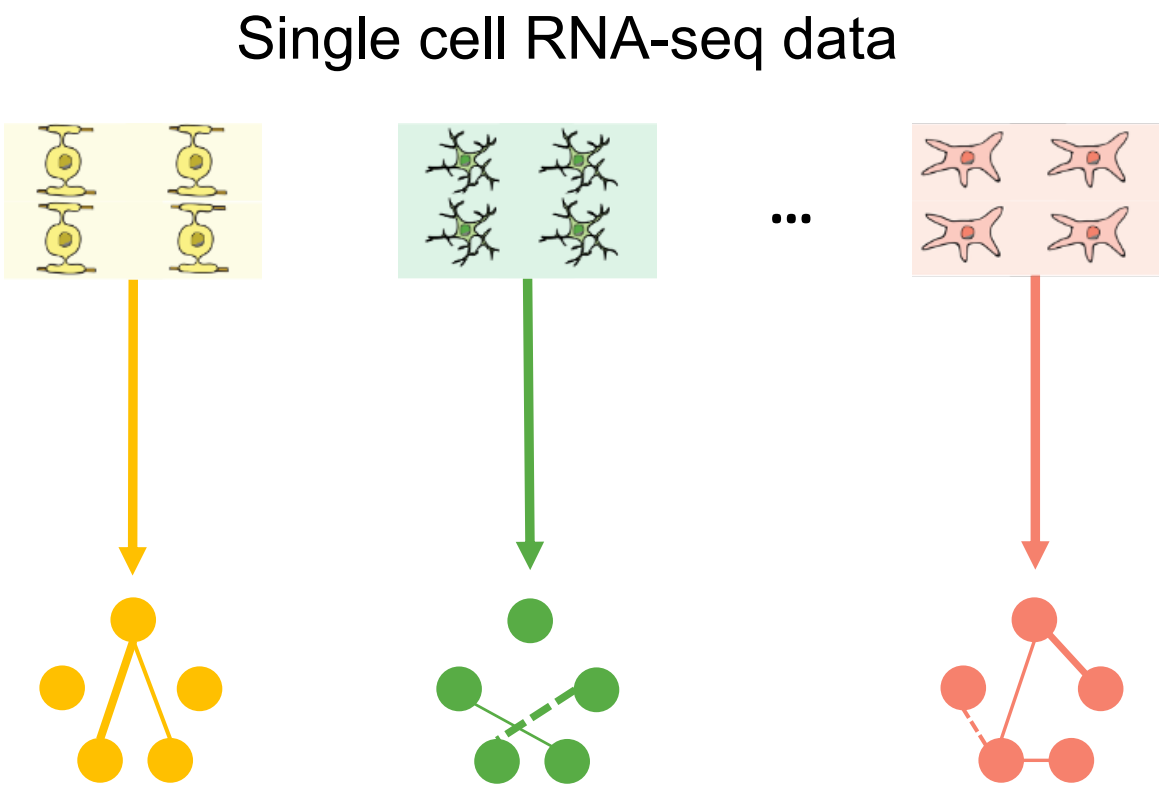
# CS-CORE: cell-type-specific co-expression inference from single cell RNA-seq data

