# Power Analyses for Single-cell Data Workshop

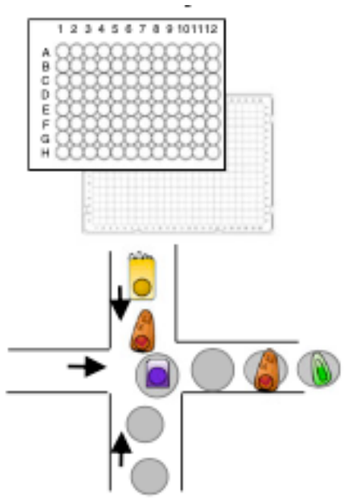
Kenong Su
Department of Computer Science
Emory University
2/10/2021

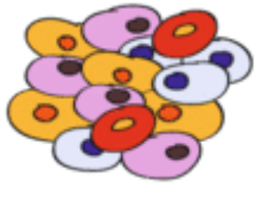
EMORY
UNIVERSITY

gmds

Deutsche Gesellschaft für
Medizinische Informatik,
Biometrie und
Epidemiologie e.V.

# Background





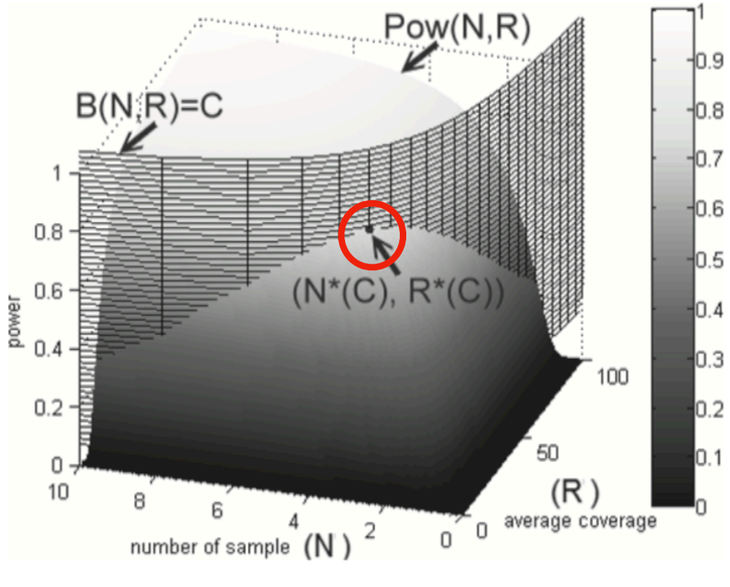
**# of cells?**

**Power evaluation**

- Under the context of differential expression (DE) tests.

**Difficulties**

- Complexities in DE analysis: type I error ($\alpha$), effect size, within group variance ($\sigma_{in}$), baseline expression level, etc.



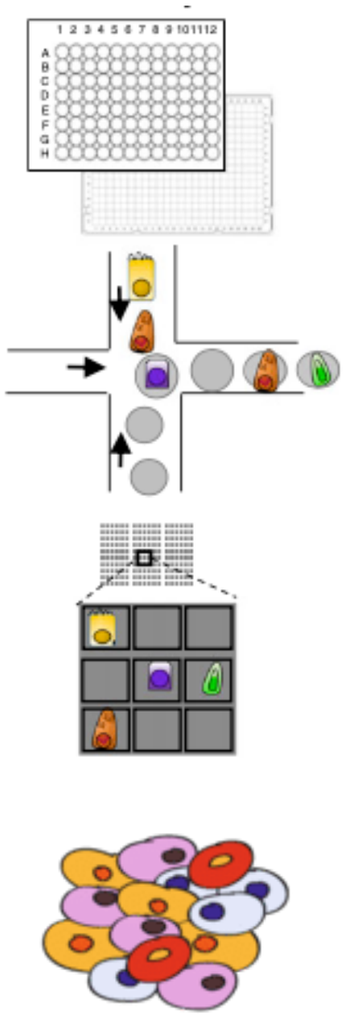

cost function

↓

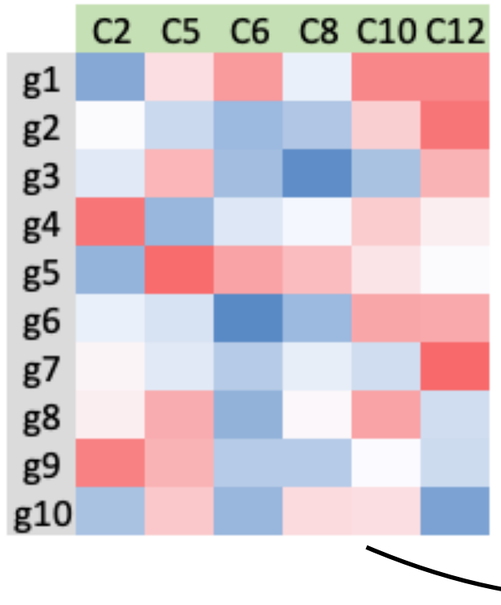
simulation

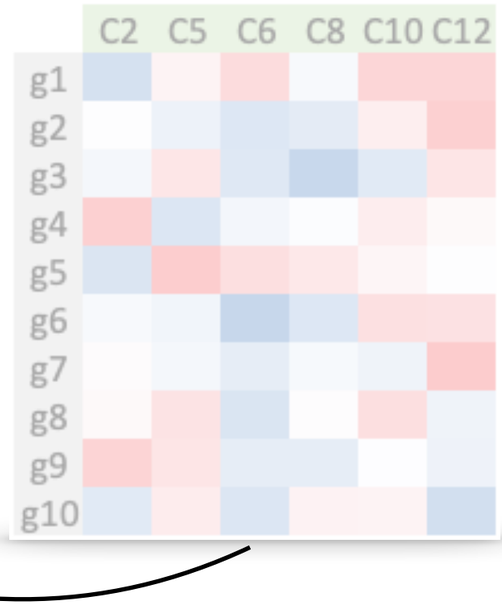
Atefeh et al. (2018) Nature Protocols
Lin et al. (2019) Appl. Statist.

