

# **Communicating scRNA-seq**

**Valentine Svensson, Serqet Therapeutics**

# Serqet Therapeutics

## Reconceiving drug development for complex disease

# Serqet Therapeutics

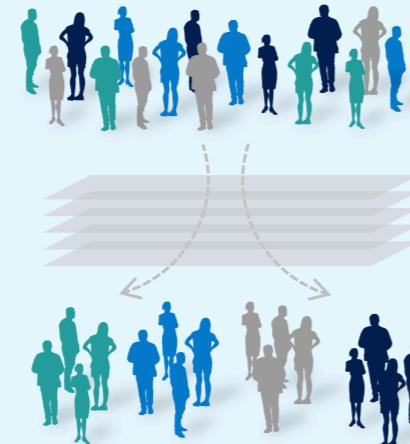
## Integrating Discovery

We are integrating computational and biological approaches to find novel genetically-based, shared biology in complex polygenic diseases.

### Inputs

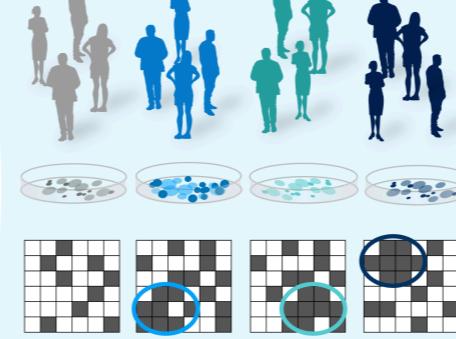
- Clinical 'Big Data'
- Genetic information
- Patient Cells and Samples

### PHENO-clusters



Analytical identification of at-risk populations into clusters

### GENO-clusters



Confirming genetically different clusters by rejecting null hypothesis

### BIO-clusters



Pheno-emulation for discovery of cluster specific biology and pathways

### Outputs

Patient clusters whose clinical phenotype is driven by a shared polygenic genomic architecture

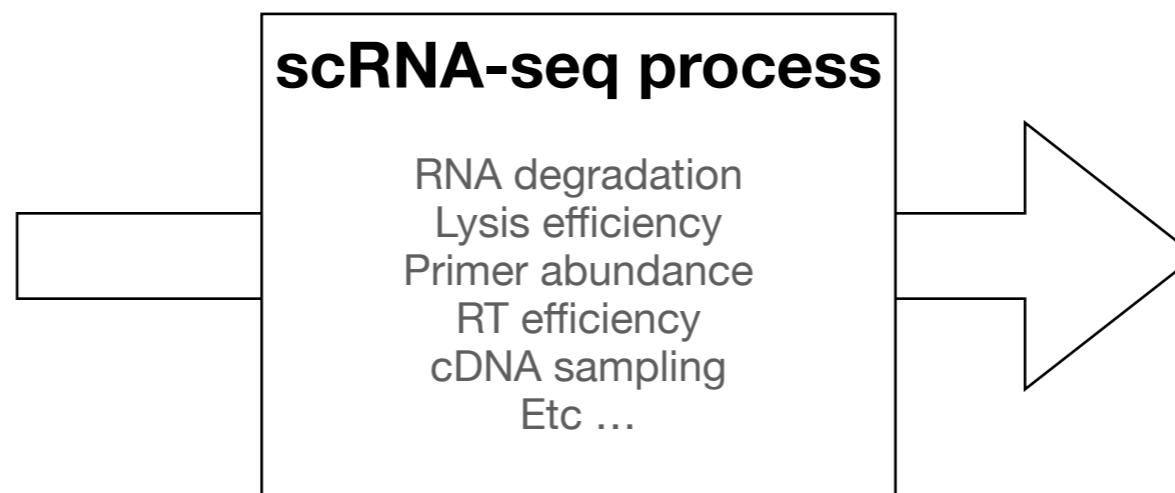
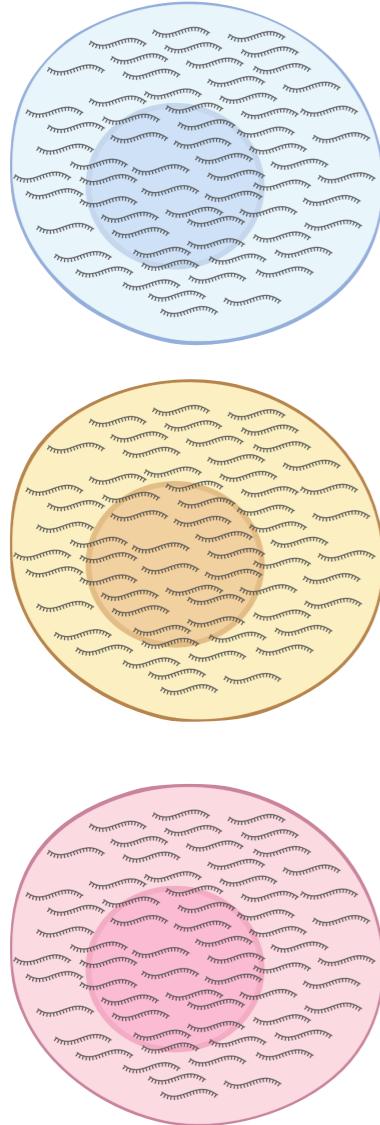
# Units and scales

## Some problems

- Figures in papers often don't report expression scale
- When discussing 'high' and 'low' expression, what do we mean?
- If an expression level is reported in a paper, how does that relate to my data?

# Single cell RNA-seq

## What are we measuring?



- Need to scale molecule counts per gene by the total number of molecules. (Unit: *Rate* or *fraction* of molecules from a gene)

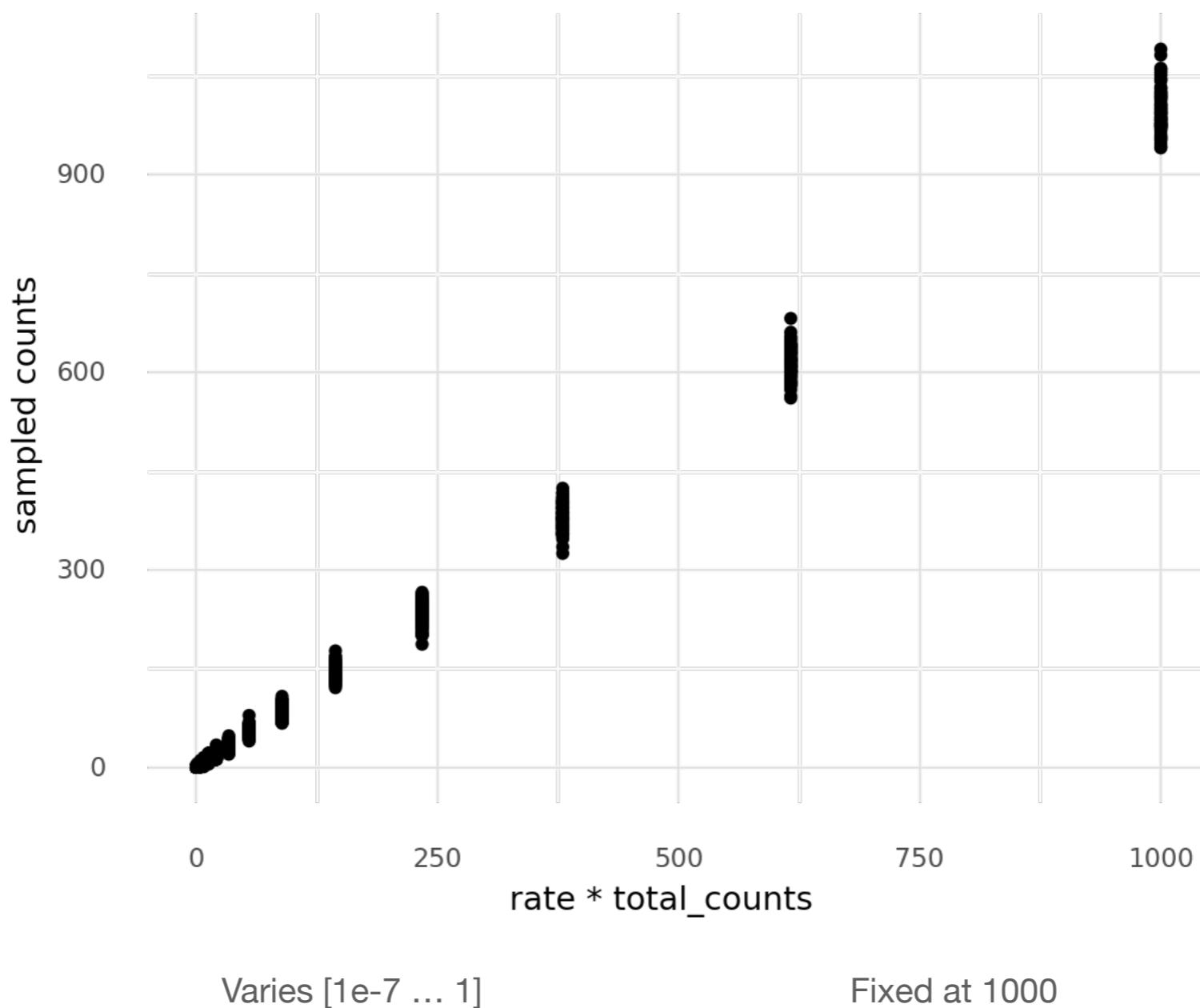
# Units and scales

- Rates are fractions -> effects are multiplicative
- Look at fraction on a log scale
- **Rate:** for each molecule we see, what is the probability it comes from gene X?
- Count / total\_count is an estimate of this probability
- 0 values are consistent with many probabilities, depending on library size

# Simulations

## Sampling from Poisson

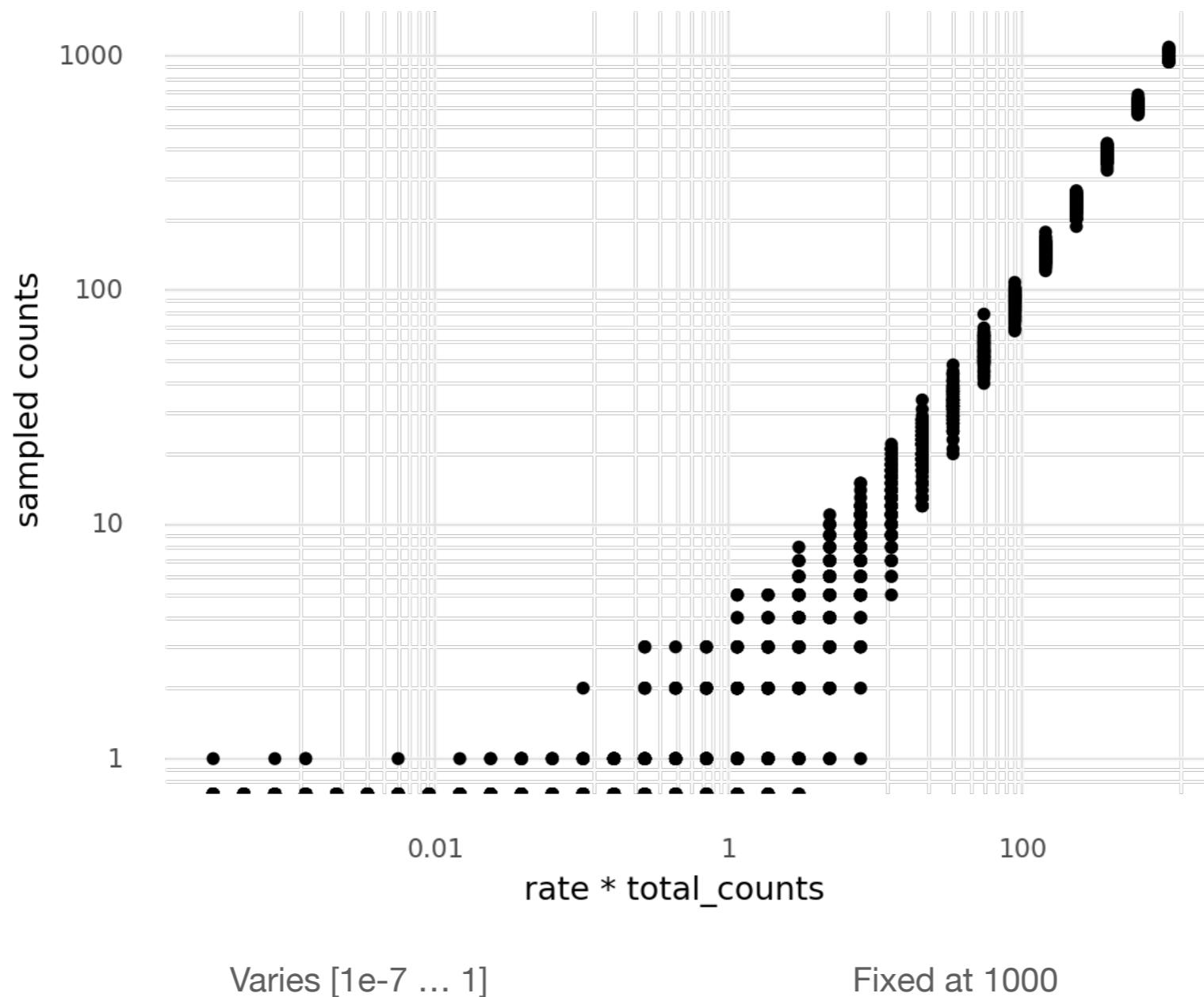
- $\text{count} \sim \text{Poisson}(\text{rate} \cdot \text{total\_count})$



