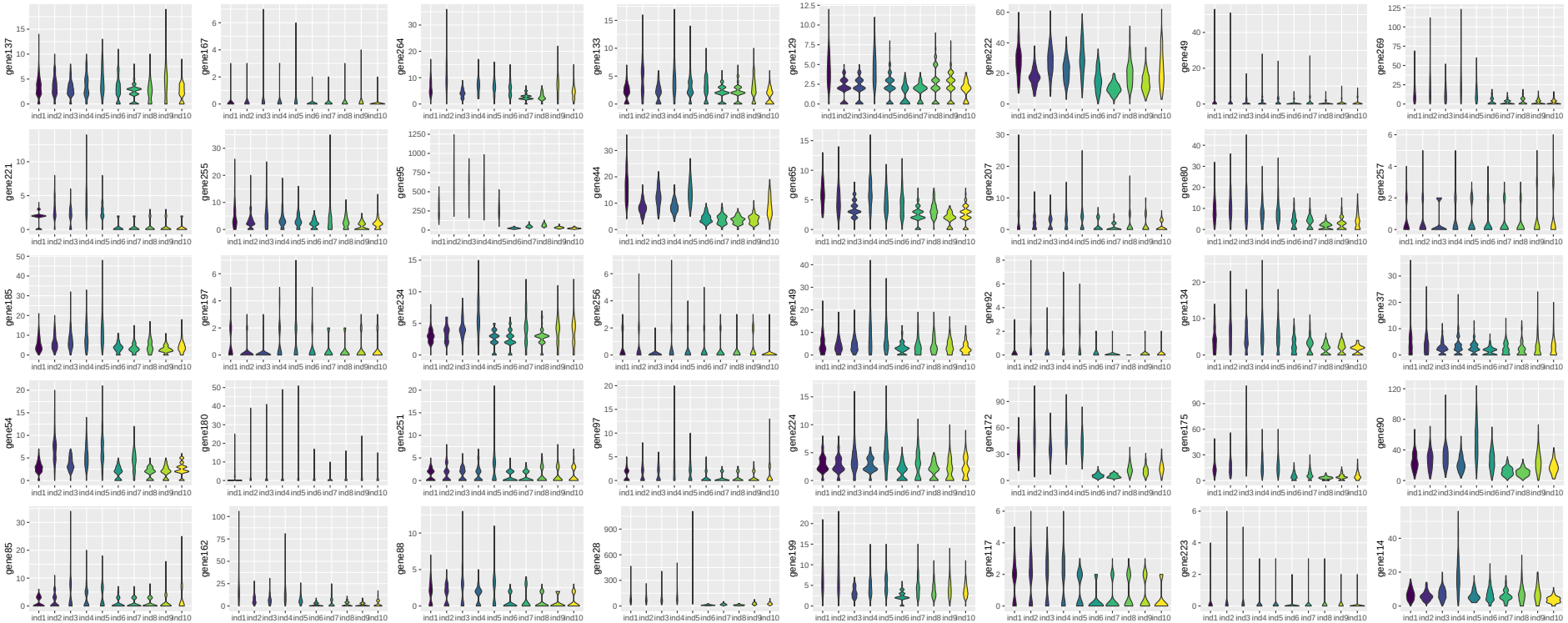


Figure 12: First batch effect: non-DE genes

DE genes after the second batch effect

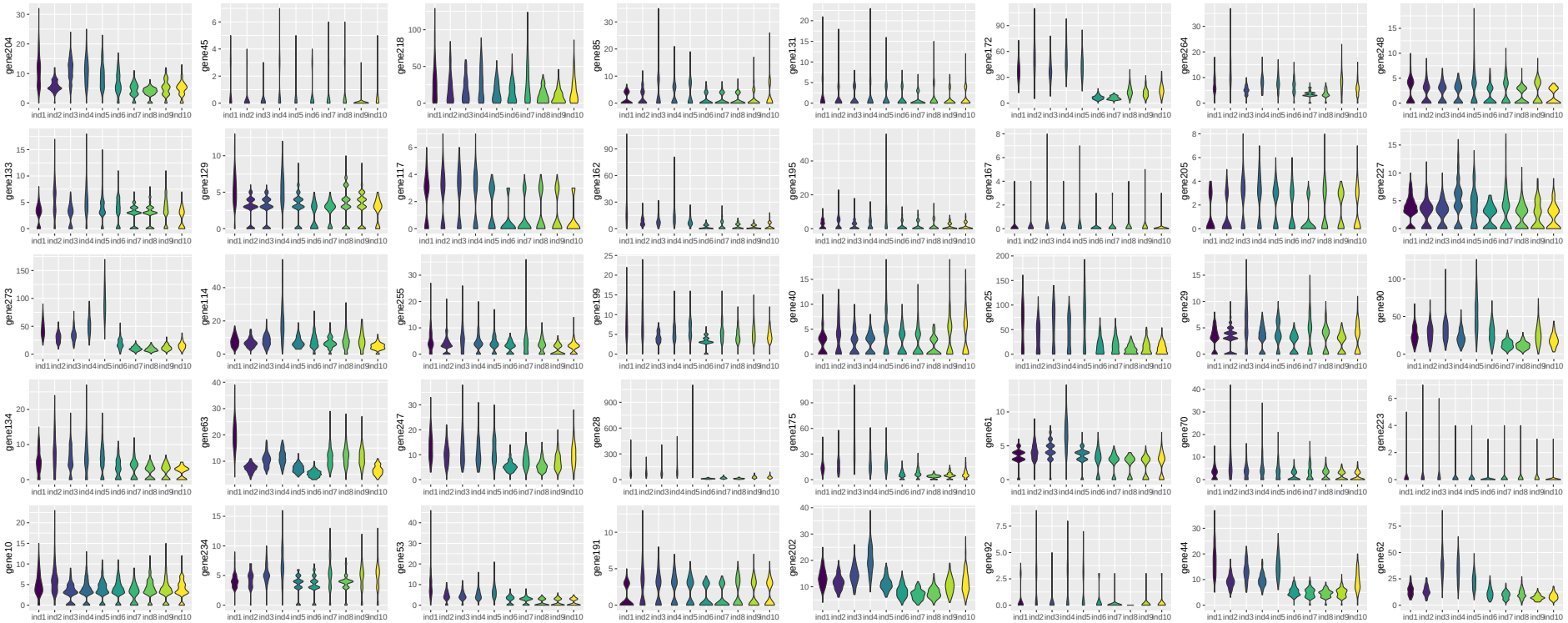


Figure 13: Second batch effect: DE genes

Non-DE genes after the second batch effect

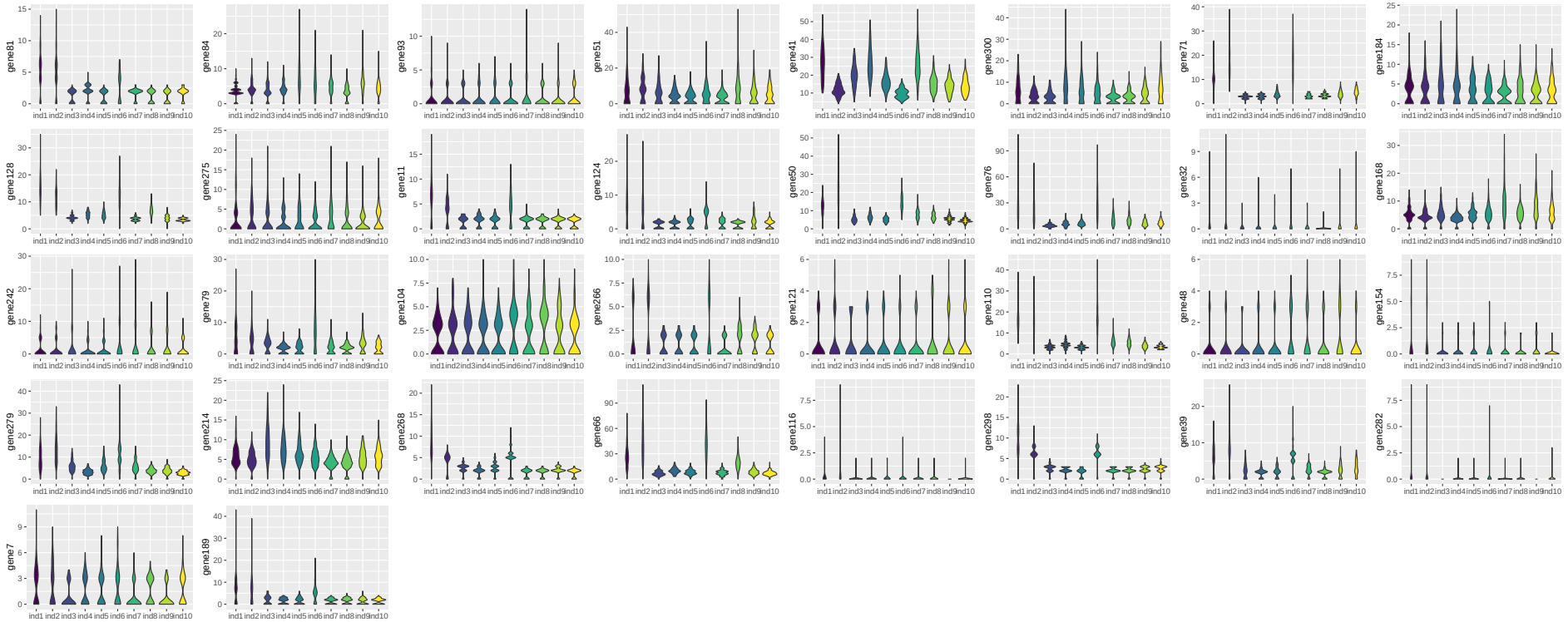


Figure 14: Second batch effect: non-DE genes

Result Summary based on the AUC

- Both MCMC and VI shows almost the same result.
- Model v1-1 and v1-2 shows almost the same result, but v1-2 is more robustness than v1-1.
- On the simulation data set, pseudobulk is better than mssc (0.75 vs 0.72).