# Scenarios



**# of cells?**

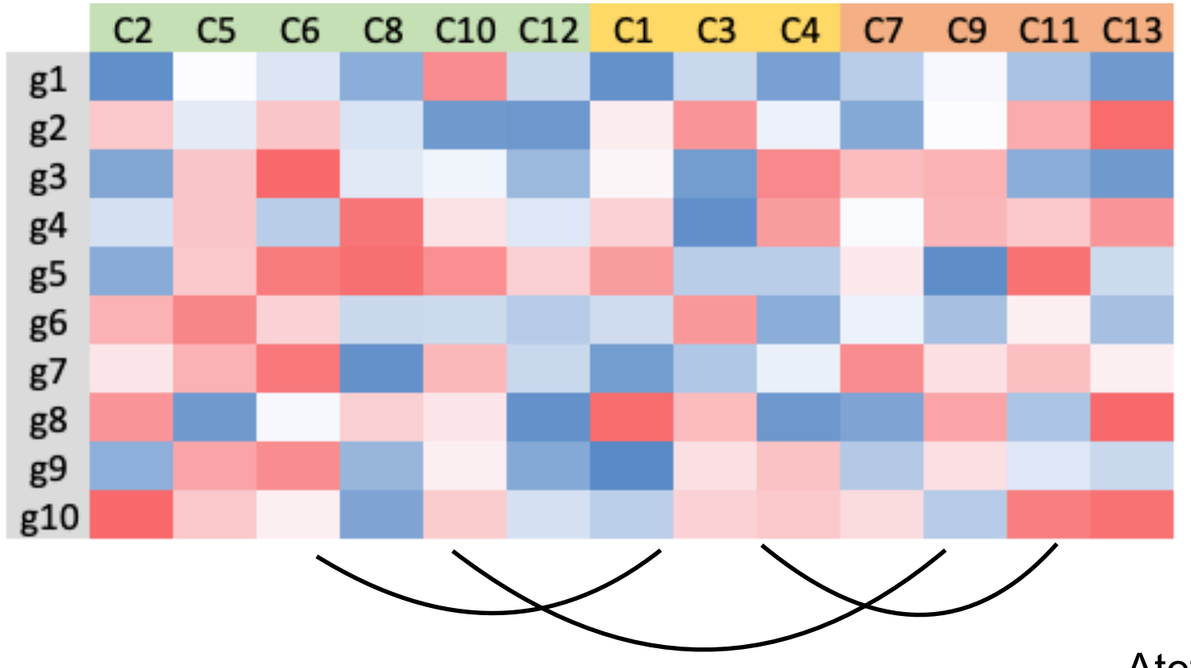



→ **scenario I**

I: The same cell type from two conditions.

II: Different cell types under one condition.




→ **scenario II**



Atefeh et al. (2018) Nature Protocols

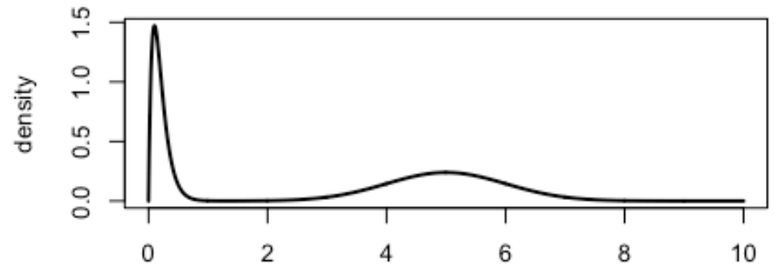
# POWSC Workflow

Estimate cell proportions given the scRNA-seq expression matrix.

Estimate model parameters for each cell type. — ***Parameter Estimator***

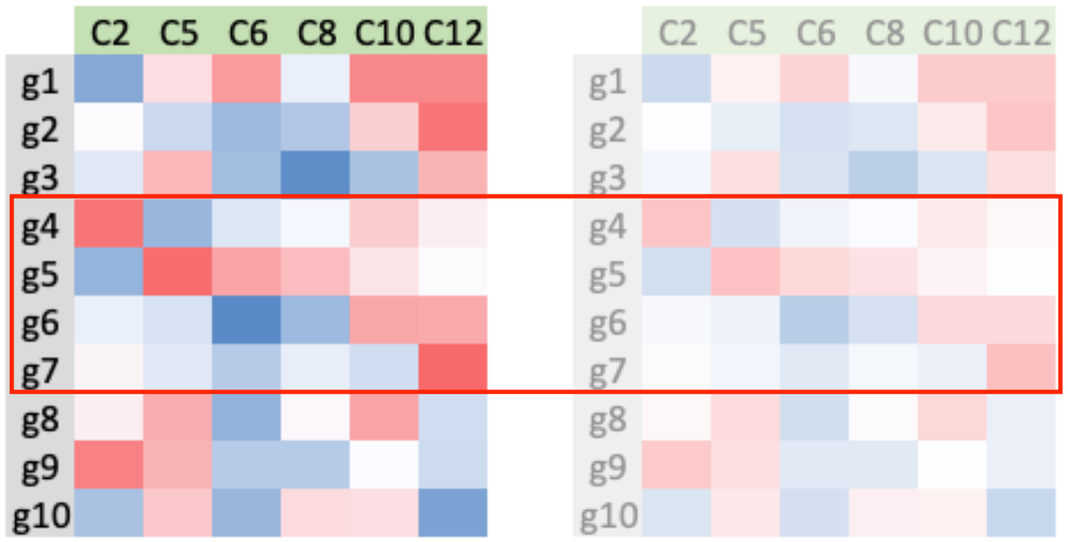
Simulate expression matrix for each cell type. — ***Data Simulator***

Based on DE tests, perform power assessments. — ***Power Assessor***




**bimodal**

$$P(Y_{gi} = y_{gi}) = (1 - \pi_g)ZIP(y_{gi} \mid p_i, \lambda_i) + \pi_g LNP(y_{gi} \mid \mu_g, \sigma_g)$$