# Simulations

## Sampling from Poisson

- $\text{count} \sim \text{Poisson}(\text{rate} \cdot \text{total\_count})$

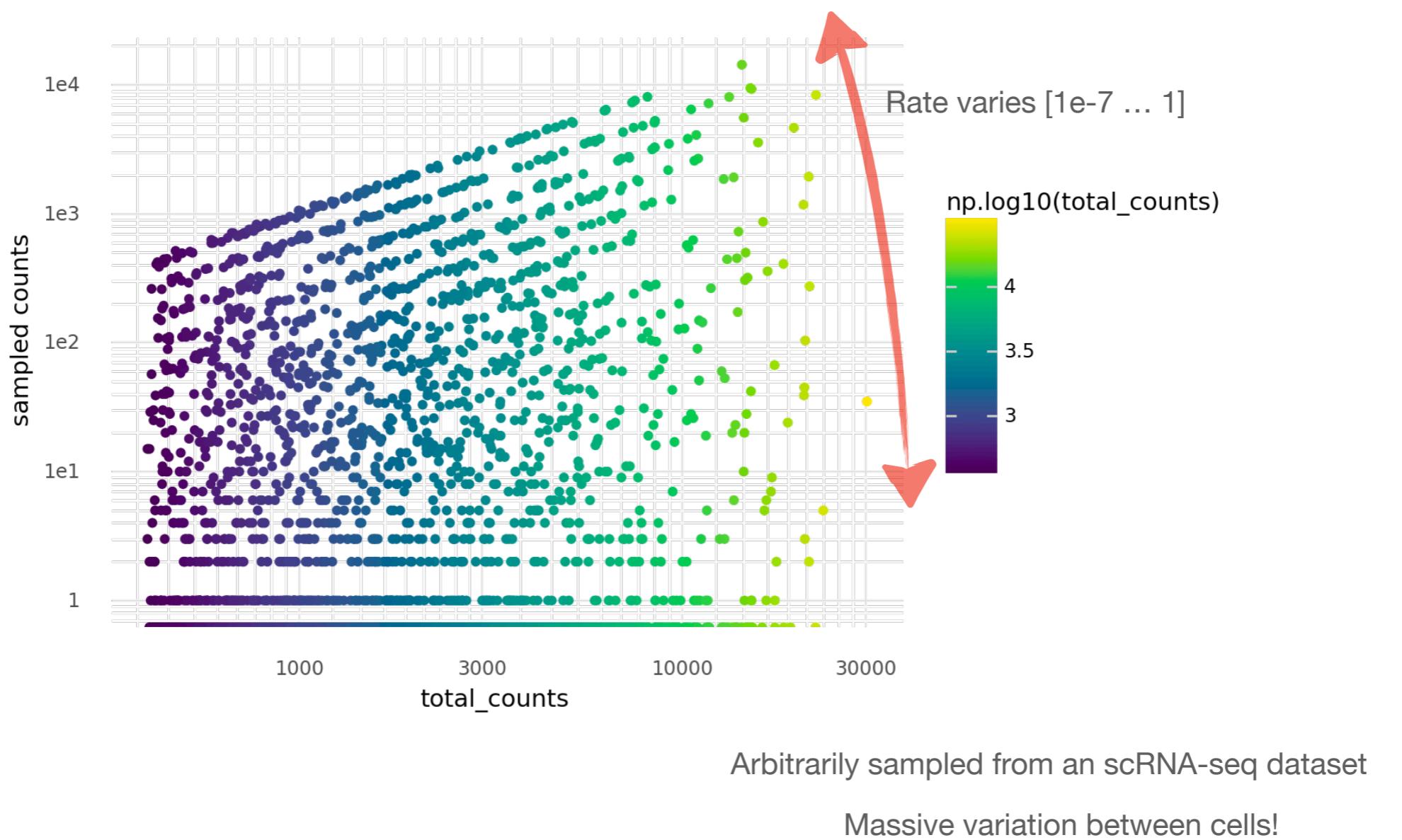


100 samples per rate value

# Simulations

## Sampling from Poisson

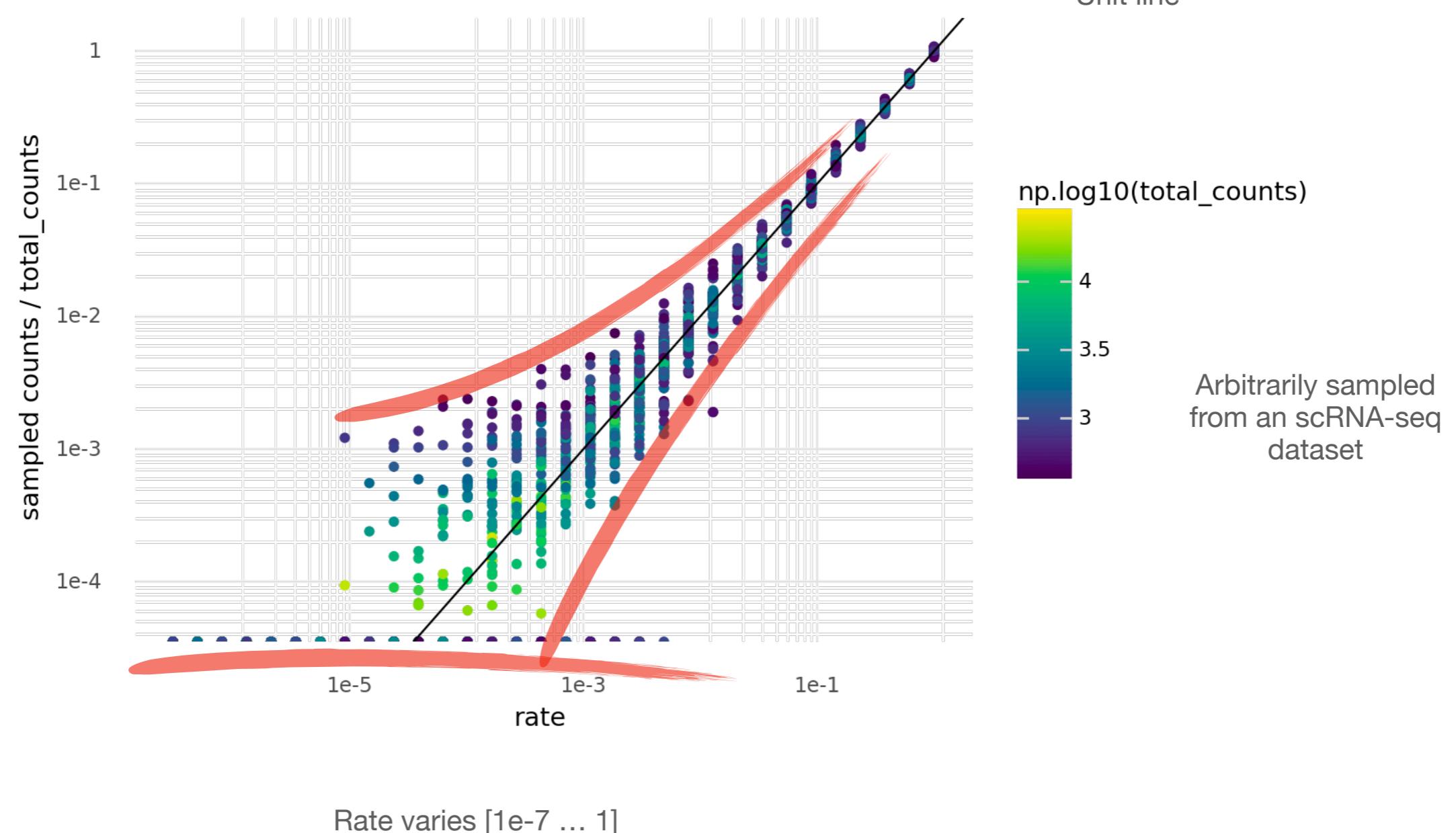
- $\text{count} \sim \text{Poisson}(\text{rate} \cdot \text{total\_count})$