Emma Jingfei Zhang, Emory University

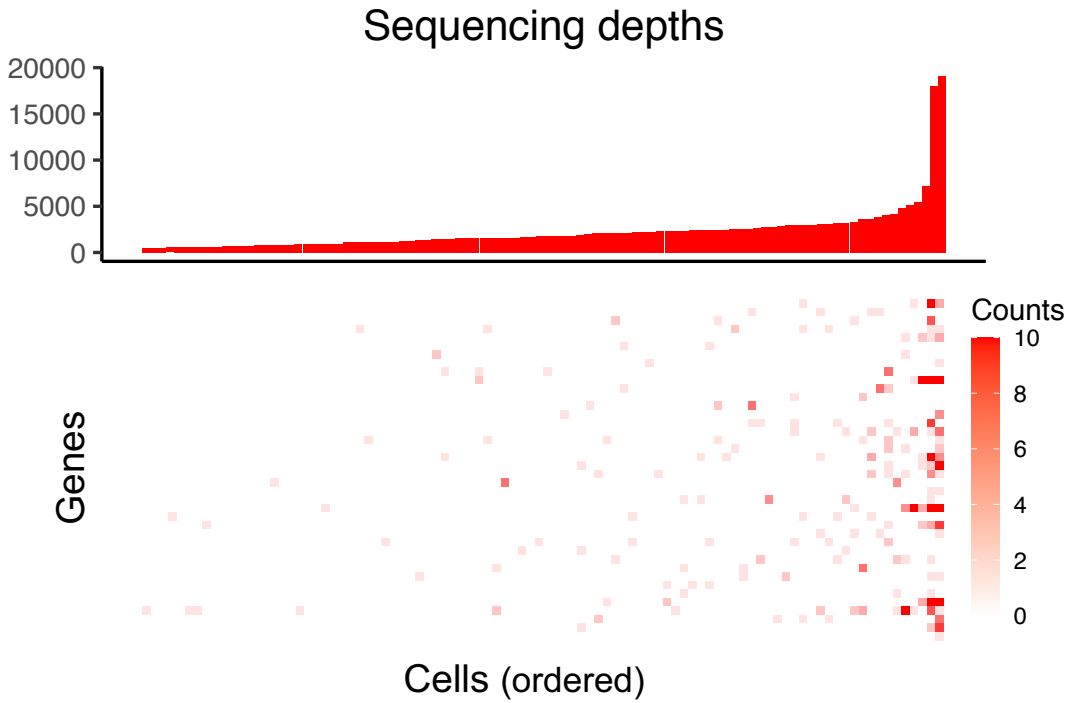
**Single Cell Plus BIRS Workshop, Banff 2023**

# Gene co-expression networks

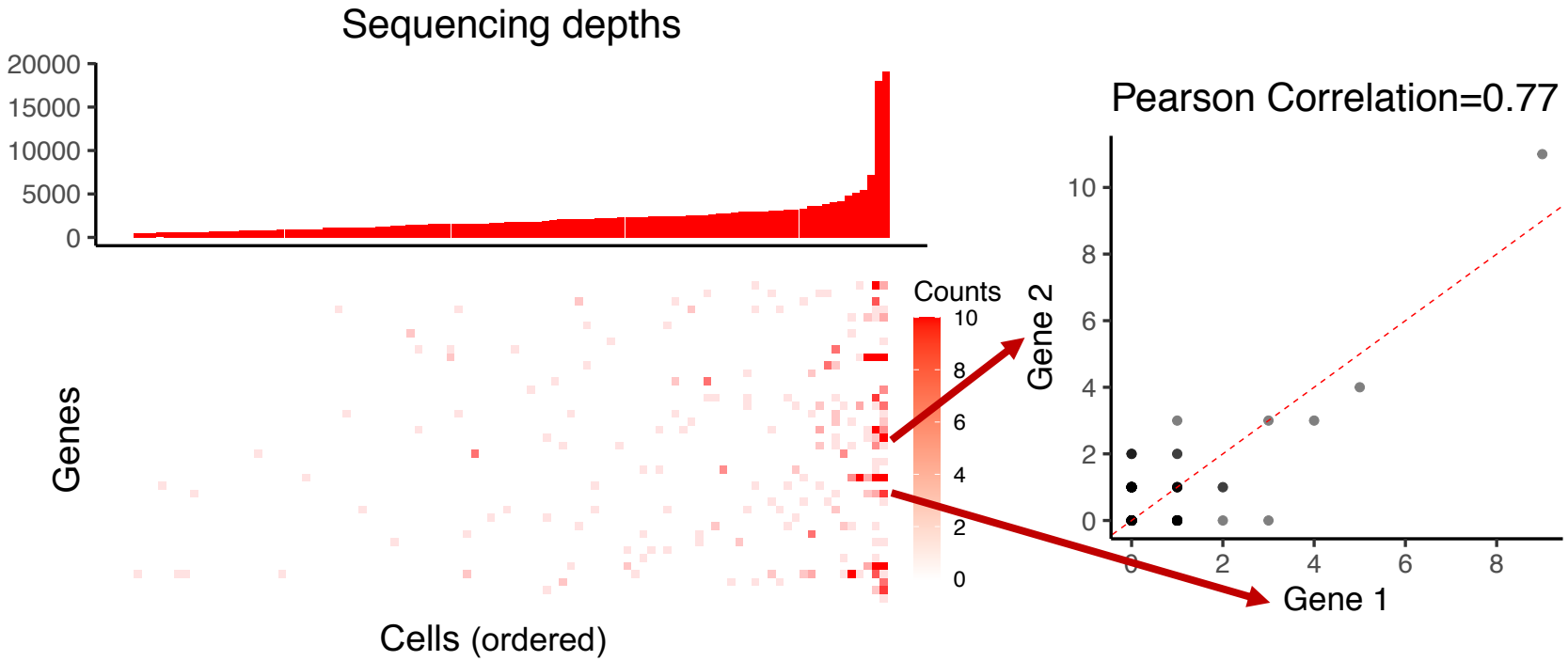
Gene co-expression networks characterize correlations of gene expression levels across biological samples.