**scenario II**

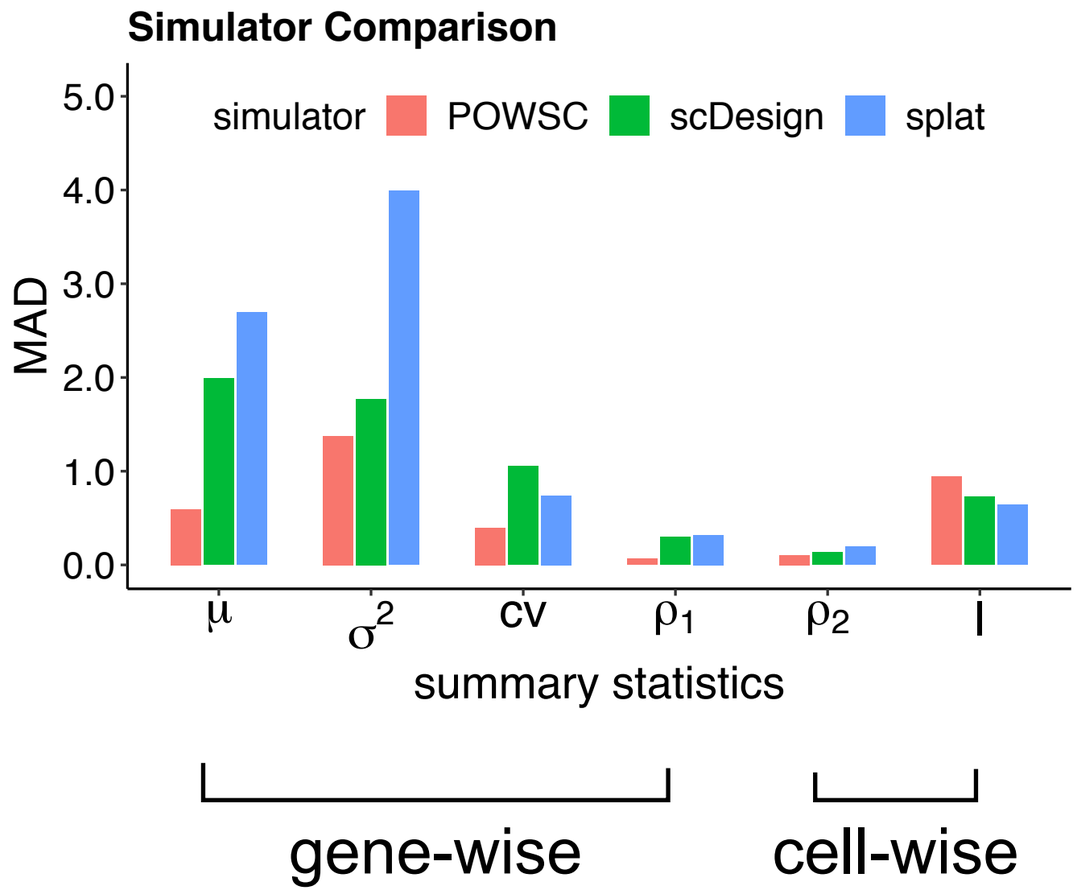
**scenario I**



Wu et al. (2018) Bioinformatics

# Data Simulator

- Comparison between simulated and real expression matrices.
  - 4 **gene-wise** parameters
  - 2 **cell-wise** parameters.
- Similar empirical distribution.
  - MAD & KS
- **Splat** (Oshlack, 2017) & **scDesign** (Li, 2019):
- **POWSC** can accurately estimate key parameters and simulate data.




Islam et al. (2011) Genome Research
Zappia et al. (2017) Genome Biology
Li et al. (2018) Bioinformatics

# Power Assessor

- Statistical test: Power = Pr(reject H0 | H1 is true)
- Confusion matrix (classification):

  Power = 1 – (false negative rate) = (sensitivity, recall, and TPR)

  Power = 1 – FN / (TP + FN) =TP / (TP + FN)

- **Stratified target power**

| | | Condition | |
|---|---|---|---|
| | | Postive | Negative |
| Prediction | Postive | **TP** | **FP** |
| | Negative | **FN** | **TN** |



Wu et al. (2014) Bioinformatics

# Power Assessor

**Stratified target power**

- **Stratify the genes by the interest categories:**
  - e.g., average expression (μ) $10 < \mu < 20$;
  - e.g., zero rate (dropout rate ρ) $0.1 < \rho < 0.2$.
- **High DE**: DE with large effect size; e.g., lfc > 1; △ > 0.1.
- Within a stratum: **Power = RD / (RD + FN).**

| Stratified power | | True DE status | | | |
|---|---|---|---|---|---|
| | | High DE | Low DE | Non-DE | Total |
| Prediction | Postive | **RD** | **RD'** | **FD** | **CD** |
| | Negative | **FN** | **FN'** | **TN** | **CN** |

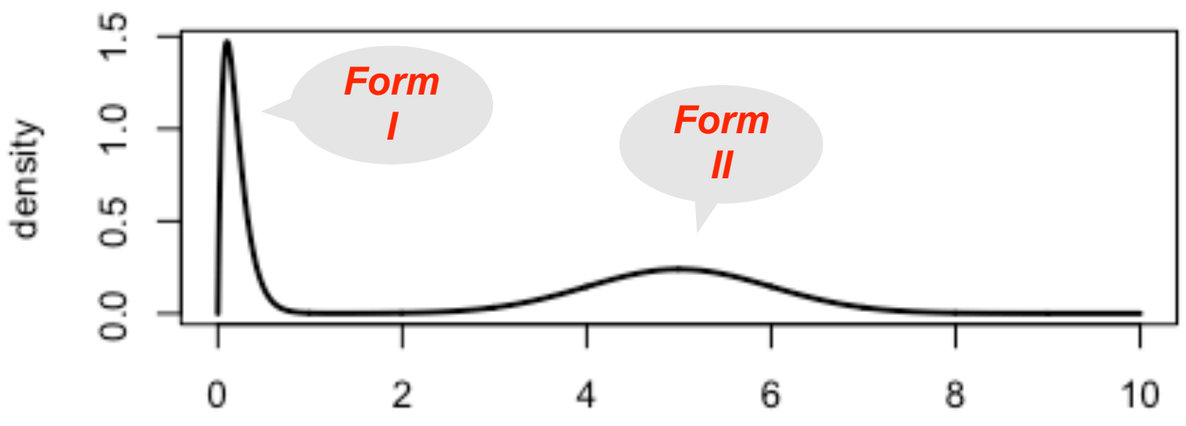
***RD:*** *Recovered high DE genes.* ***FN:*** *False negative high DE genes.*

***TD:*** *Total high DE genes.* ***CD:*** *Called DE genes.*

# DE Analysis

**Two forms of differentially expressed (DE) genes:**

- ***Form I*** (phase transition): test the differences in proportions of cells expressing a certain gene (e.g. $\rho$ 90% VS 30%).
- ***Form II*** (magnitude turning): test the quantitative changes given the gene is expressed (e.g. $\mu$ 1 VS10).