Posterior t statistics vs p-value from DESeq2

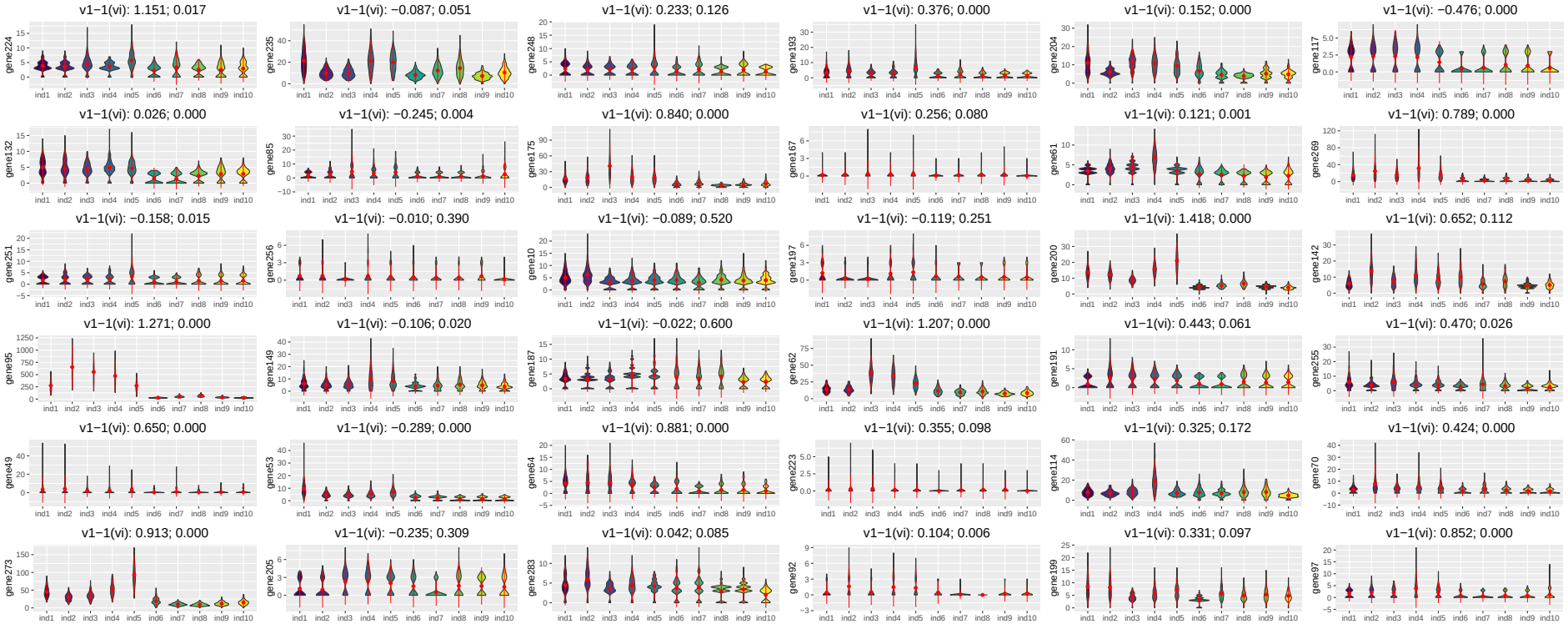


Figure 15: DE genes: mssc vs pseudo