# Confounding by sequencing depth variations



# Confounding by sequencing depth variations



# Marginal normalization?

location: (a,b)
sequencing depth: $s_1, s_2, \ldots, s_n$

$\Rightarrow$

$(a/s_1, b/s_1)$
$(a/s_2, b/s_2)$
....
$(a/s_n, b/s_n)$
$\times 10^4$

$\Rightarrow$

slope b/a
intercept 0

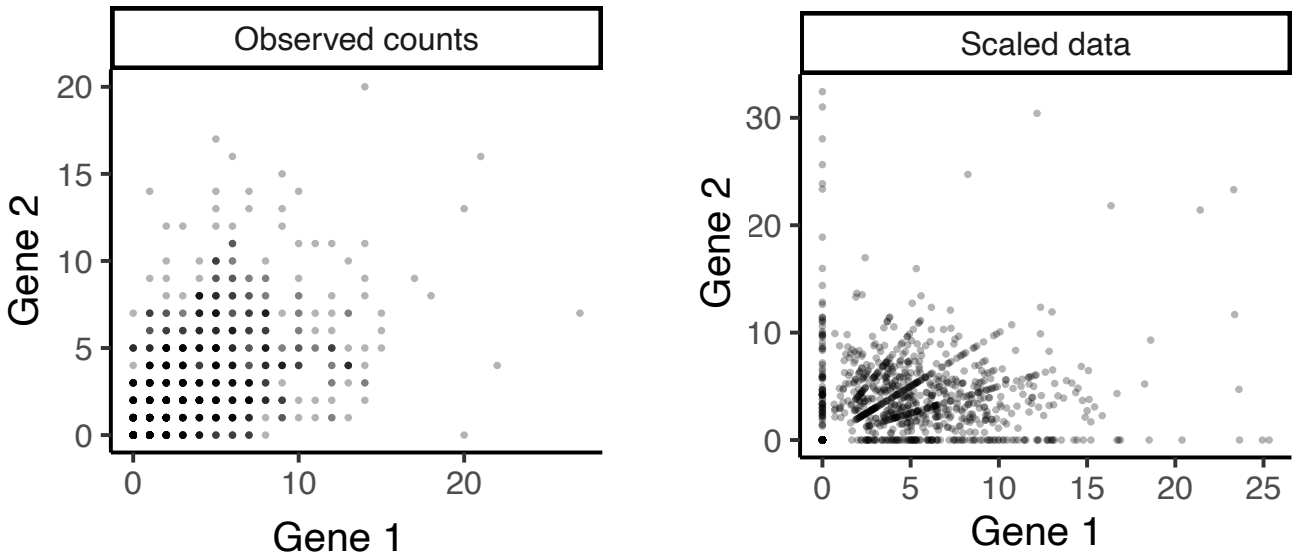



Figure: Expressions of a simulated independent gene pair in original UMI counts and scaled counts calculated as $10^4 \times x_i/s_i$, where $s_i$ is sequencing depth.

# Marginal normalization?

location: (a,b)
sequencing depth: $s_1, s_2, \ldots, s_n$

$\Rightarrow$

$(\log(a/s_1), \log(b/s_1))$
$(\log(a/s_2), \log(b/s_2))$
$\ldots$
$(\log(a/s_n), \log(b/s_n))$
$+\log(10^4)$

$\Rightarrow$

slope 1
intercept log(b)-log(a)