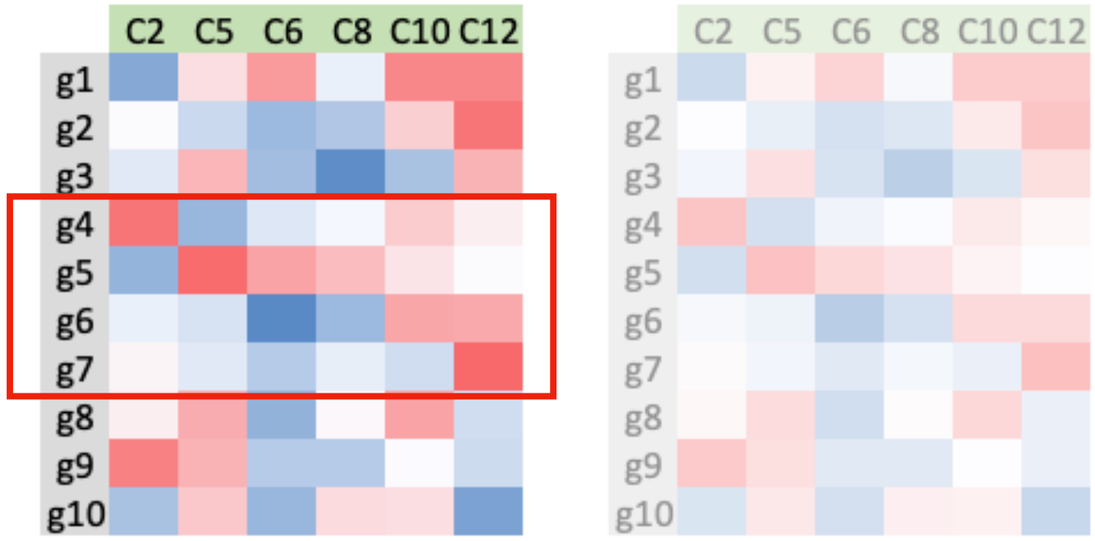



| Form | Stratum |
|---|---|
| I | rates ($\rho$) (0.8, 1] |
| II | levels ($\mu$) (0, 10] |

Wu et al. (2018) Bioinformatics

# Scenario I: One Cell Type

$$P(Y_{gi} = y_{gi}) = (1 - \pi_g)ZIP(y_{gi} \mid p_i, \lambda_i) + \pi_g LNP(y_{gi} \mid \mu_g, \sigma_g)$$







- Simulate the DE genes.
- Call DE genes using ***MAST***.

**Form I**

**perturbations**

**Form II**

$$\pi'_g = \begin{cases} \pi_g + \delta \ (\pi_g < 0.5) \\ \pi_g - \delta \ (\pi_g \geq 0.5) \end{cases}$$

$$\delta \sim uniform(0.1, 0.3)$$

$$\mu'_g = \kappa_g \mu_g$$

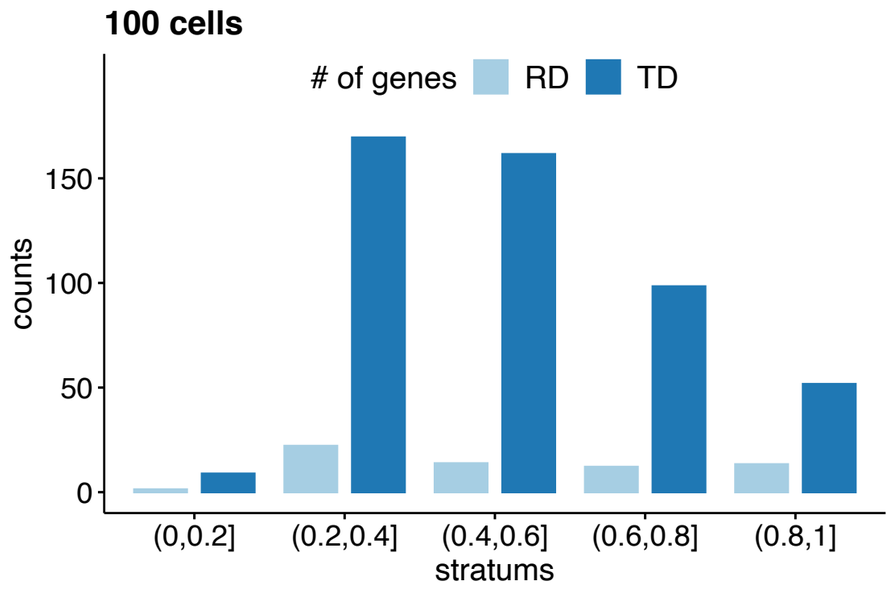
$$\kappa_g \sim 0.5 \times N(-1, 1) + 0.5 \times N(1, 1)$$