# Simulations

## Sampling from Poisson

- $\text{count} \sim \text{Poisson}(\text{rate} \cdot \text{total\_count})$

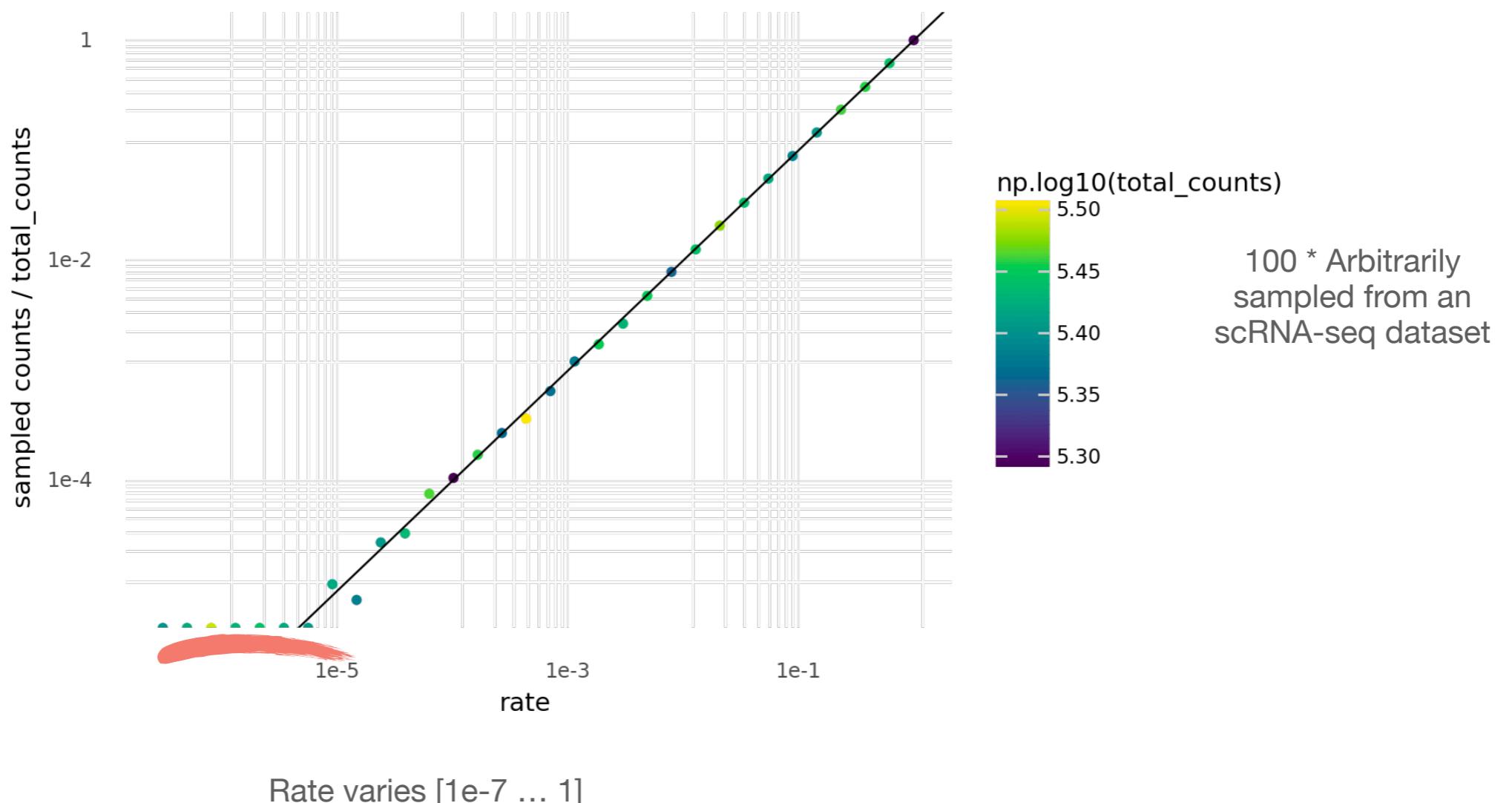


# Simulations

## Sampling from Poisson

- $\text{count} \sim \text{Poisson}(\text{rate} \cdot \text{total\_count})$

Unit line



# Units and scales

- Rates are fractions -> effects are multiplicative
- Look at fraction on a log scale
- **Rate:** for each molecule we see, what is the probability it comes from gene X?
- Count / total\_count is an estimate of this probability
- 0 values are consistent with many probabilities, depending on library size

# **Units and scales**

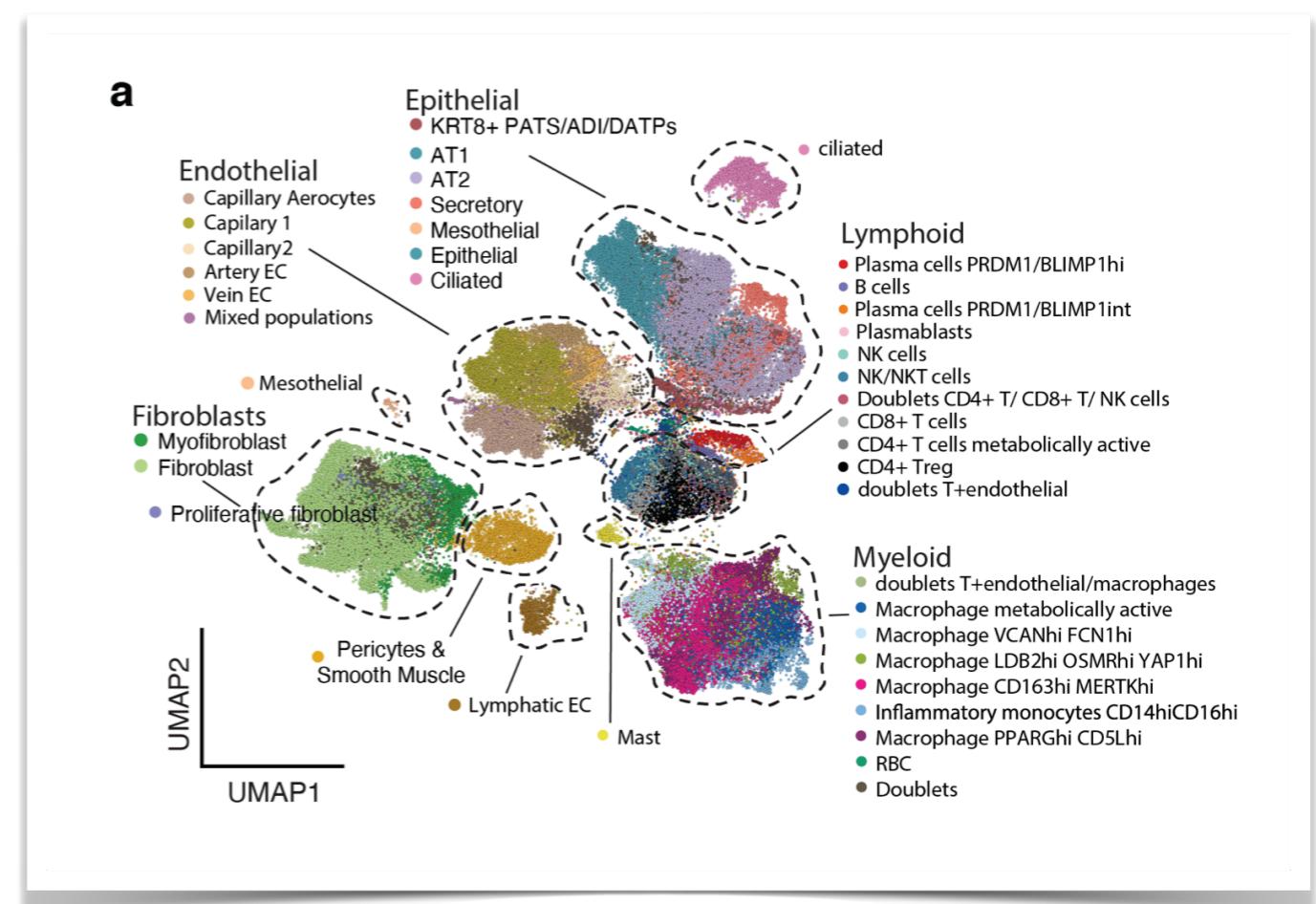
# Example data

## A single-cell and spatial atlas of autopsy tissues reveals pathology and cellular targets of SARS-CoV-2

Toni M. Delorey<sup>1\*</sup>, Carly G. K. Ziegler<sup>2,3,4,5,6,7\*</sup>, Graham Heimberg<sup>1\*</sup>, Rachelly Normand<sup>2,8,9,10,11\*</sup>, Yiming Yang<sup>1,8\*</sup>, Asa Segerstolpe<sup>1\*</sup>, Domenic Abbondanza<sup>1\*</sup>, Stephen J. Fleming<sup>12,13\*</sup>, Ayshwarya Subramanian<sup>1\*</sup>, Daniel T. Montoro<sup>2\*</sup>, Karthik A. Jagadeesh<sup>1\*</sup>, Kushal K. Dey<sup>14\*</sup>, Pritha Sen<sup>2,8,15,16\*</sup>, Michal Slyper<sup>1\*</sup>, Yered H. Pita-Juárez<sup>2,10,17,18,19\*</sup>, Devan Phillips<sup>1\*</sup>, Zohar Bloom-Ackerman<sup>20^†</sup>, Nick Barkas<sup>12^†</sup>, Andrea Ganna<sup>21,22^†</sup>, James Gomez<sup>20^†</sup>, Erica Normandin<sup>2,10^†</sup>, Pourya Naderi<sup>10,17,18^†</sup>, Yury V. Popov<sup>10,23,24^†</sup>, Siddharth S. Raju<sup>2,25,26^†</sup>, Sebastian Niezen<sup>10,23,24^†</sup>, Linus T.-Y. Tsai<sup>2,10,23,27,28^†</sup>, Katherine J. Siddle<sup>2,29^†</sup>, Malika Sud<sup>1^†</sup>, Victoria M. Tran<sup>20^†</sup>, Shamsudheen K. Vellarikkal<sup>2,30^†</sup>, Liat Amir-Zilberstein<sup>1</sup>, Deepak S. Atri<sup>2,30</sup>, Joseph Beechem<sup>31</sup>, Olga R. Brook<sup>32</sup>, Jonathan Chen<sup>2,33</sup>, Prajan Divakar<sup>31</sup>, Phylicia Dorceus<sup>1</sup>, Jesse M. Engreitz<sup>2,34</sup>, Adam Essene<sup>23,27,28</sup>, Donna M. Fitzgerald<sup>35</sup>, Robin Fropf<sup>31</sup>, Steven Gazal<sup>36</sup>, Joshua Gould<sup>12</sup>, John Grzyb<sup>37</sup>, Tyler Harvey<sup>1</sup>, Jonathan Hecht<sup>10,17</sup>, Tyler Hether<sup>31</sup>, Judit Jane-Valbuena<sup>1</sup>, Michael Leney-Greene<sup>2</sup>, Hui Ma<sup>1,8</sup>, Cristin McCabe<sup>1</sup>, Daniel E. McLoughlin<sup>35</sup>, Eric M. Miller<sup>31</sup>, Christoph Muus<sup>2,38</sup>, Mari Niemi<sup>21</sup>, Robert Padera<sup>37,39,40</sup>, Liiliu Pan<sup>31</sup>, Deepti Pant<sup>23,27,28</sup>, Carmel Pe'er<sup>1</sup>, Jenna Pfiffner-Borges<sup>2</sup>, Christopher J. Pinto<sup>16,35</sup>, Jacob Plaisted<sup>37</sup>, Jason Reeves<sup>31</sup>, Marty Ross<sup>31</sup>, Melissa Rudy<sup>2</sup>, Erroll H. Rueckert<sup>31</sup>, Michelle Siciliano<sup>37</sup>, Alexander Sturm<sup>20</sup>, Ellen Todres<sup>1</sup>, Avinash Waghray<sup>41,42</sup>, Sarah Warren<sup>31</sup>, Shuting Zhang<sup>20</sup>, Daniel R. Zollinger<sup>31</sup>, Lisa Cosimi<sup>43</sup>, Rajat M. Gupta<sup>2,30</sup>, Nir Hacohen<sup>2,9,44</sup>, Winston Hide<sup>10,17,18,19</sup>, Alkes L. Price<sup>14</sup>, Jayaraj Rajagopal<sup>35</sup>, Purushothama Rao Tata<sup>45</sup>, Stefan Riedel<sup>10,17</sup>, Gyongyi Szabo<sup>2,10,23</sup>, Timothy L. Tickle<sup>1,12</sup>, Deborah Hung<sup>20,46,47†</sup>, Pardis C. Sabeti<sup>2,29,48,49,50†</sup>, Richard Novak<sup>51†</sup>, Robert Rogers<sup>23,52†</sup>, Donald E. Ingber<sup>38,51,53†</sup>, Z. Gordon Jiang<sup>10,23,24†</sup>, Dejan Juric<sup>16,35†</sup>, Mehrtash Babadi<sup>12,13†</sup>, Samouil L. Farhi<sup>1†</sup>, James R. Stone<sup>33†</sup>, Ioannis S. Vlachos<sup>2,10,17,18,19†</sup>, Isaac H. Solomon<sup>37†</sup>, Orr Ashenberg<sup>1†</sup>, Caroline B.M. Porter<sup>1†</sup>, Bo Li<sup>1,8,16†</sup>, Alex K. Shalek<sup>2,3,4,5,6,7,10,41,54,55,56†</sup>, Alexandra-Chloé Villani<sup>2,8,9,16†</sup>, Orit Rozenblatt-Rosen<sup>1,57†</sup>, Aviv Regev<sup>1,5,49,57†</sup>