Posterior t statistics vs p-value from DESeq2

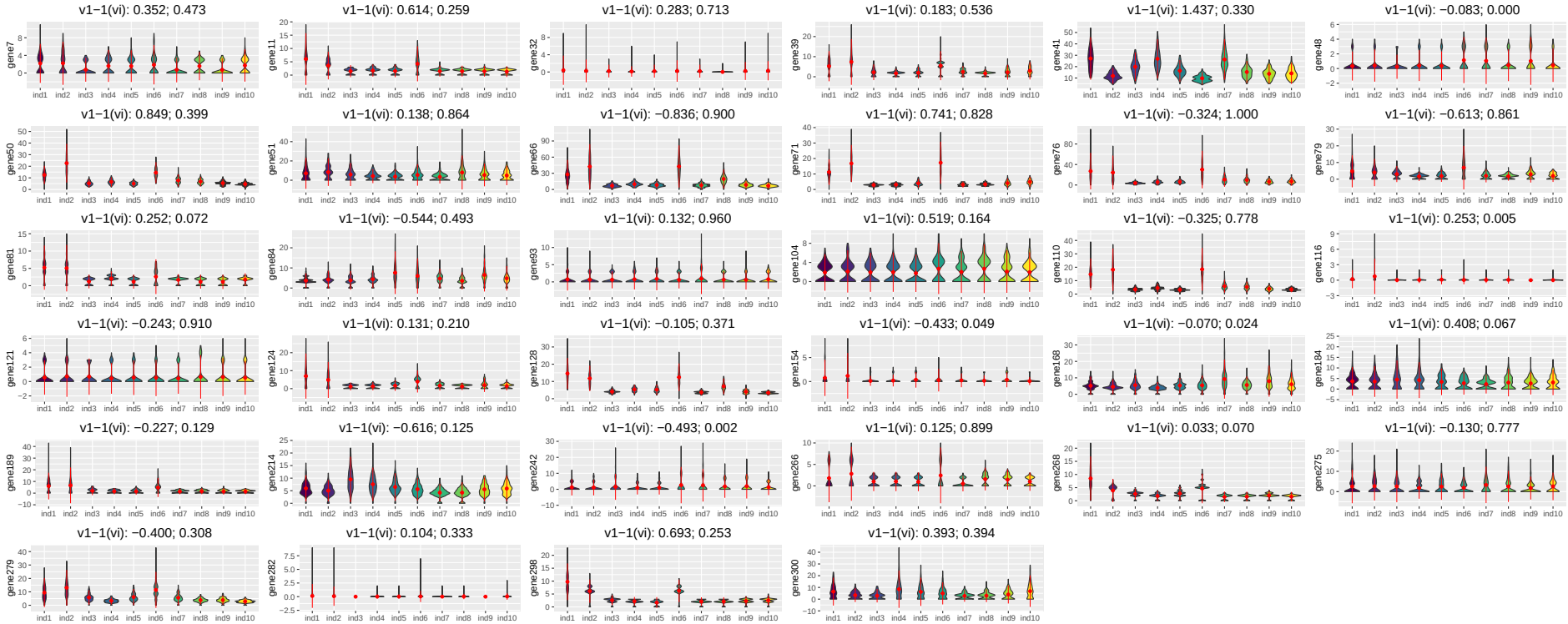
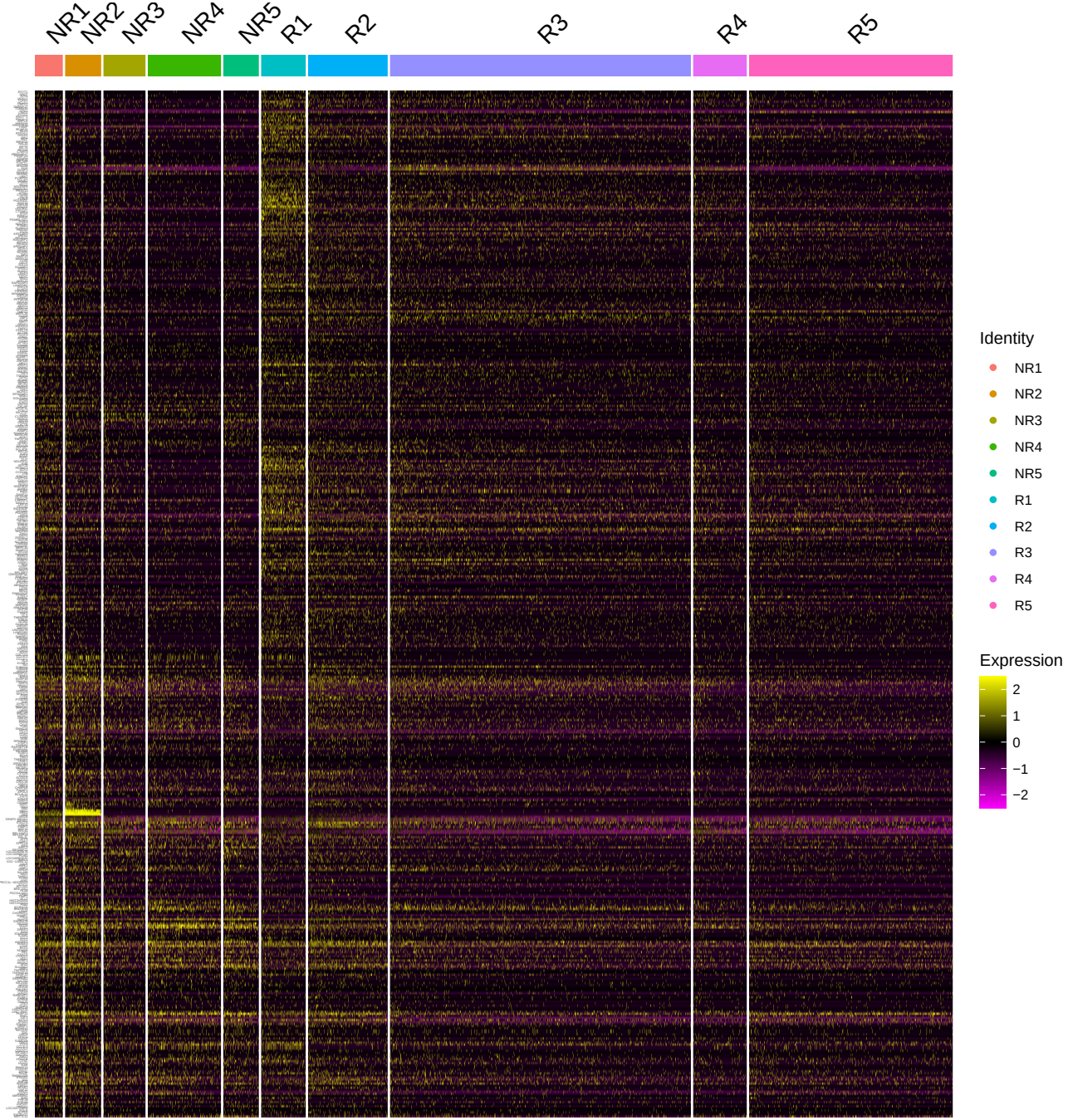