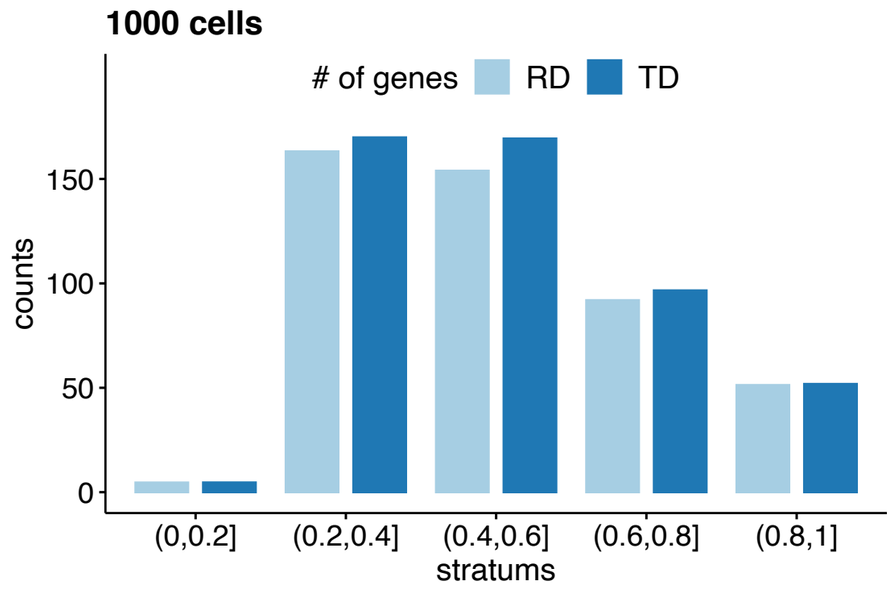


Finak et al. (2016) Genome Biology

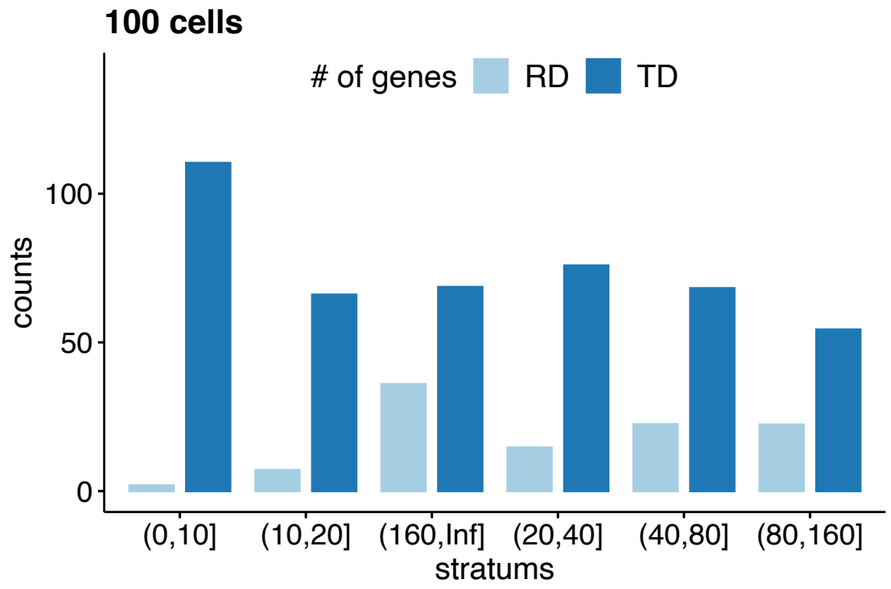
# Result for Scenario I

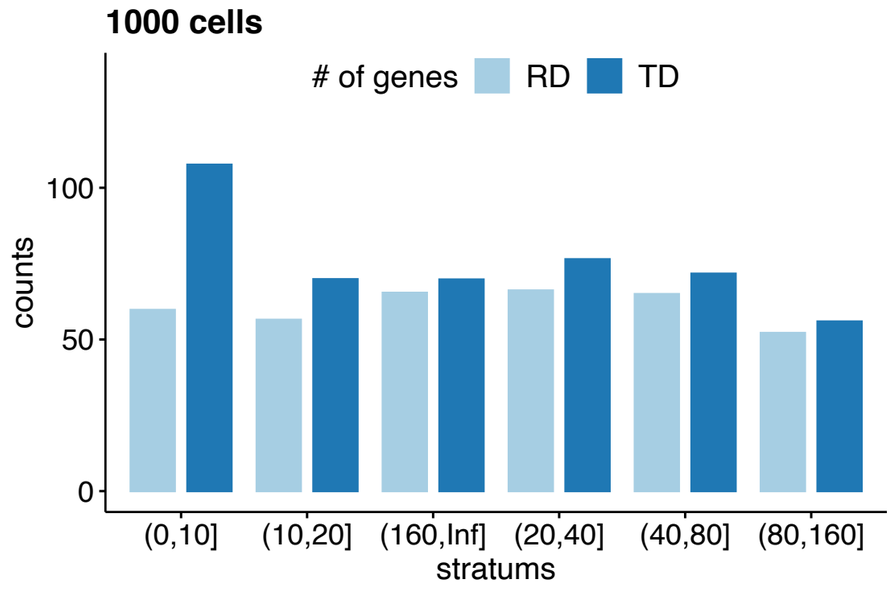