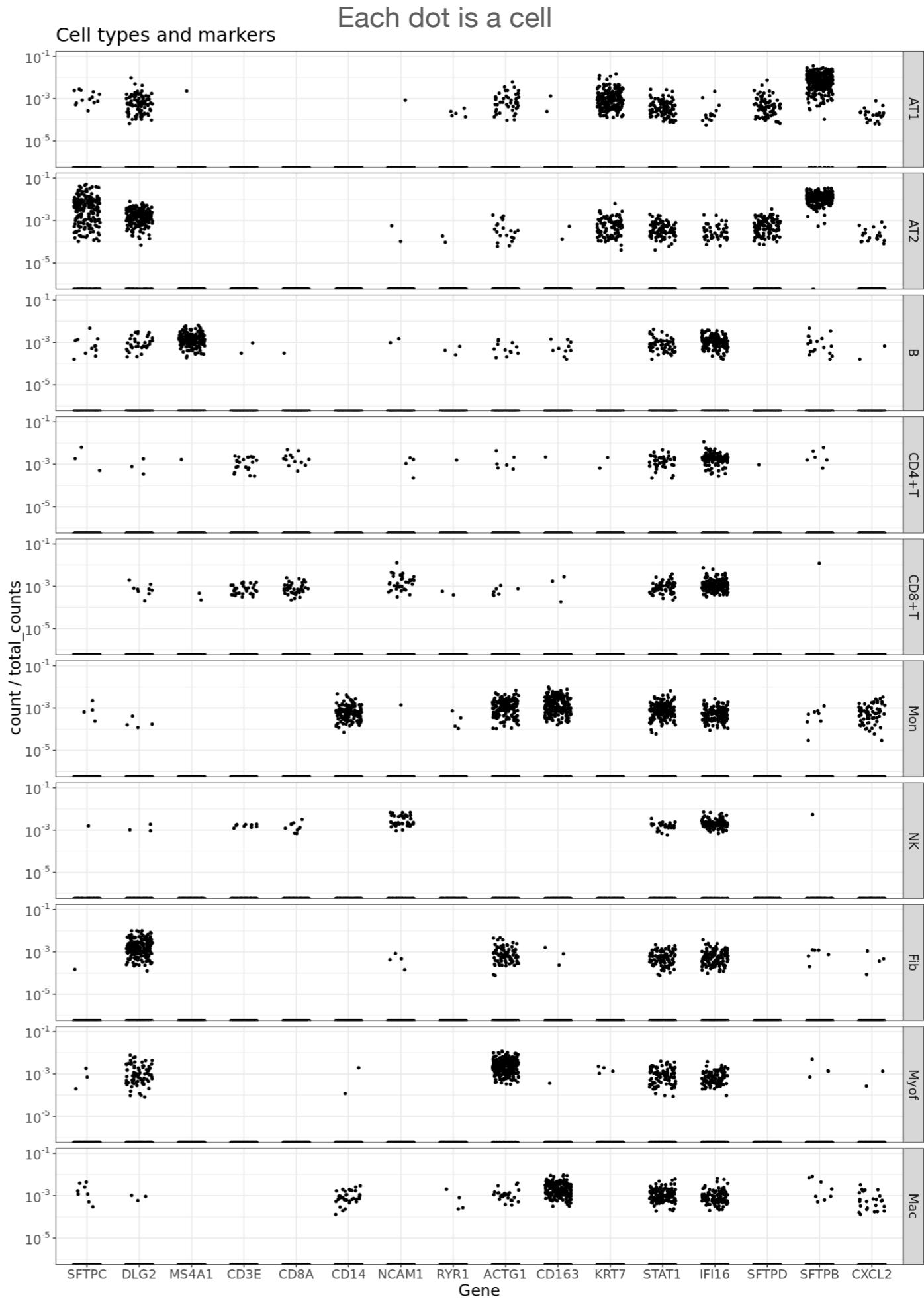
# Example data

- Lung autopsies
- ~100,000 cells
- 16 donors
- ~24 batches
- Well annotated and organized data!



# Example data

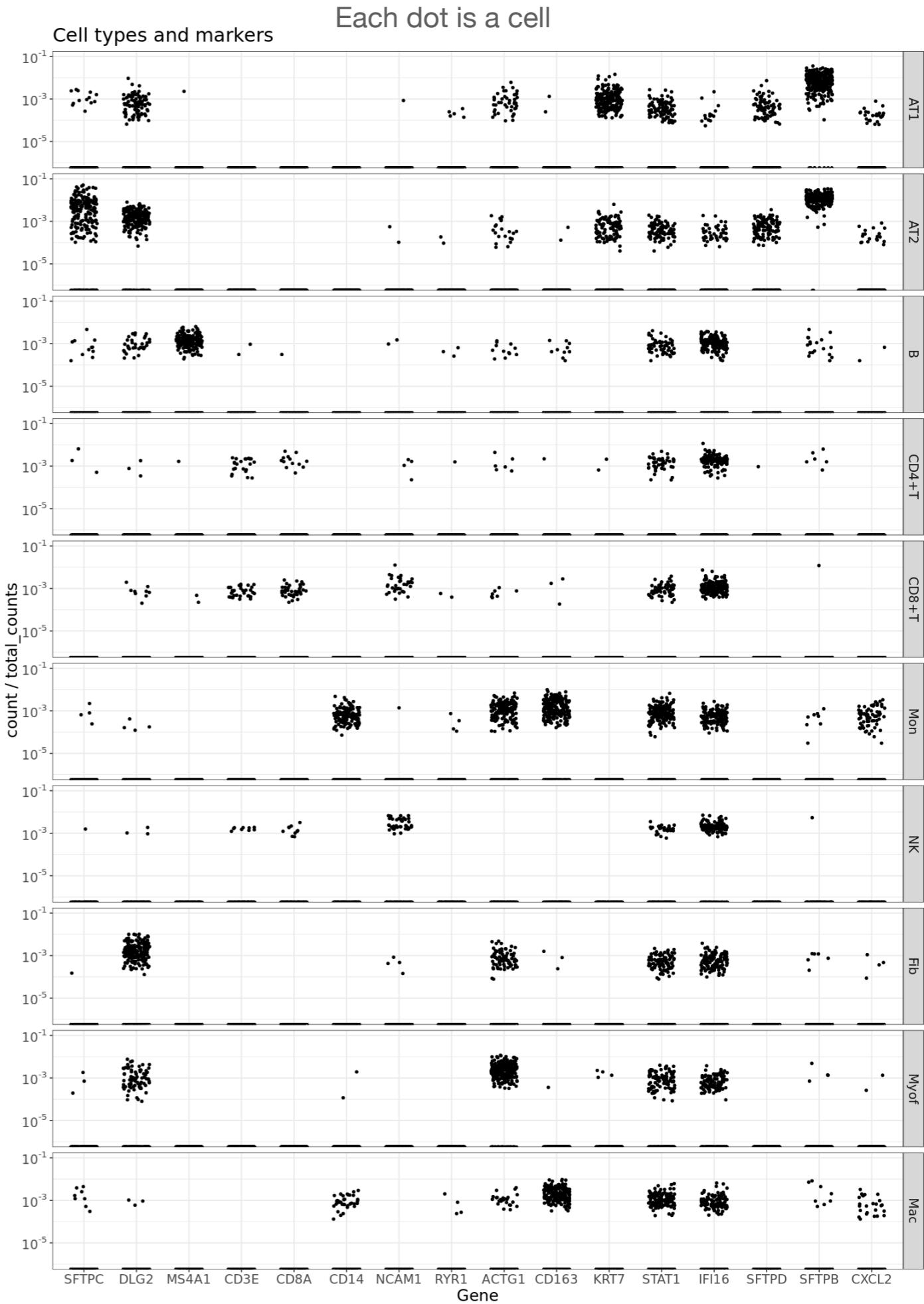
- Contains more information than can be seen
- Compare 0 / 100 and 0 / 100,000
- 0 / 100,000 gives stronger evidence on the *true* (small) rate of expression



# Generalized linear regression

$$\lambda = \exp(\beta_{\text{cell type}} + \log(\text{total\_count}))$$

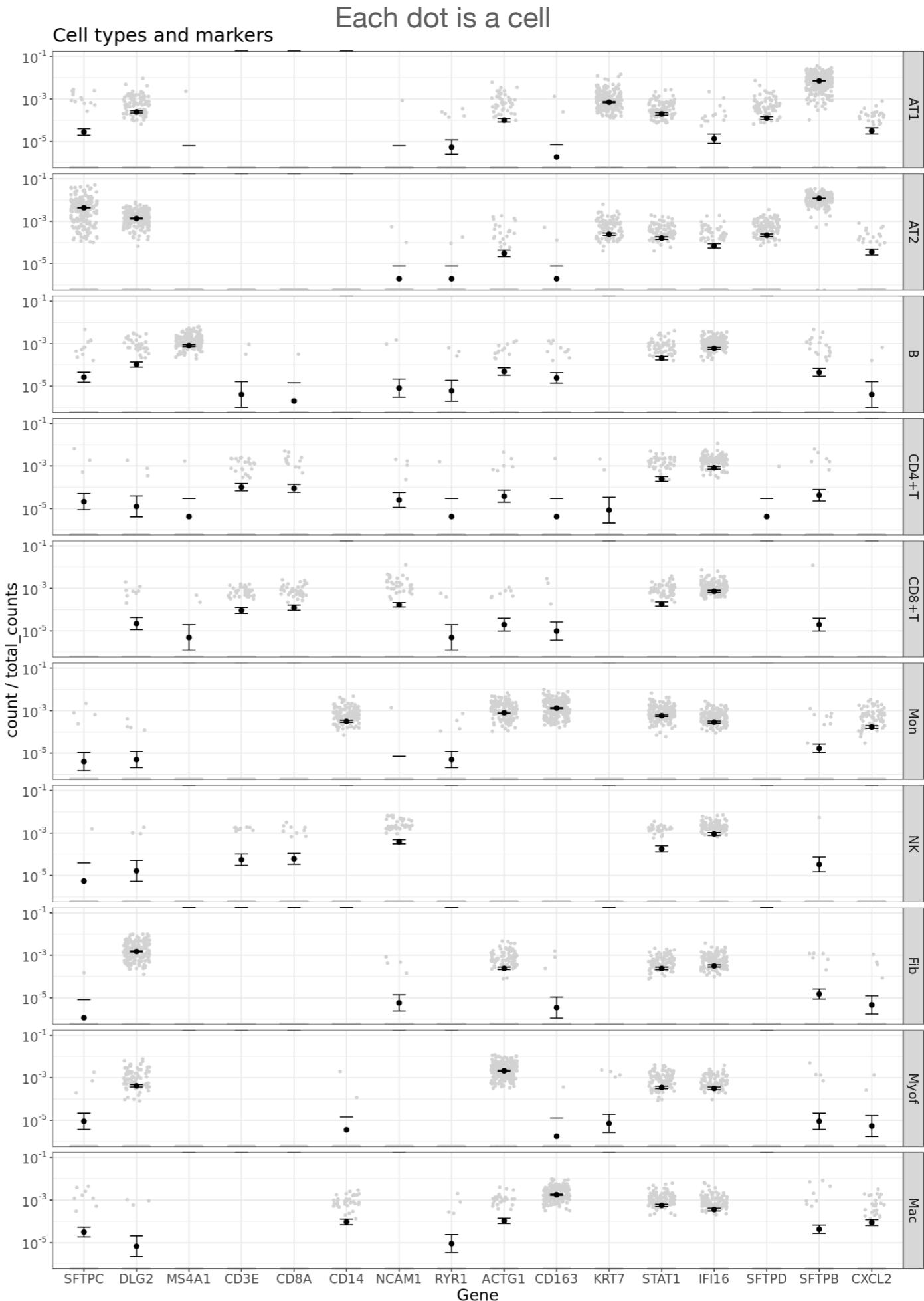
- count  $\sim$  Poisson( $\lambda$ )
- Makes use of both the count and the total\_count to estimate  $\beta_{\text{cell type}}$  values
- Also provide uncertainty of the estimates
- Set total\_count = 1, then rate<sub>cell type</sub> =  $\exp(\beta_{\text{cell type}})$