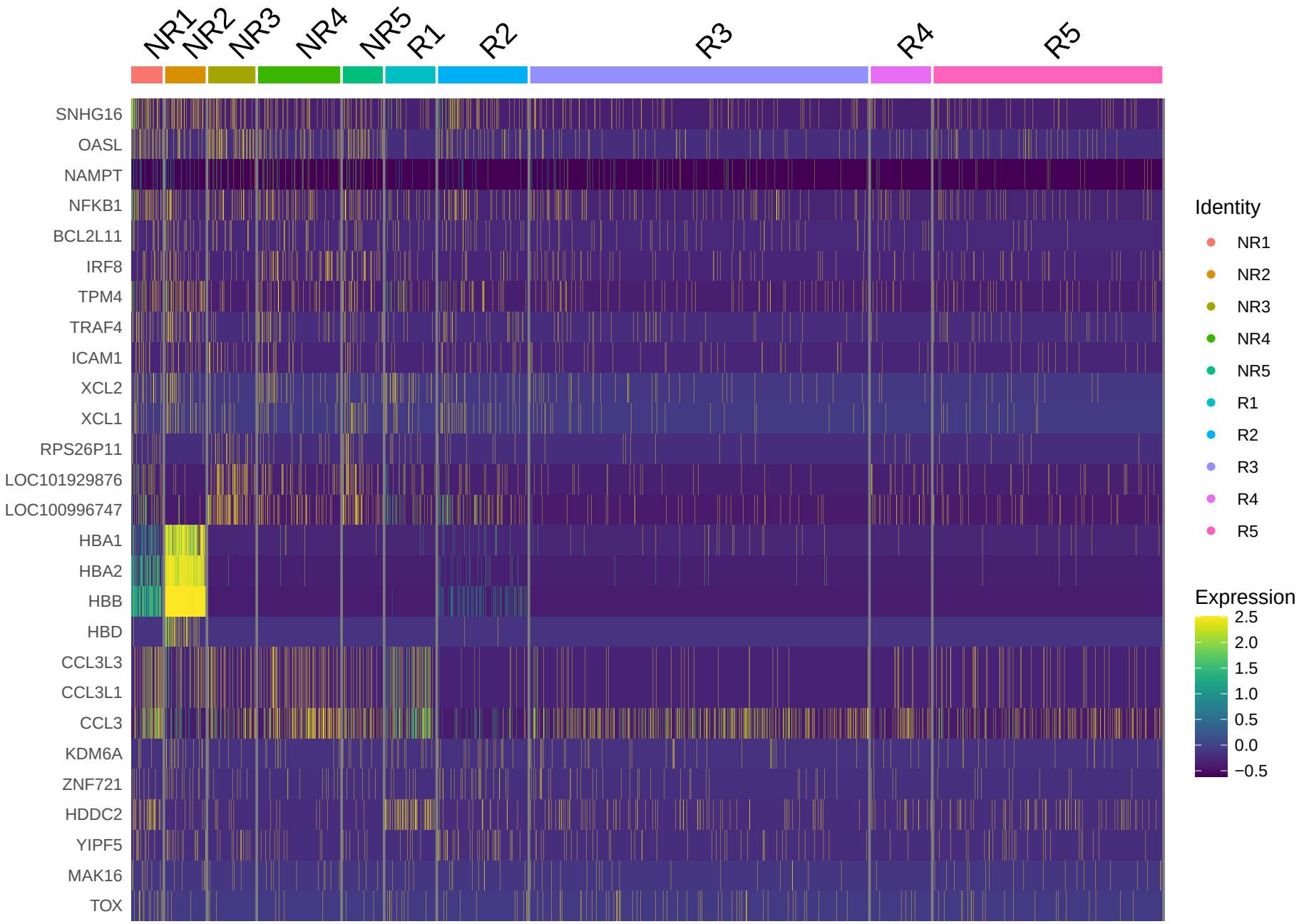
Figure 16: Non-DE genes: mssc vs pseudo

Outline

DataSet

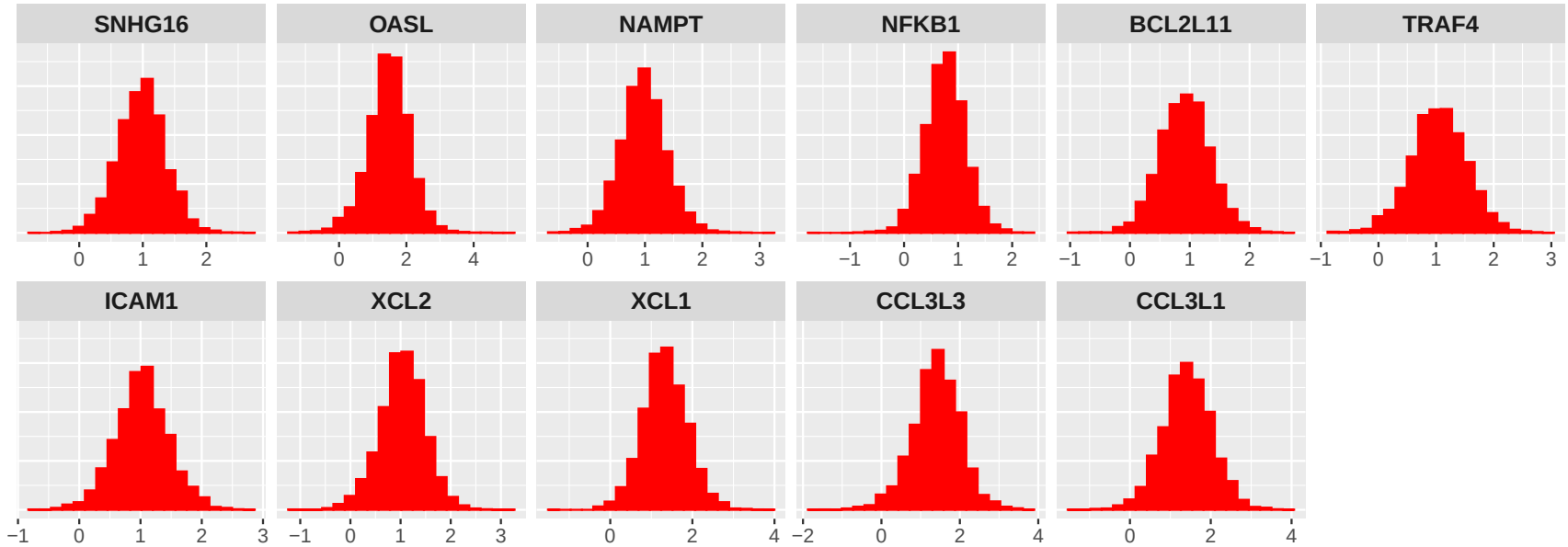
- UMI-based scRNAseq Data: 5 cases vs. 5 controls
  - In total, 26,000 cells (over 2,000 cells / individual).
  - Identify cell types: using Harmony; choose cluster 2, which is cytotoxic T-cells, with 3,885 cells.
- Select 27 genes from both top-ranked and low-ranked genes based on pseudo bulk analysis with DESeq2.
  - Gene modules by hierarchical clustering of genes on cluster 2.
    - Group1: ICAM1(DESeq2 Rank 6), XCL2 (DESeq2 Rank 13), XCL1 (DESeq2 Rank 43)
    - Group2: RPS26P11 (DESeq2 Rank 15), LOC101929876 (DESeq2 Rank 22),LOC100996747 (DESeq2 Rank 97)
    - Group3: HBA1 (DESeq2 Rank 14), HBA2, HBB, HBD
    - Group4: CCL3L3 (DESeq2 Rank 99), CCL3L1 (DESeq2 Rank 102), CCL3
  - Top-ranked genes: SNHG16,OASL,NAMPT,NFKB1,BCL2L11,IRF8,TPM4,TRAF4
  - Low-ranked genes: KDM6A,ZNF721,HDDC2,YIPF5,MAK16, TOX