**Form I**



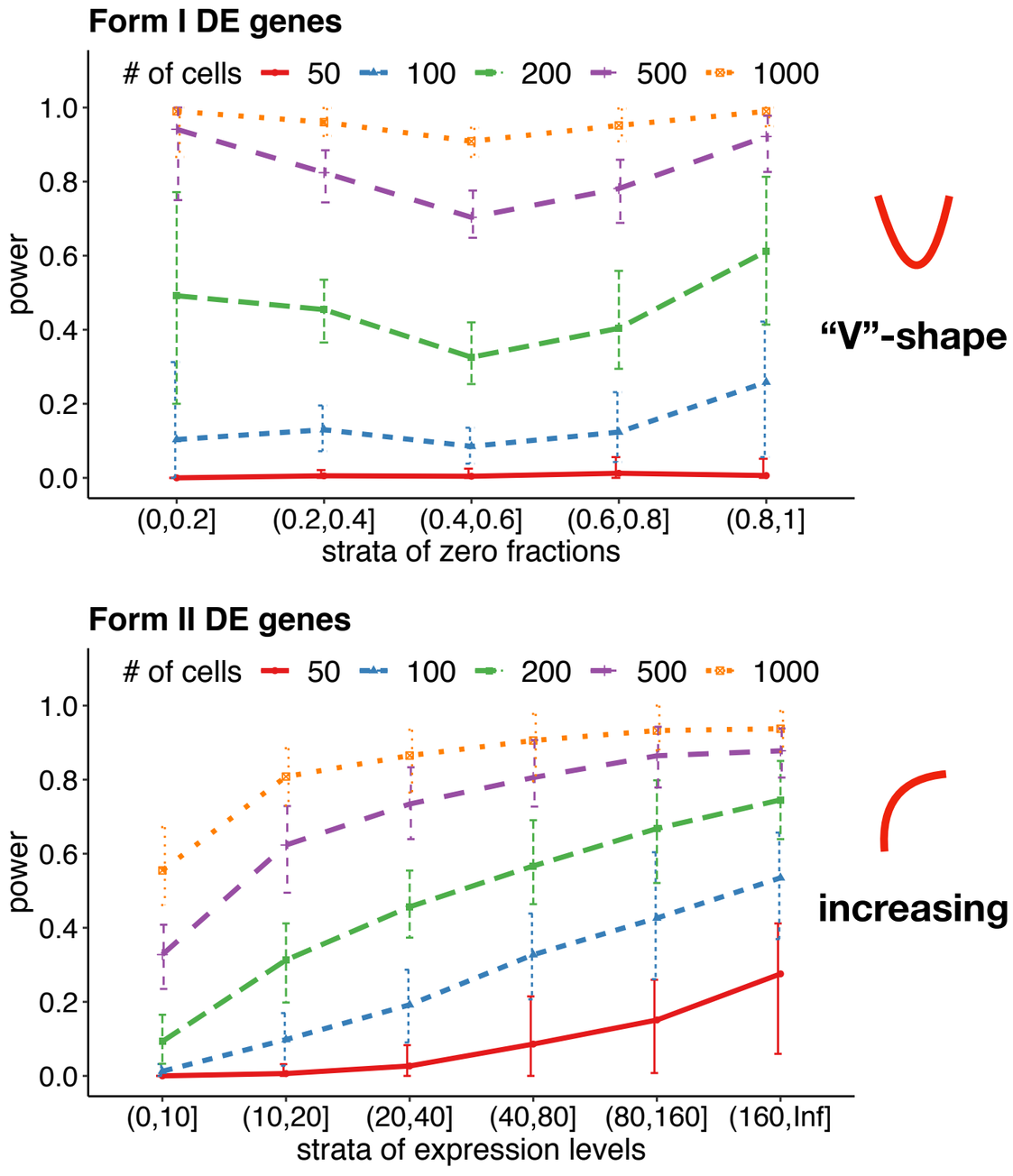

**Form II**



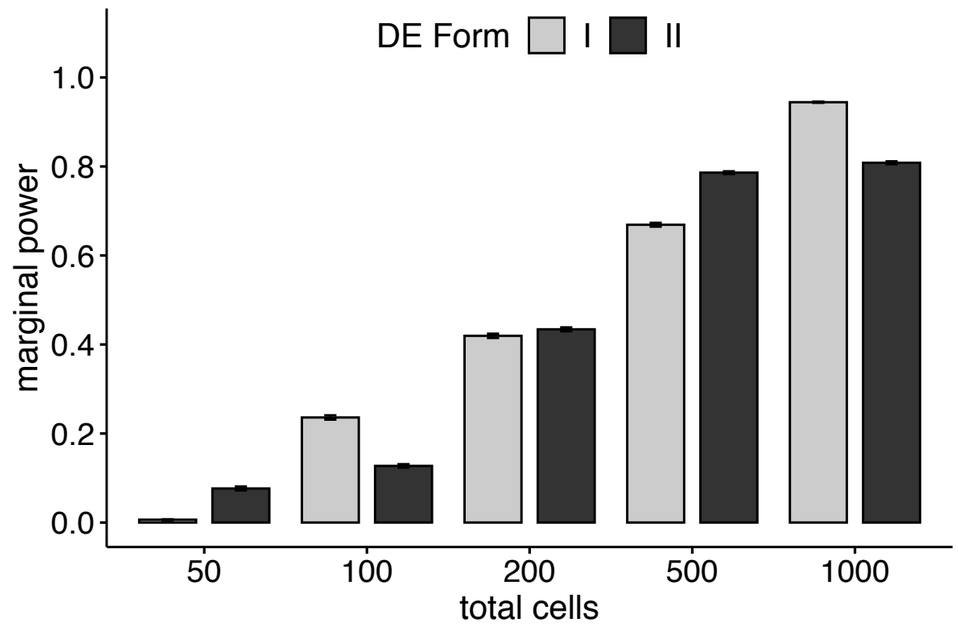

Simulate a series of datasets with different sample sizes.