# Generalized linear regression

$$\lambda = \exp(\beta_{\text{cell type}} + \log(\text{total\_count}))$$

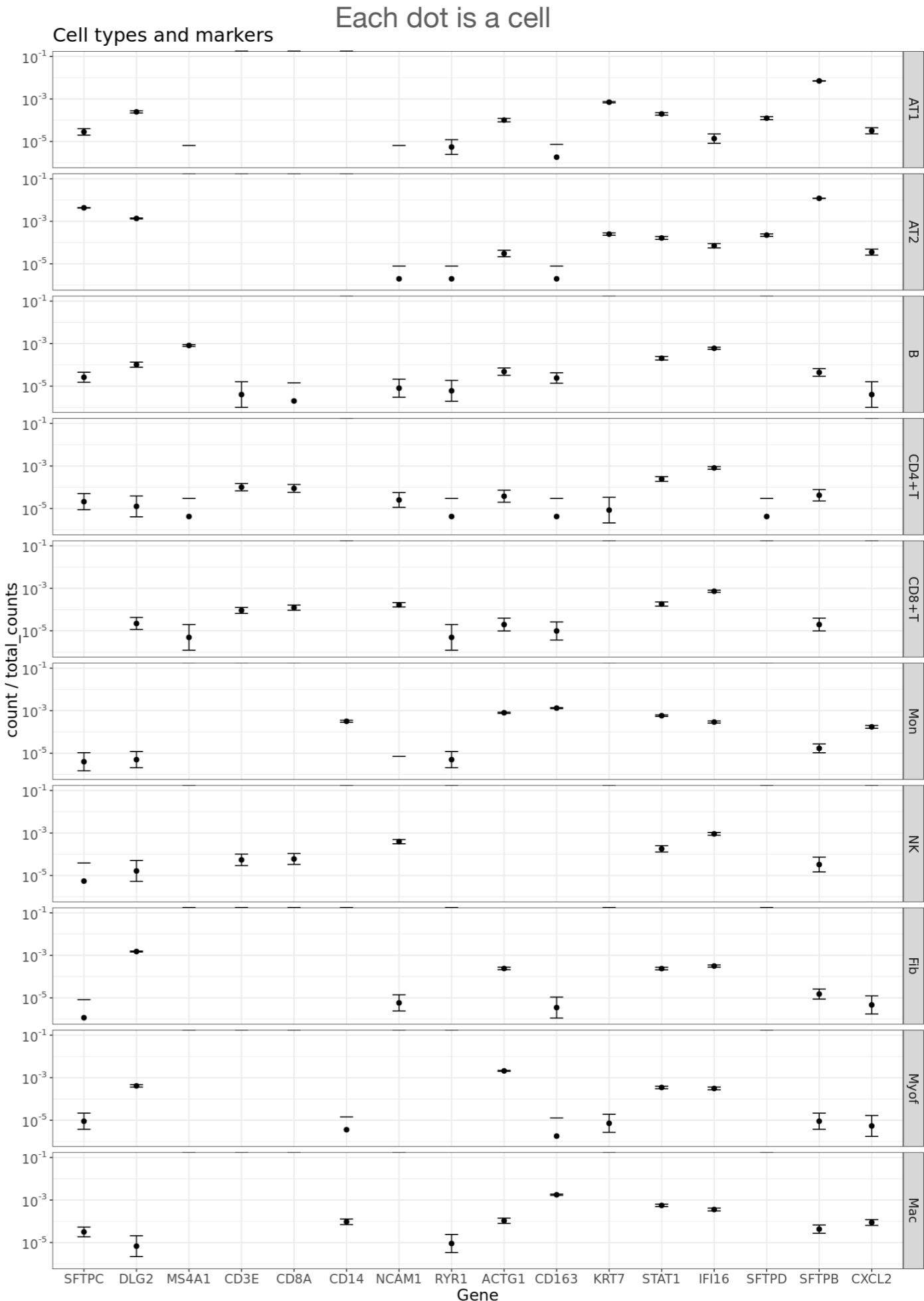
- count  $\sim$  Poisson( $\lambda$ )
- Makes use of both the count and the total\_count to estimate  $\beta_{\text{cell type}}$  values
- Also provide uncertainty of the estimates
- Set total\_count = 1, then rate<sub>cell type</sub> =  $\exp(\beta_{\text{cell type}})$



# Generalized linear regression

$$\lambda = \exp(\beta_{\text{cell type}} + \log(\text{total\_count}))$$

- count  $\sim$  Poisson( $\lambda$ )
- Makes use of both the count and the total\_count to estimate  $\beta_{\text{cell type}}$  values
- Also provide uncertainty of the estimates
- Set total\_count = 1, then rate<sub>cell type</sub> =  $\exp(\beta_{\text{cell type}})$



# Units and scales

- Rates put data on comparable scale
- When data and methods work on the same scale, can inspect performance and potential issues.

# Generalized linear regression

- Can be used for pairwise differential expression
- $\exp(\beta_1)/\exp(\beta_2)$  is the *fold change* between cell types 1 and 2
- Ratio of *rates*
- Uncertainties of the  $\beta$  estimates goes into uncertainty of the fold change
- This also gives p-value

contrast	ratio		SE	df	z.ratio	p.value	gene
	<fct>	<dbl>					
AT2 / (CD8+T)	5.965872e+07	6.679780e+09	Inf	0.15990627	1.000000e+00	SFTPC	
AT1 / (CD8+T)	3.906268e+05	4.373718e+07	Inf	0.11499410	1.000000e+00	SFTPC	
B / (CD8+T)	3.613813e+05	4.046273e+07	Inf	0.11429888	1.000000e+00	SFTPC	
(CD4+T) / (CD8+T)	2.901920e+05	3.249205e+07	Inf	0.11233890	1.000000e+00	SFTPC	
AT2 / Fib	3.707219e+03	3.707642e+03	Inf	8.21710129	1.235678e-13	SFTPC	
AT2 / Mon	1.086733e+03	5.436143e+02	Inf	13.97549640	0.000000e+00	SFTPC	
AT2 / NK	7.899900e+02	7.900800e+02	Inf	6.67126029	1.140538e-09	SFTPC	
AT2 / Myof	4.824935e+02	2.159005e+02	Inf	13.80872786	0.000000e+00	SFTPC	
AT2 / (CD4+T)	2.055836e+02	9.199215e+01	Inf	11.90219144	0.000000e+00	SFTPC	
AT2 / B	1.650853e+02	4.585417e+01	Inf	18.38440405	0.000000e+00	SFTPC	
AT2 / Mac	1.368293e+02	3.662745e+01	Inf	18.37493575	0.000000e+00	SFTPC	
AT1 / Fib	2.427372e+01	2.466213e+01	Inf	3.13916450	5.381351e-02	SFTPC	
B / Fib	2.245639e+01	2.330410e+01	Inf	2.99838897	8.074024e-02	SFTPC	
(CD4+T) / Fib	1.803266e+01	1.975379e+01	Inf	2.64019111	1.988891e-01	SFTPC	
Mac / Mon	7.942256e+00	4.502836e+00	Inf	3.65501263	9.686787e-03	SFTPC	
AT1 / Mon	7.115594e+00	3.780370e+00	Inf	3.69351372	8.407786e-03	SFTPC	
B / Mon	6.582862e+00	3.763895e+00	Inf	3.29584178	3.317184e-02	SFTPC	
Mac / NK	5.773543e+00	5.976185e+00	Inf	1.69383514	7.993630e-01	SFTPC	
(CD4+T) / Mon	5.286090e+00	3.546017e+00	Inf	2.48215294	2.779697e-01	SFTPC	
AT1 / NK	5.172609e+00	5.255376e+00	Inf	1.61749566	8.400001e-01	SFTPC	
B / NK	4.785345e+00	4.965988e+00	Inf	1.50860962	8.892221e-01	SFTPC	
(CD4+T) / NK	3.842670e+00	4.209434e+00	Inf	1.22887706	9.678995e-01	SFTPC	
Mac / Myof	3.526243e+00	1.837131e+00	Inf	2.41892890	3.142438e-01	SFTPC	
AT1 / Myof	3.159218e+00	1.522527e+00	Inf	2.38690386	3.335536e-01	SFTPC	
B / Myof	2.922693e+00	1.538023e+00	Inf	2.03807375	5.718054e-01	SFTPC	
(CD4+T) / Myof	2.346945e+00	1.484338e+00	Inf	1.34889251	9.421219e-01	SFTPC	
Myof / NK	1.637307e+00	1.793580e+00	Inf	0.45009360	9.999884e-01	SFTPC	
AT1 / (CD4+T)	1.346098e+00	6.487270e-01	Inf	0.61670549	9.998315e-01	SFTPC	
B / (CD4+T)	1.245318e+00	6.553296e-01	Inf	0.41690671	9.999940e-01	SFTPC	
AT1 / B	1.080927e+00	3.571661e-01	Inf	0.23551145	1.000000e+00	SFTPC	
AT1 / Mac	8.959159e-01	2.884886e-01	Inf	-0.34132707	9.999990e-01	SFTPC	
B / Mac	8.288403e-01	3.192398e-01	Inf	-0.48739637	9.999770e-01	SFTPC	
Mon / NK	7.269399e-01	8.127435e-01	Inf	-0.28524306	9.999998e-01	SFTPC	
(CD4+T) / Mac	6.655653e-01	3.467516e-01	Inf	-0.78143550	9.988389e-01	SFTPC	
Mon / Myof	4.439851e-01	2.978343e-01	Inf	-1.21040488	9.709420e-01	SFTPC	
Fib / Mon	2.931398e-01	3.277403e-01	Inf	-1.09755666	9.850748e-01	SFTPC	
Fib / NK	2.130950e-01	3.013619e-01	Inf	-1.09319917	9.854850e-01	SFTPC	
Fib / Myof	1.301497e-01	1.425719e-01	Inf	-1.86140766	6.953367e-01	SFTPC	
Fib / Mac	3.690888e-02	3.820432e-02	Inf	-3.18742955	4.652775e-02	SFTPC	
AT1 / AT2	6.547690e-03	1.180146e-03	Inf	-27.89994254	0.000000e+00	SFTPC	
(CD8+T) / Fib	6.214044e-05	6.957928e-03	Inf	-0.08650555	1.000000e+00	SFTPC	
(CD8+T) / Mon	1.821584e-05	2.039585e-03	Inf	-0.09746761	1.000000e+00	SFTPC	
(CD8+T) / NK	1.324182e-05	1.482700e-03	Inf	-0.10031285	1.000000e+00	SFTPC	
(CD8+T) / Myof	8.087560e-06	9.055433e-04	Inf	-0.10471959	1.000000e+00	SFTPC	
(CD8+T) / Mac	2.293534e-06	2.567998e-04	Inf	-0.11597554	1.000000e+00	SFTPC	

# Generalized linear regression

- Can evaluate markers

- $\exp(\beta_1)/(\exp(\beta_2) \cdot \dots \cdot \exp(\beta_n))$  is the *fold change* between cell type 1 and the rest of the cells.