NR1
NR2
NR3
NR4
NR5
R1
R2
R3
R4
R5
Identity
NR1
NR2
NR3
NR4
NR5
R1
R2
R3
R4
R5
Expression
2
1
0
−1
−2


NR1
NR2
NR3
NR4
NR5
R1
R2
R3
R4
R5
SNHG16
OASL
NAMPT
NFKB1
BCL2L11
IRF8
TPM4
TRAF4
ICAM1
XCL2
XCL1
RPS26P11
LOC101929876
LOC100996747
HBA1
HBA2
HBB
HBD
CCL3L3
CCL3L1
CCL3
KDM6A
ZNF721
HDDC2
YIPF5
MAK16
TOX
Identity
NR1
NR2
NR3
NR4
NR5
R1
R2
R3
R4
R5
Expression
2.5
2.0
1.5
1.0
0.5
0.0
−0.5

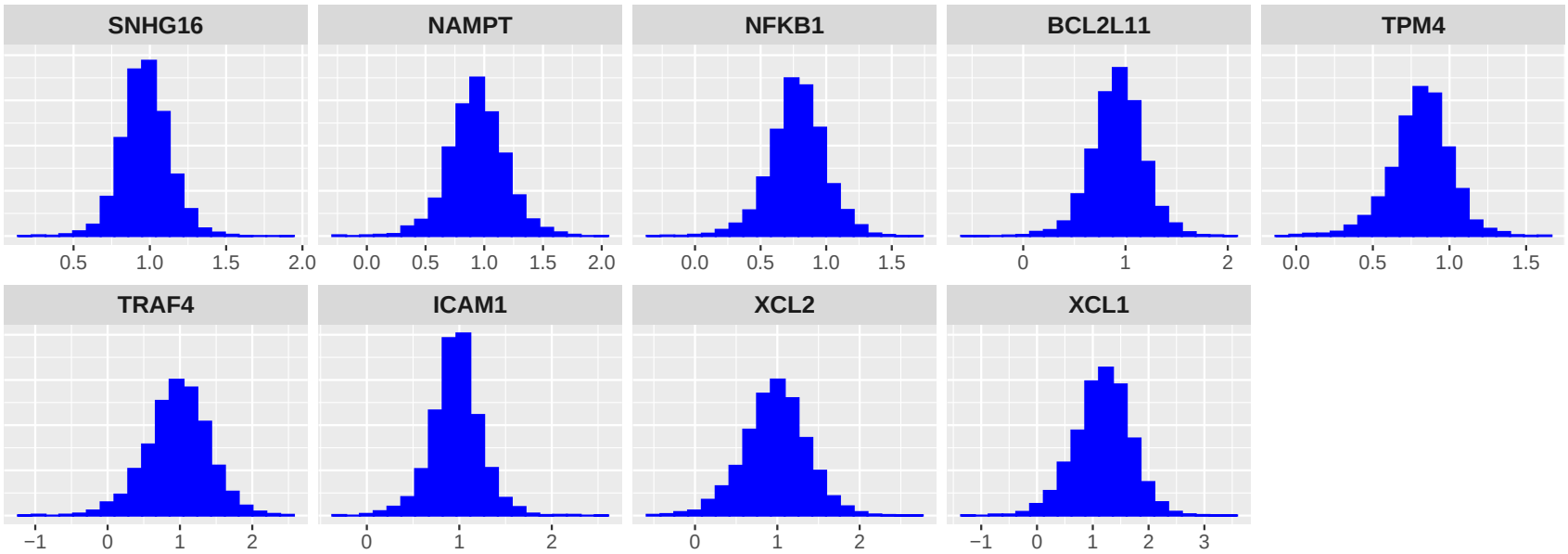
Differential genes detected by modeling gene−wise invididual effect indepdently

Histogram of the differences of log fold changes between control and case.