# Result for Scenario I




"V"-shape

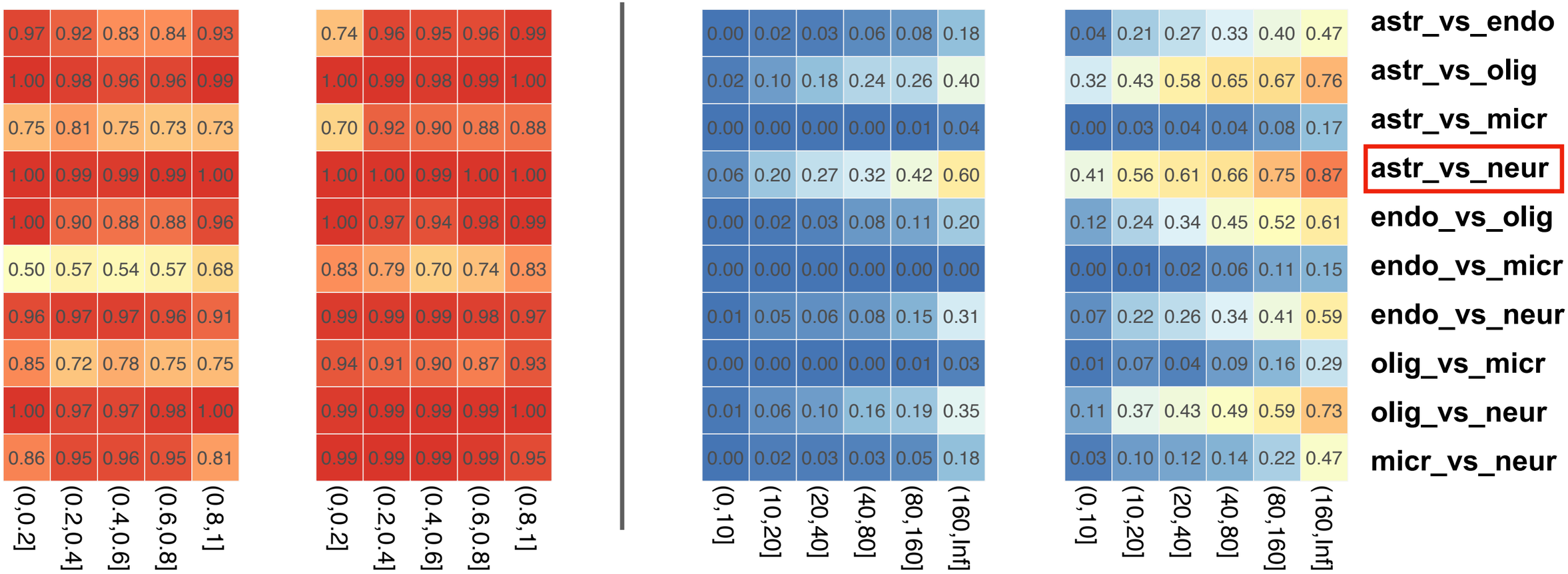
increasing




**Findings:**

1. Large sample size: better power evaluation.
2. Power curve shape.
3. Guidance for Form II: remove the lowly-expressed genes.

# Result for Scenario II




# of cells 2000 -> 6000

**Findings**: the power for cell type ***astr_vs_neur*** comparison (0.49, 0.23) is the highest.

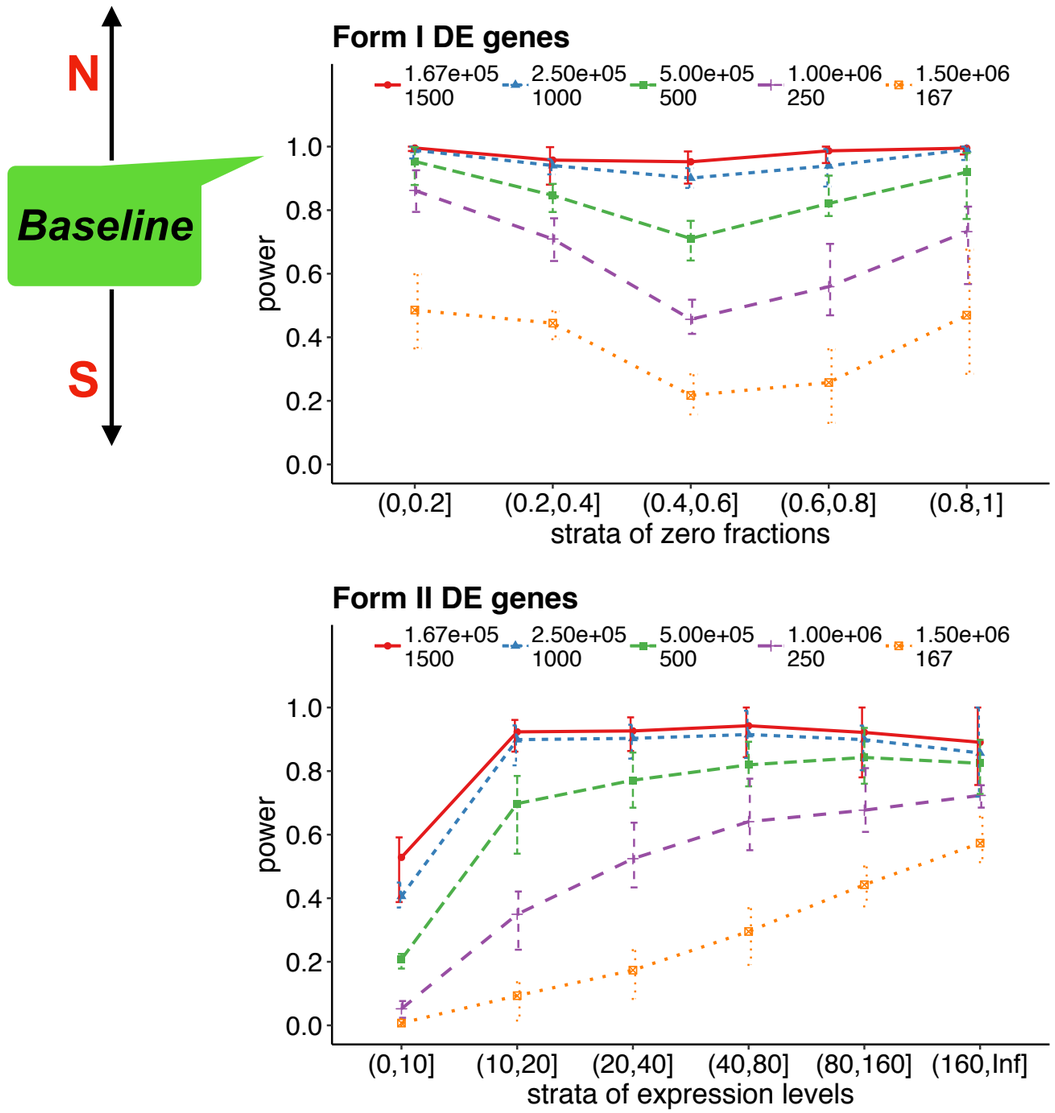
**Notes**: marker genes detection (cell proportion).