- Ratio of *rates*

- Uncertainties of the  $\beta$  estimates goes into uncertainty of the fold change

- This also gives p-value

contrast	ratio	SE	df	z.ratio	p.value	gene
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
AT2 effect	1.590070e+03	1.978387e+04	Inf	0.592465082	0.9973265	SFTPC
Mac effect	6.727967e+00	8.372931e+01	Inf	0.153176238	0.9973265	SFTPC
AT1 effect	5.954529e+00	7.409466e+01	Inf	0.143381280	0.9973265	SFTPC
B effect	5.461298e+00	6.796686e+01	Inf	0.136413144	0.9973265	SFTPC
(CD4+T) effect	4.279854e+00	5.328447e+01	Inf	0.116780023	0.9973265	SFTPC
Myof effect	1.658666e+00	2.065051e+01	Inf	0.040643476	0.9973265	SFTPC
NK effect	9.590400e-01	1.197050e+01	Inf	-0.003350692	0.9973265	SFTPC
Mon effect	6.728933e-01	8.378900e+00	Inf	-0.031815532	0.9973265	SFTPC
Fib effect	1.721104e-01	2.148240e+00	Inf	-0.140975283	0.9973265	SFTPC
(CD8+T) effect	3.645680e-06	4.081947e-04	Inf	-0.111836543	0.9973265	SFTPC

# Units and scales

- Rates put data on comparable scale
- When data and methods work on the same scale, can inspect performance and potential issues.
- Changes and differences in rates are interpretable fold changes

# Towards summary statistics

- How will we use data when the HCA has 1,000,000,000 cells cataloging 1,000,000 cell types?
- Different levels of data: FASTQ, Count matrices, summary statistics?
- Typical question: “Is gene X highly expressed in cell type Y?”
- Can summarize per cell type expression rate based on all data.
- If also including confidence intervals, meta-analys methods can be used to gain new insights from the summarized data!
- The generalized linear model is one way to summarize the data

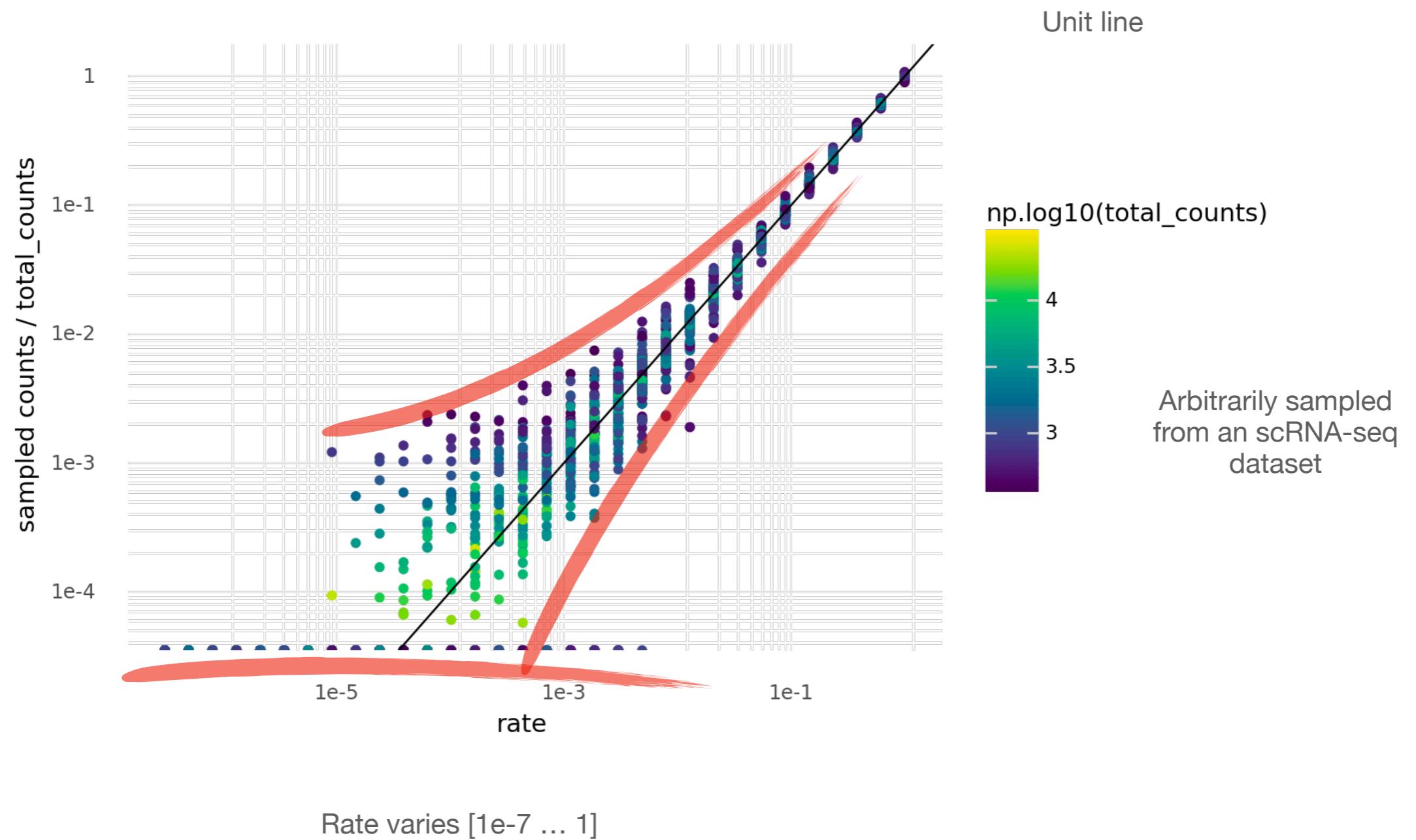
# Generalized linear regression

- **Many available packages:** glm in R, statsmodels in Python, brms, DEseq, EdgeR, ...
- Useful package for interpreting models: emmeans
  - A couple of great posts describing emmeans:
    - <https://aosmith.rbind.io/2019/03/25/getting-started-with-emmeans/>
    - <https://timmastny.rbind.io/blog/tests-pairwise-categorical-mean-emmeans-contrast/>
  - Out of scope for this presentation: generalized linear mixed models. Enables analysis of data with hierarchical structure.

# Units and scales

- Rates put data on comparable scale
- When data and methods work on the same scale, can inspect performance and potential issues.
- Changes and differences in rates are interpretable fold changes
- Summarized rates with confidence intervals will enable meta-analysis
- Standardized expression unit will help the community build intuition about expression levels

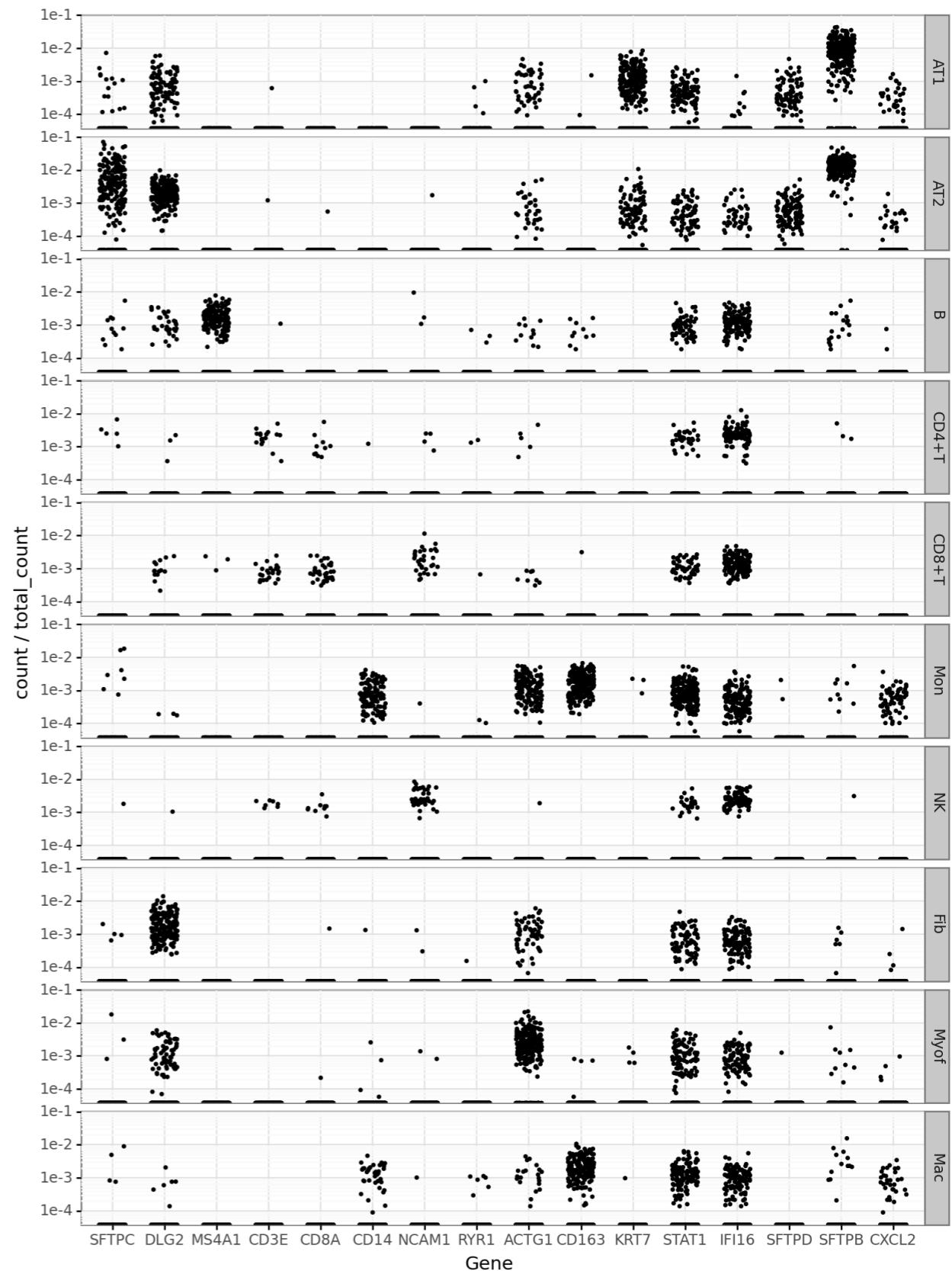
# Sampling from Poisson



# scVI

- Representation model (variational autoencoder)
  - Learn to represent each cell as a vector  $z_i$ .
- $$\lambda_i = f_\theta(z_i, \text{total\_count}_i)^*$$
- $\text{count}_i \sim \text{Poisson}(\lambda_i)$
  - Set  $\text{total\_count}_i = 1$ , then  

$$\text{rate}_i = f_\theta(z_i, 1).$$
  - (Warning! This  $\text{rate}_i$  is a mean estimate! Do not plug it in to another analysis method directly, will likely cause overfitting!)