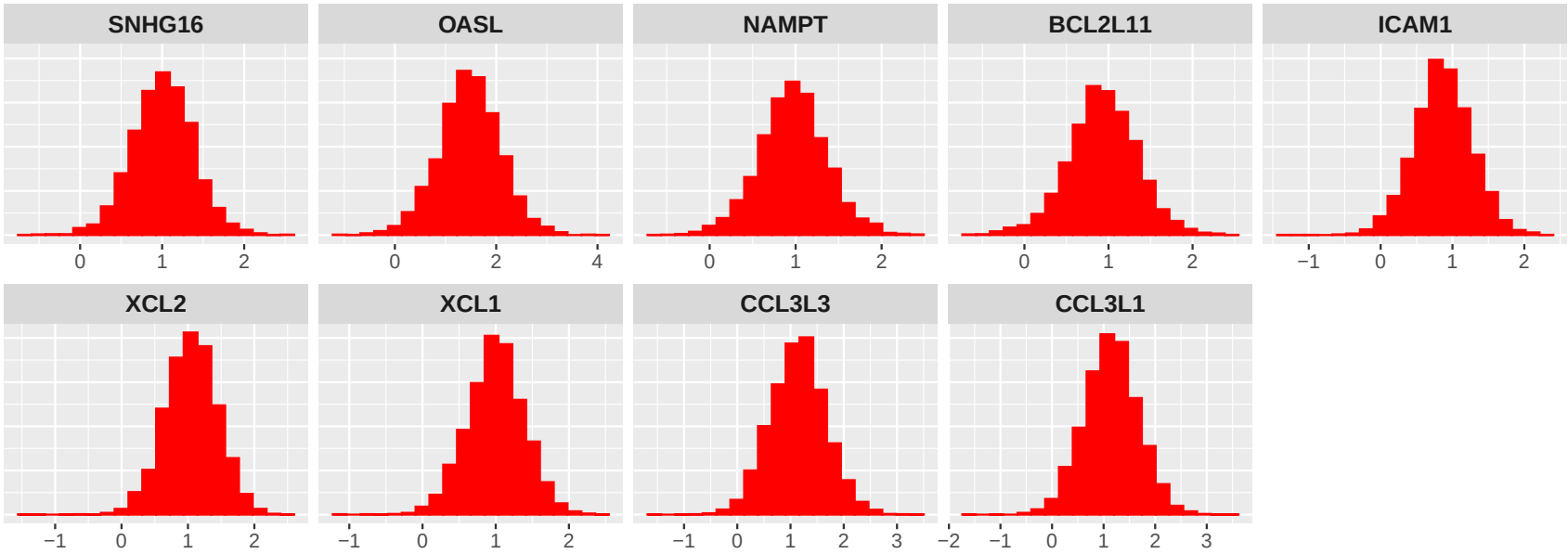
Differential genes detected by modeling gene−wise invididual effect, and sharing hyper prior for variances.

Histogram of the differences of log fold changes between control and case.



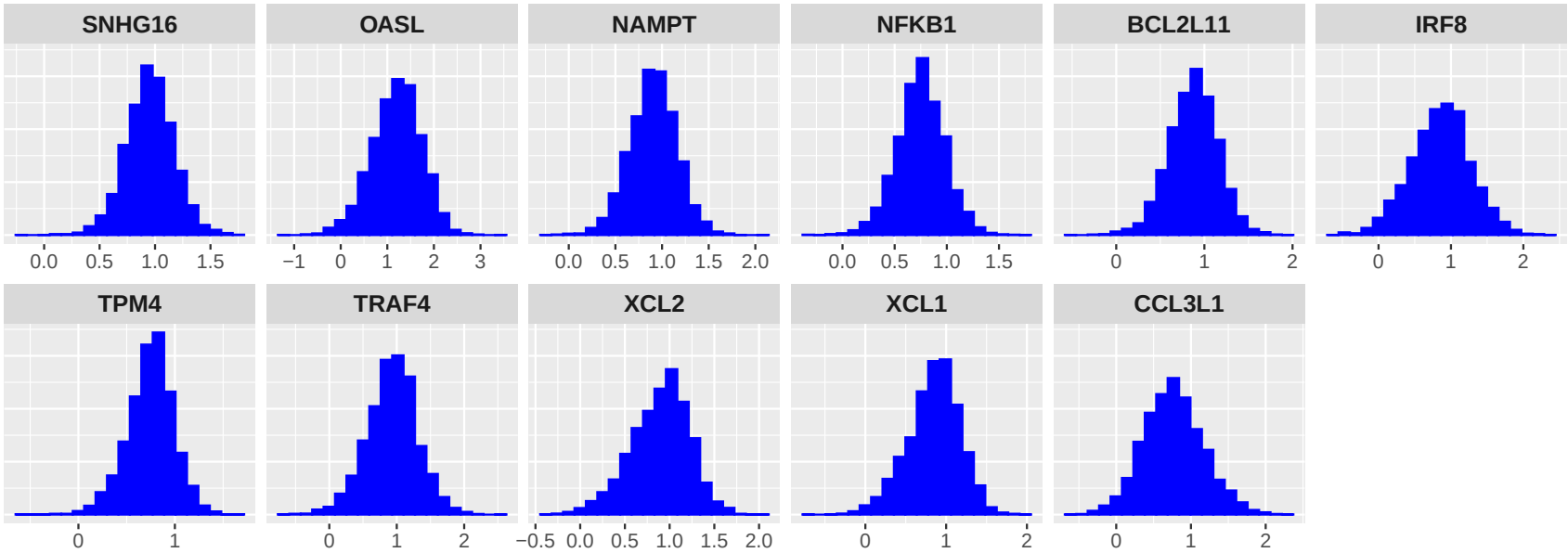
Differential genes detected by modeling gene−module invididual effect indepdently

Histogram of the differences of log fold changes between control and case.