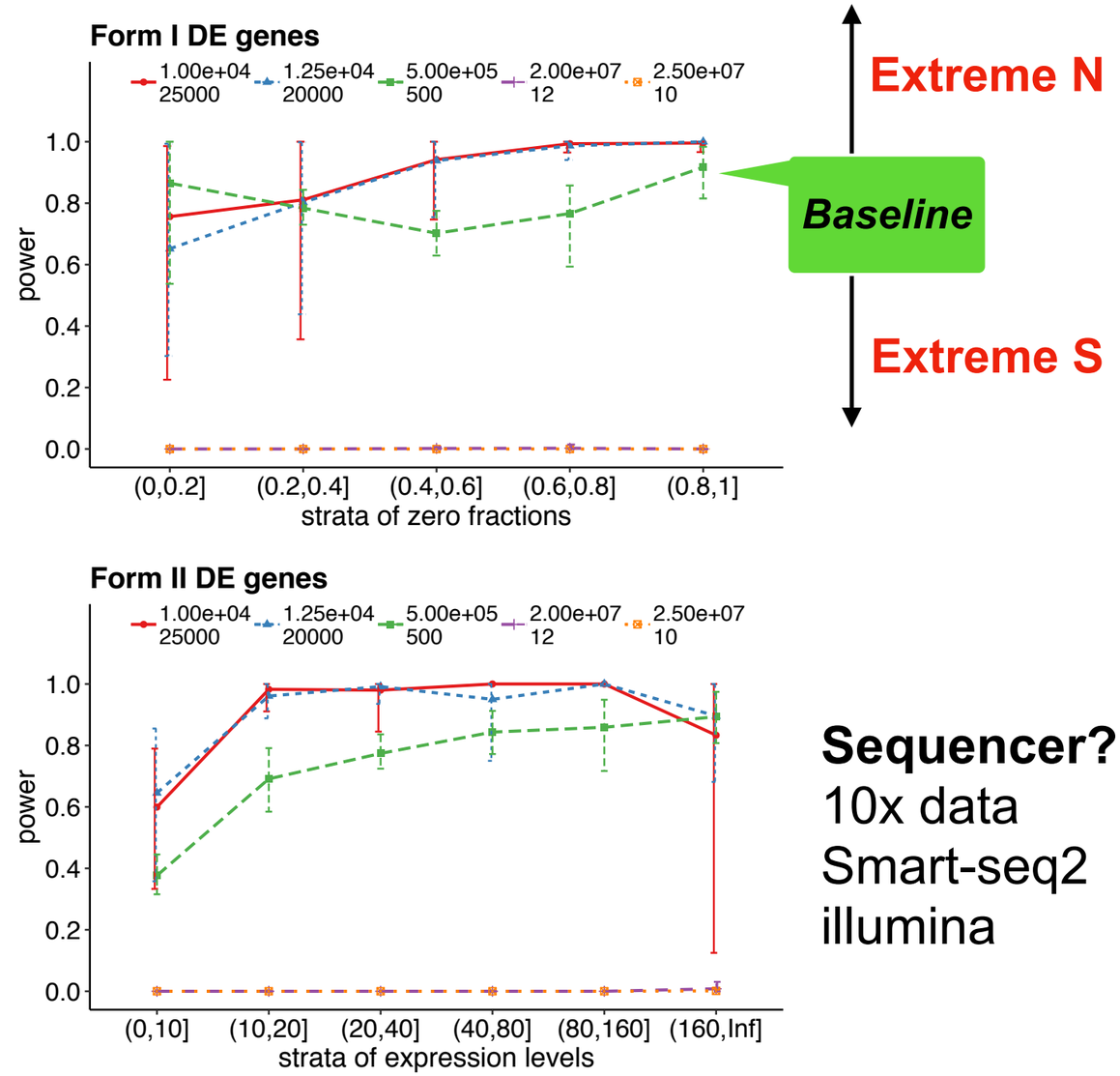


Villani et al. (2017) Science

# One Question

- **Tradeoffs**: average sequencing depths (**S**) V.S. total number of cells (**N**).
- **Baseline model**: **N** is 500 * **S** is 0.5 million; 2**S** * (1/2)**N;** (1/2)**S** * 2**N**.
- **Conclusion**: larger sample sizes (breadths) are more preferable. (***Really?***)

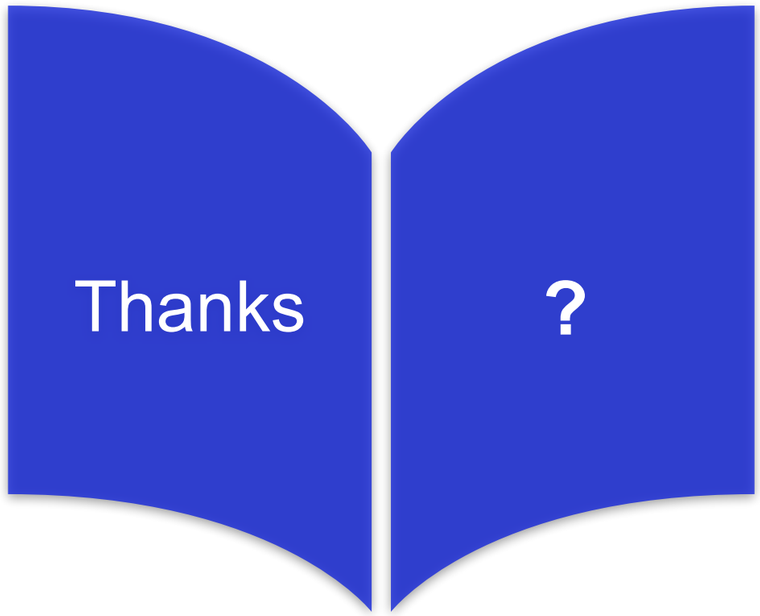



**Sequencer?**
10x data
Smart-seq2
illumina

# Acknowledge


- Prof. Hao Wu and Wu's Lab
- http://www.haowulab.org/
- Prof. Zhijin Wu @Brown


## Reference:


*Su, Kenong, Zhijin Wu, and Hao Wu. "Simulation, power evaluation and sample size recommendation for single-cell RNA-seq." Bioinformatics 36.19 (2020): 4860-4868.*