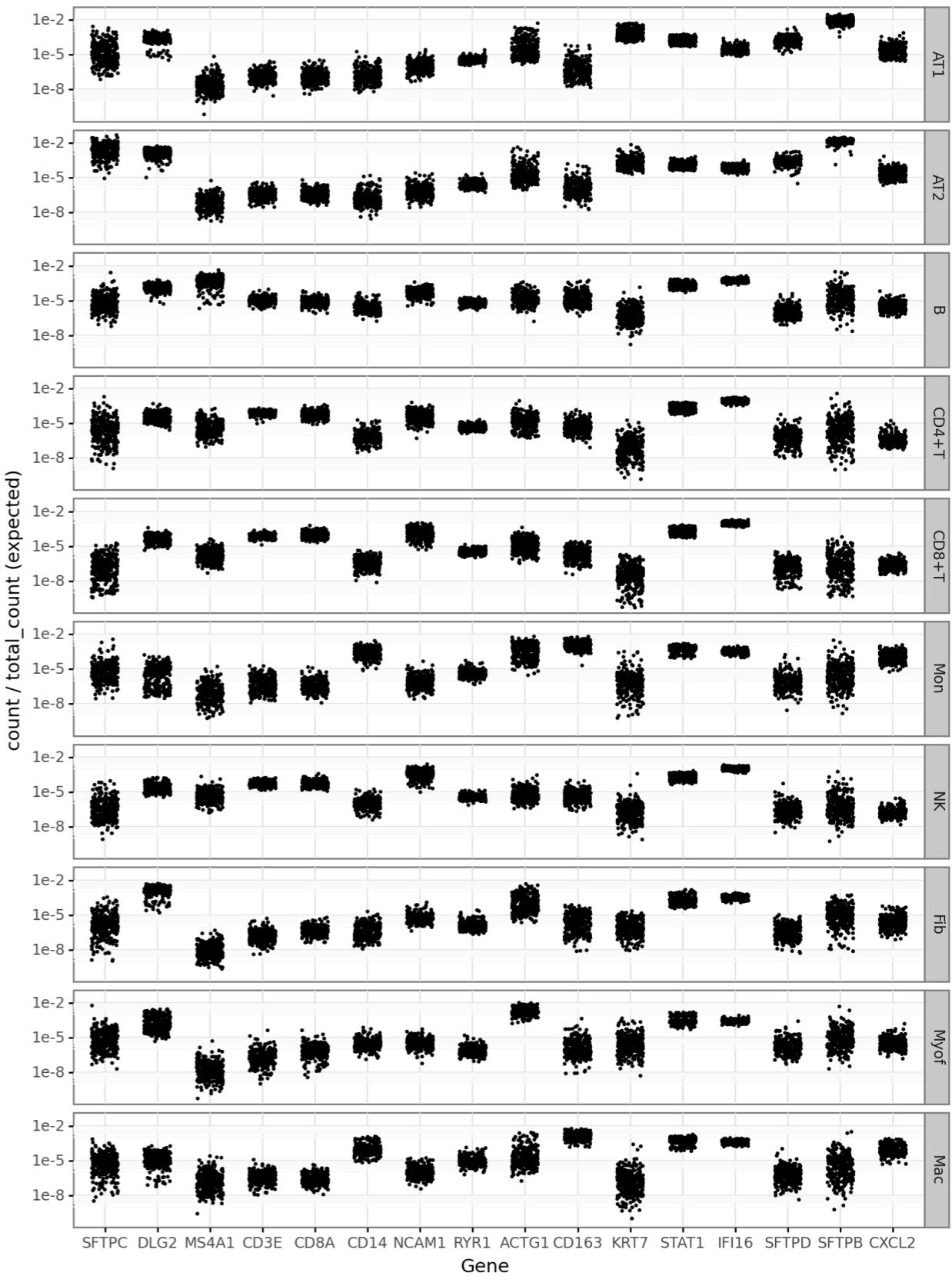


\* This is a simplification for the sake of clarity. For example, scVI uses NB distribution rather than Poisson.

# scVI

- Representation model (variational autoencoder)
  - Learn to represent each cell as a vector  $z_i$ .
- $$\lambda_i = f_\theta(z_i, \text{total\_count}_i)^*$$
- $\text{count}_i \sim \text{Poisson}(\lambda_i)$
  - Set  $\text{total\_count}_i = 1$ , then  

$$\text{rate}_i = f_\theta(z_i, 1).$$
  - (Warning! This  $\text{rate}_i$  is a mean estimate! Do not plug it in to another analysis method directly, will likely cause overfitting!)

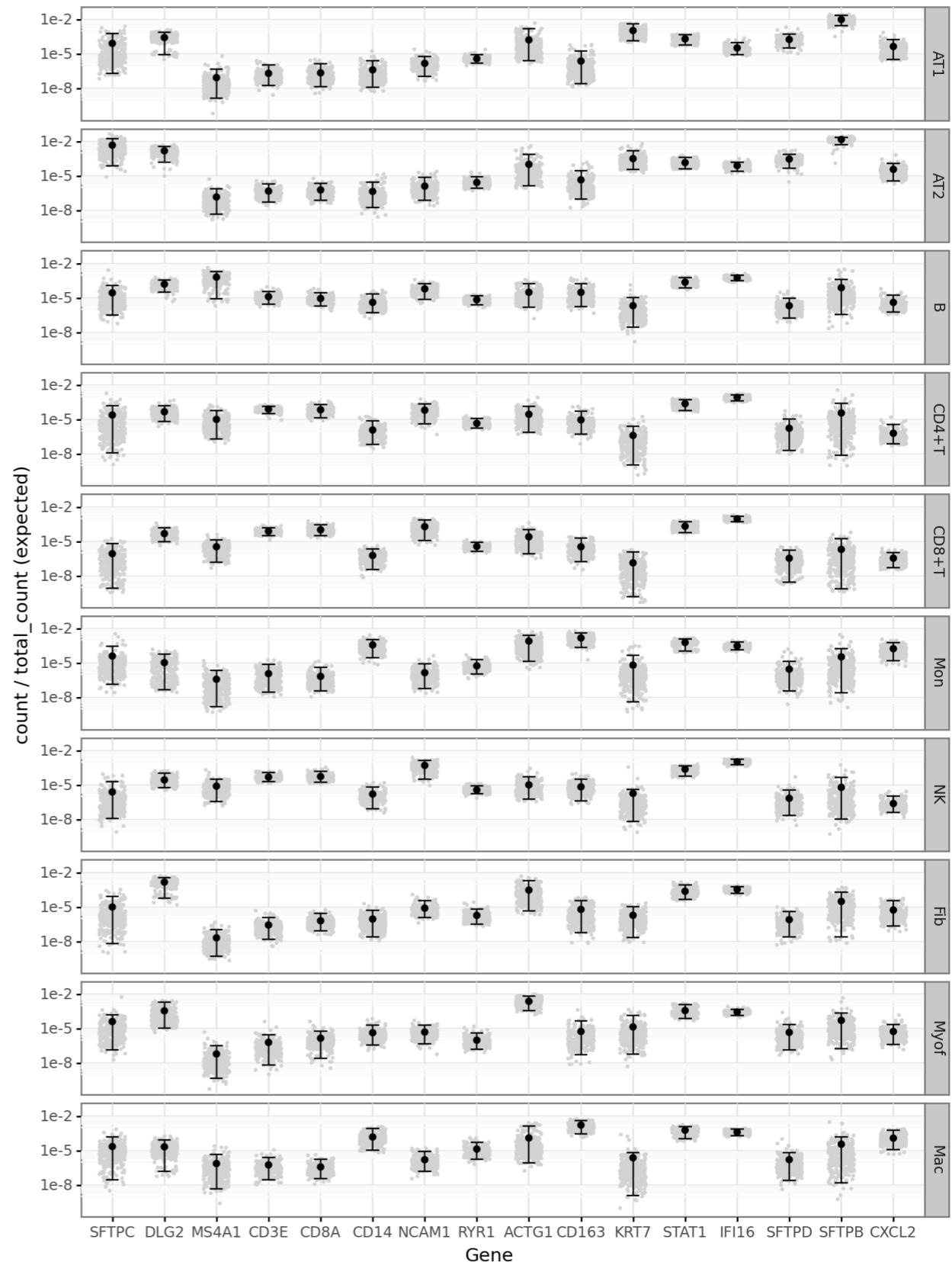


\* This is a simplification for the sake of clarity. For example, scVI uses NB distribution rather than Poisson.

# scVI

- Representation model (variational autoencoder)
  - Learn to represent each cell as a vector  $z_i$ .
- $$\lambda_i = f_\theta(z_i, \text{total\_count}_i)^*$$
- $\text{count}_i \sim \text{Poisson}(\lambda_i)$
  - Set  $\text{total\_count}_i = 1$ , then  

$$\text{rate}_i = f_\theta(z_i, 1).$$
  - (Warning! This  $\text{rate}_i$  is a mean estimate! Do not plug it in to another analysis method directly, will likely cause overfitting!)

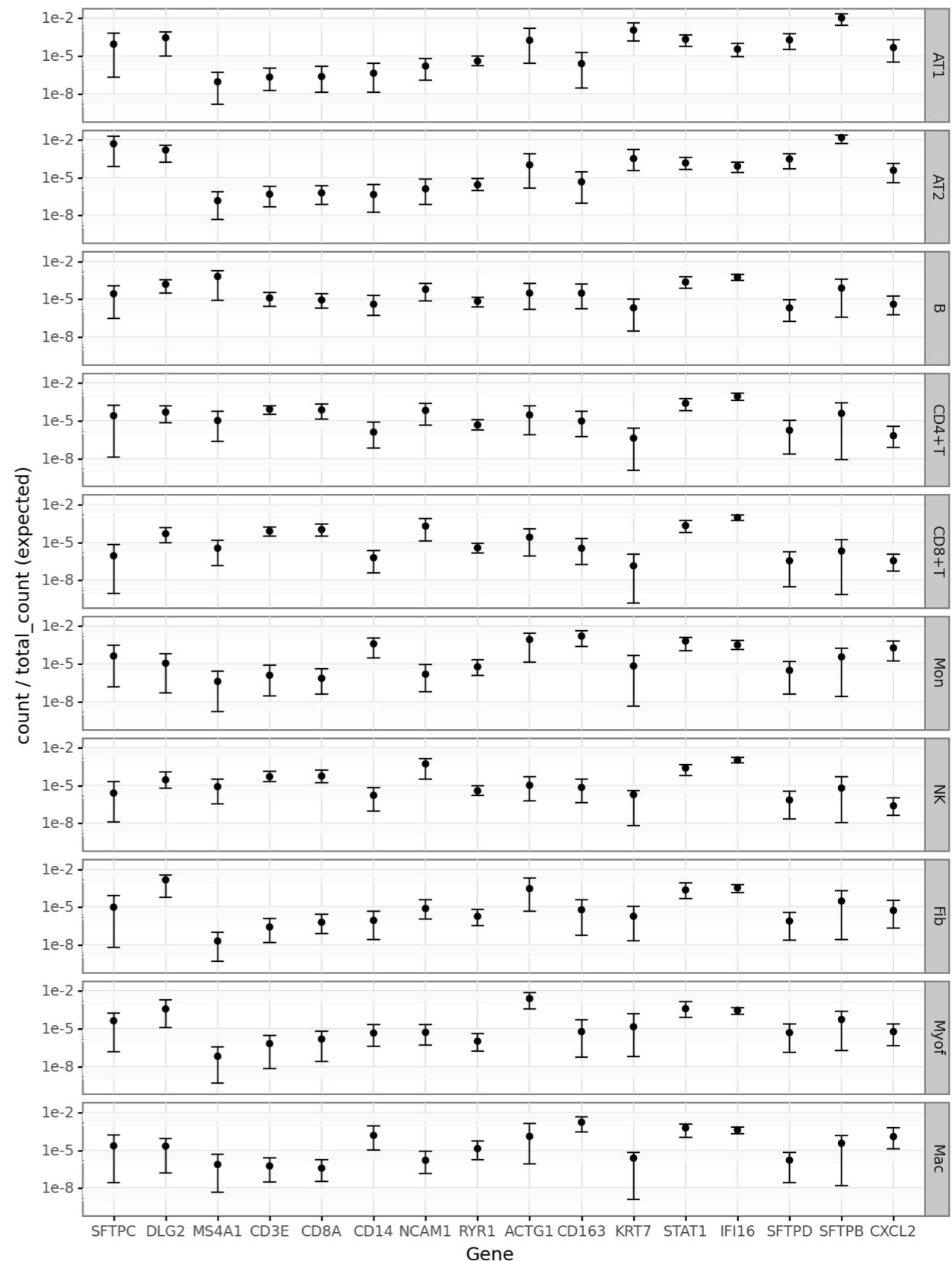


\* This is a simplification for the sake of clarity. For example, scVI uses NB distribution rather than Poisson.