Differential genes detected by modeling gene−module invididual effect, and sharing hyper prior for variances.

Histogram of the differences of log fold changes between control and case.



Simulation limitations

- Simulation: current batch effect is added after count generation, while in our model, we model batch effects on gene expression level.
- Gene modules: batch effect estimation might be hard. So how about consider two gene modules: one is for the false positive ones, and the other is for the positive ones. Then let each of them share the same batch effect. And we estimate the batch effect on gene-module level.
- Pseudobulk analysis might over estimate the genes if the genes have lots of zero counts.

Droplet scRNAseq is not zero-inflated [6]

- The original "dropout" problem of yielding an inflation of zero values in scRNA-seq data was in the description of single-cell differential expression (SCDE) [10], which was on low-throughput plate-based methods, like Smart-seq and STRT-seq.
- Investigate the number of zeros value using negative control data (like endogenous RNA, ERCC spike-ins) with no biological variation.
- When scaling the counts by total counts in cells, even a Poisson distribution can explain the fraction of zeros well.
- Heterogeneous cells showed bad fittings, i.e., more zeros than expected.
- Suggest that additional zero values in biological data are likely due to biological variation.

Demystifying drop-outs in single-cell UMI data [7]

- They observe that most drop-outs disappear once cell-type heterogeneity is resolved. The simple Poisson distribution is then sufficient to fully leverage the biological information in the UMI data.
- Zero proportions are effective measures for cell-type heterogeneity.
- Sequencing depths are confounded with cell types and size factor-based adjustment can obscure biological information.
- Proposed a zero ratio based t-test to select the zero-inflated genes as cell features.

By involving cell-specific gene module expressions, we can estimate $\boldsymbol{B}$ from the data.

$$
\begin{aligned}
X_{ijk}^{g} &\sim Poisson(S_{ijk} \cdot \lambda_{ijk}^{g}) \\
\ln \lambda_{ijk}^{g} &= \mu^{g} + \mu_{i}^{g} + \mu_{k}^{g} + \mu_{j}^{g} \\
\mu^{g} &\sim \mathcal{N}(\mu, \sigma^{2}) \\
\mu_{i}^{g} &= \boldsymbol{b}_{g}^{t} \cdot \boldsymbol{f}_{i}, \quad \boldsymbol{b}_{g}, \boldsymbol{f}_{i} \in \mathcal{R}^{p \times 1} \\
\boldsymbol{f}_{i} &\sim \mathcal{MVN}(\boldsymbol{f}_{cell}, diag(\boldsymbol{\sigma}_{cell}^{2}) \cdot \boldsymbol{I}) \\
\mu_{k}^{g} &= \boldsymbol{b}_{g}^{t} \cdot \boldsymbol{f}_{k}, \quad \boldsymbol{f}_{k} \in \mathcal{R}^{p \times 1} \\
\boldsymbol{f}_{k} &\sim \mathcal{MVN}(\boldsymbol{f}_{ind}, diag(\boldsymbol{\sigma}_{ind}^{2}) \cdot \boldsymbol{I}) \\
\mu_{j}^{g} &\sim \mathcal{N}(\mu_{cond_0}, \sigma_{cond_0}^{2})
\end{aligned}
$$

[1] Michael A Durante, Daniel A Rodriguez, Stefan Kurtenbach, Jeffim N Kuznetsov, Margaret I Sanchez, Christina L Decatur, Helen Snyder, Lynn G Feun, Alan S Livingstone, and J William Harbour. Single-cell analysis reveals new evolutionary complexity in uveal melanoma. *Nature communications*, 11(1):1–10, 2020.

[2] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, pages 1–8, 2019.

[3] Katharina T Schmid, Cristiana Cruceanu, Anika Boettcher, Heiko Lickert, Elisabeth B Binder, Fabian J Theis, and Matthias Heinig. Design and power analysis for multi-sample single cell genomics experiments. *bioRxiv*, 2020.

[4] Kobe C Yuen, Li-Fen Liu, Vinita Gupta, Shravan Madireddi, Shilpa Keerthivasan, Congfen Li, Deepali Rishipathak, Patrick Williams, Edward E Kadel, Hartmut Koeppen, et al. High systemic and tumor-associated il-8 correlates with reduced clinical benefit of pd-l1 blockade. *Nature Medicine*, 26(5):693–698, 2020.

[5] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):1–16, 2019.

[6] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[7] Tae Kim, Xiang Zhou, and Mengjie Chen. Demystifying" drop-outs" in single cell umi data. *bioRxiv*, 2020.

[8] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[9] Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, 10(1):1–16, 2019.

[10] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[11] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[12] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[13] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.