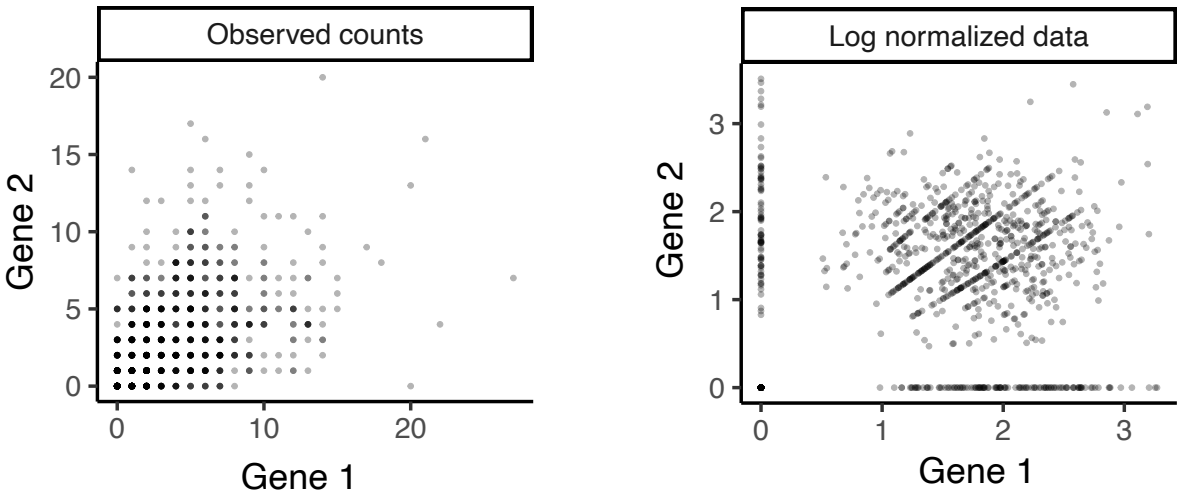



Figure: Expressions of a simulated independent gene pair in original UMI counts and log normalized counts calculated as $\log(10^4 \times x_i/s_i + 1)$, where $s_i$ is sequencing depth.

# Existing methods

- Generic methods applied on log normalized data
  - Pearson
  - Spearman
- Methods developed for single cell data
  - baredSC [Lopez-Delisle and Delisle, 2022]
  - locCSN [Wang et al., 2021]
  - Noise Regularization [Zhang et al., 2021]
  - Normalisr [Wang, 2021]
  - propr [Quinn et al., 2017]
  - $\rho$-sctransform [Hafemeister and Satija, 2019]
  - $\rho$-analytic Pearson residual [Lause et al., 2021]
  - SpQN [Wang et al., 2022]
  - Dozer [Lu and Keleș, 2023] (to be added)