# scVI

- Representation model (variational autoencoder)
  - Learn to represent each cell as a vector  $z_i$ .
- $$\lambda_i = f_\theta(z_i, \text{total\_count}_i)^*$$
- $\text{count}_i \sim \text{Poisson}(\lambda_i)$
  - Set  $\text{total\_count}_i = 1$ , then  

$$\text{rate}_i = f_\theta(z_i, 1).$$
  - (Warning! This  $\text{rate}_i$  is a mean estimate! Do not plug it in to another analysis method directly, will likely cause overfitting!)

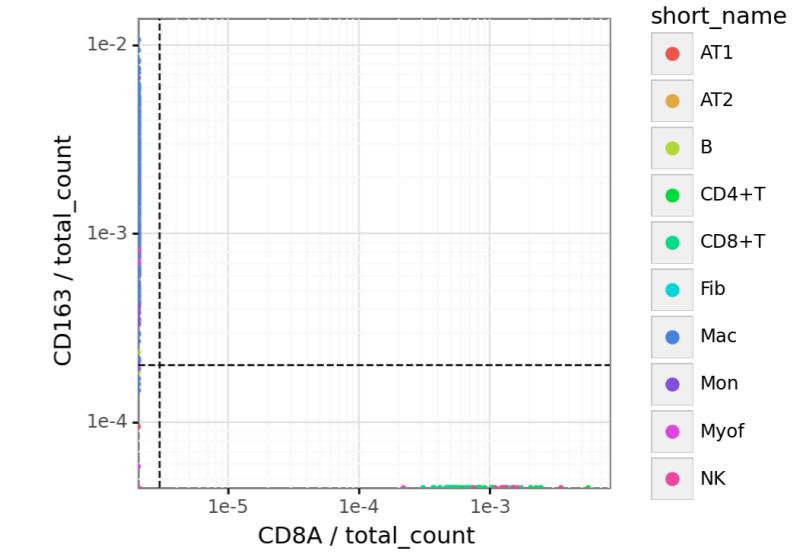
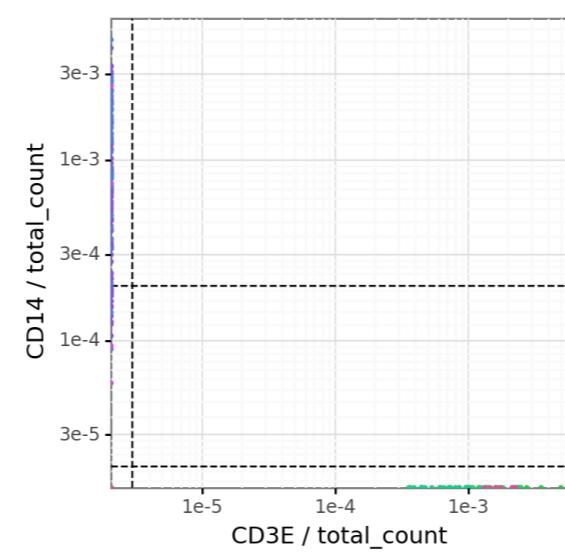
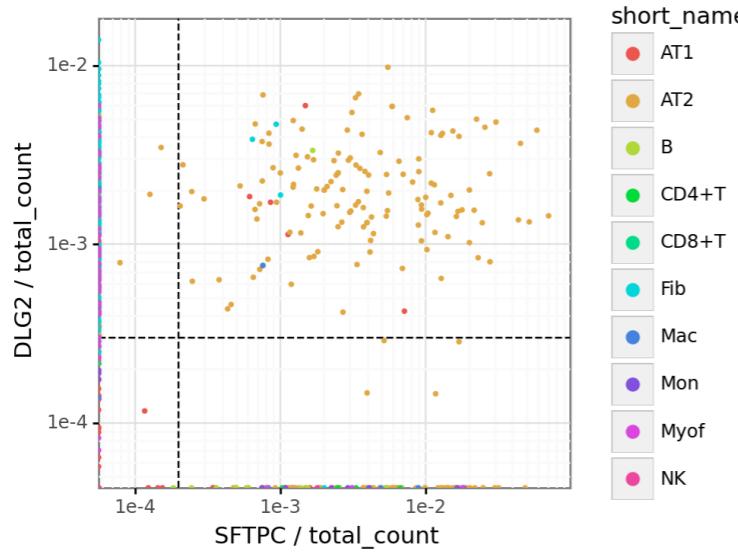


\* This is a simplification for the sake of clarity. For example, scVI uses NB distribution rather than Poisson.

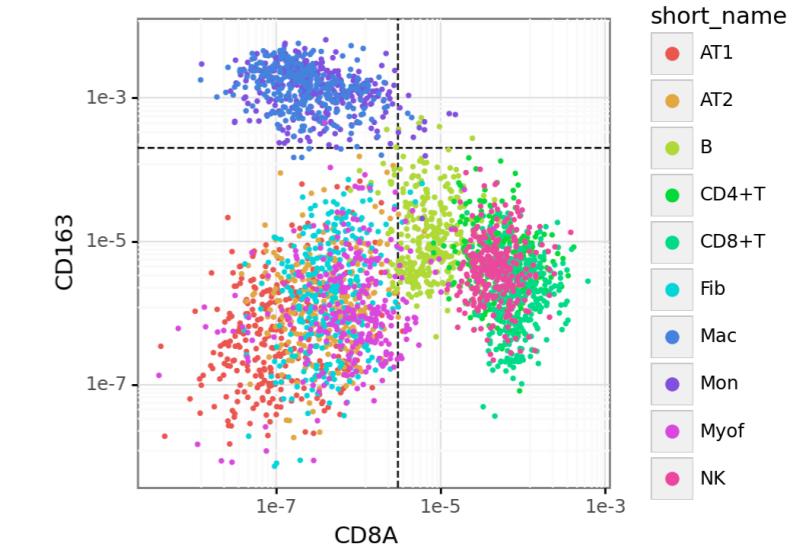
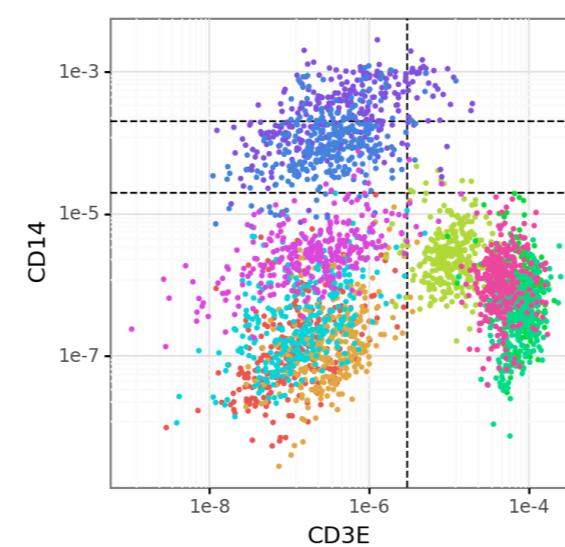
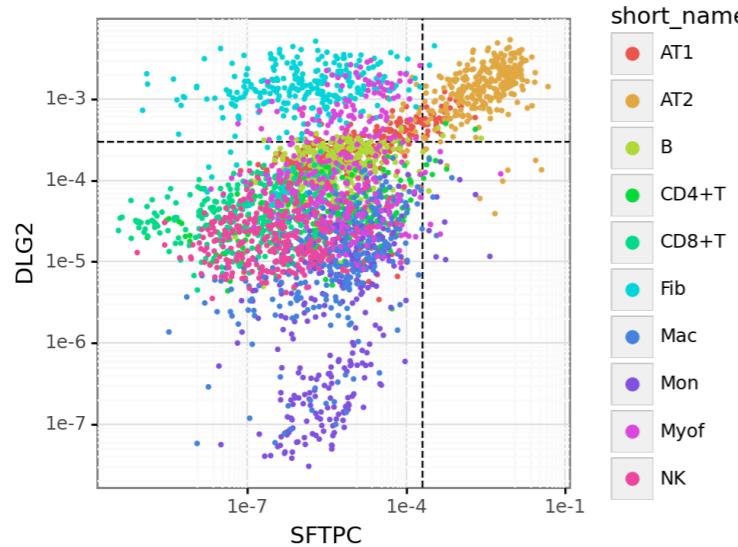
# scVI

## Pairwise estimated *rate* plots useful for exploration

Observed count / total\_count



Estimated count / total\_count



# Units and scales

- Rates put data on comparable scale
- When data and methods work on the same scale, can inspect performance and potential issues
- Changes and differences in rates are interpretable fold changes
- Summarized rates with confidence intervals will enable meta-analysis
- Standardized expression unit will help the community build intuition about expression levels
- Methods that tie rates to observed counts can also work on the single cell level
- Single cell expression rates are great visualization aids

# scVI

- Can do differential expression analysis by sampling  $z_i$  values from representation space
- From sampled pairs of  $z_i$ , generate pairs of  $\text{rate}_i = f_\theta(z_i, 1)$
- Compare ratios of rates, give distribution of *fold changes*.
- This gives mean fold change
- This also gives posterior probability of significant fold change

	comparison	featurekey	proba_de	lfc_mean	fc_mean
10005	AT2 vs Rest	SFTPC	0.9994	10.343406	1299.198120
1202	AT1 vs Rest	SFTPC	0.9646	1.214242	2.320189
60758	Mac vs Rest	SFTPC	0.9642	0.695837	1.619824
82486	Myof vs Rest	SFTPC	0.9622	0.476230	1.391104
21258	B vs Rest	SFTPC	0.9562	0.406211	1.325200
71262	Mon vs Rest	SFTPC	0.9624	0.270618	1.206324
30367	CD4+T vs Rest	SFTPC	0.9612	-0.913049	0.531062
51039	Fib vs Rest	SFTPC	0.9672	-2.217187	0.215060
90159	NK vs Rest	SFTPC	0.9744	-4.586297	0.041628
40097	CD8+T vs Rest	SFTPC	0.9728	-5.638350	0.020076

# Units and scales

- Rates put data on comparable scale
- When data and methods work on the same scale, can inspect performance and potential issues
- Changes and differences in rates are interpretable fold changes
- Summarized rates with confidence intervals will enable meta-analysis
- Standardized expression unit will help the community build intuition about expression levels
- Methods that tie rates to observed counts can also work on the single cell level
- Single cell expression rates are great visualization aids
- Distributions of single cell expression rates can be used for statistical analysis

# Communication experience

- Put everything that is a fraction on a log axis.
- Can develop intuition about expression levels by consistently using rate scale for expression levels.
- Pairwise estimated rate plots can be tied to intuition from flow cytometry.
- Can multiply rate by e.g. 10,000 to get most genes of interest to the 0.1-10.0 scale. But best do this at the plotting stage.

# References

- Example data source: Delorey et al. bioRxiv, 2021, <https://doi.org/10.1101/2021.02.25.430130>
- Scvi-tools pre-print: Gayoso et al. bioRxiv, 2021, <https://doi.org/10.1101/2021.04.28.441833>
- Differential expression with scVI: Boyneau et al. MLCB, 2019, [https://mlcb.github.io/mlcb2019\\_proceedings/](https://mlcb.github.io/mlcb2019_proceedings/)
- scvi-tools: <https://scvi-tools.org/>
- All analysis and plots available at <https://github.com/vals/210527-czi-talk>
-