# Confounding by sequencing depth variations

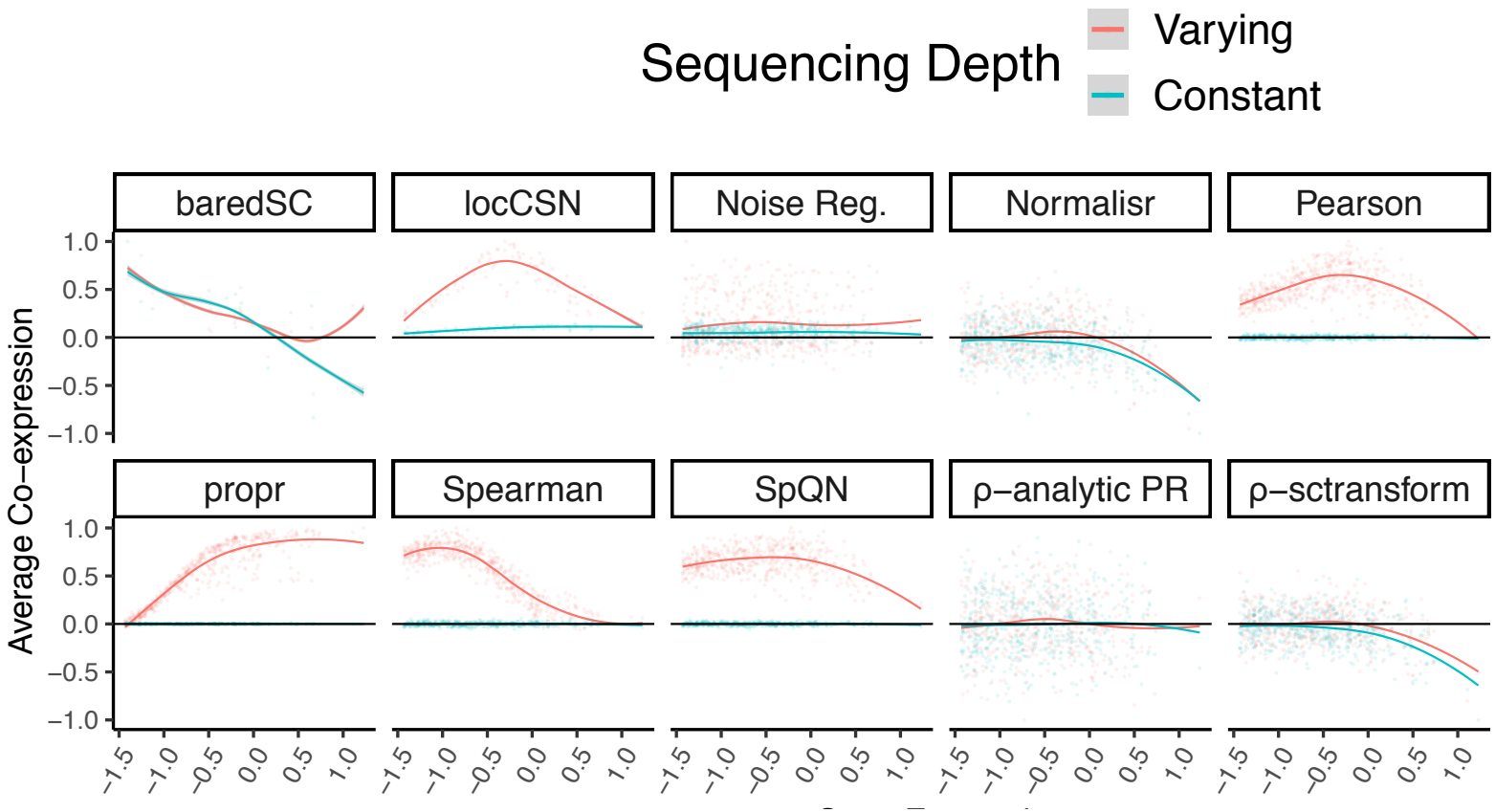
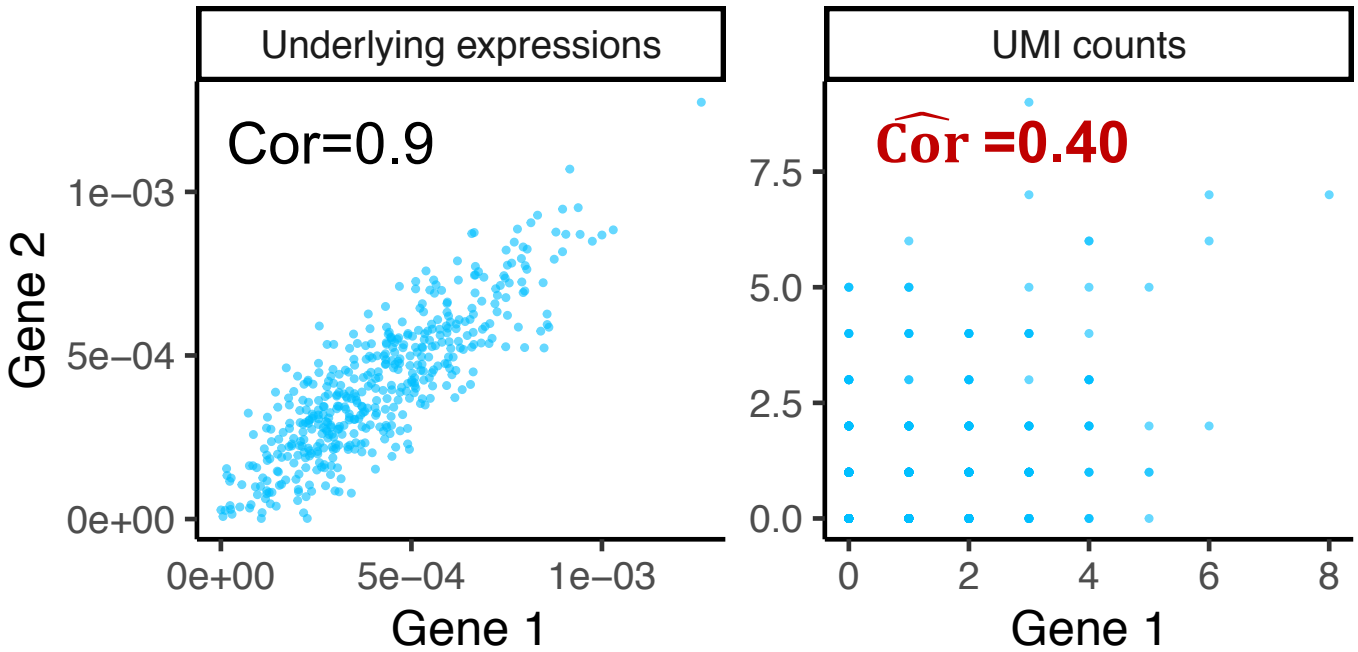



Figure: A permutation-based experiment where all gene pairs have co-expression=0.

# Attenuation by measurement noises



# Attenuation by measurement noises

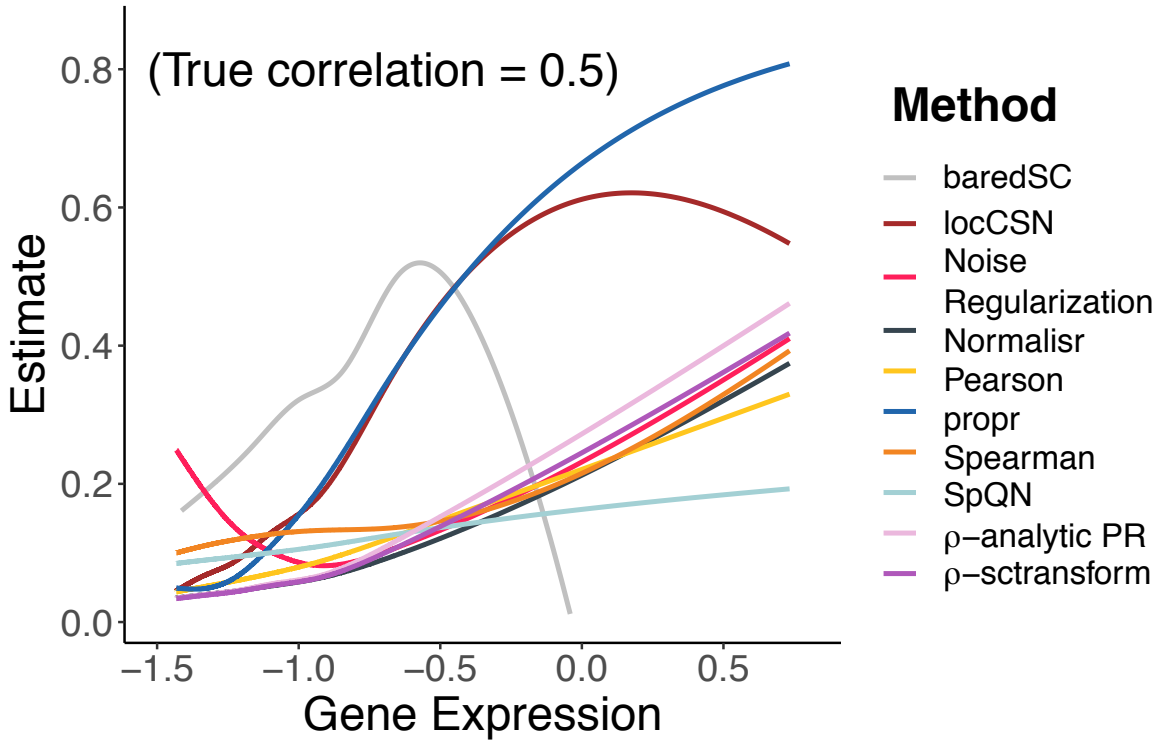



Figure: Simulated gene pairs. True correlation=$0.5$.

# Expression-measurement model in CS-CORE

- For cell $i = 1, \ldots, n$, gene $j = 1, \ldots, p$, we assume

$$(z_{i1}, \ldots, z_{ip}) \sim F_p, \quad x_{ij}|z_{ij} \sim \text{Poisson}(s_i z_{ij}),$$

  where $F_p$ is some nonnegative $p$-variate distribution.

- We measure co-expression via:

$$\rho_{jj'} = \text{Cor}(z_{ij}, z_{ij'}).$$

# Moment conditions

- For **expression level** $(z_{i1}, \ldots, z_{ip})$, denote

$$\begin{aligned} \text{Mean:} \quad & \mu_j = \mathbb{E}[z_{ij}], \\ \text{Variance:} \quad & \sigma_{jj} = \mathsf{Var}[z_{ij}], \\ \text{Covariance:} \quad & \sigma_{jj'} = \mathsf{Cov}(z_{ij}, z_{ij'}). \end{aligned}$$

- We can show for **UMI counts** $(x_{i1}, \ldots, x_{ip})$ that

$$\begin{aligned} \text{Mean:} \quad & \mathbb{E}[x_{ij}] = s_i \mu_j, \\ \text{Variance:} \quad & \mathsf{Var}[x_{ij}] = s_i \mu_j + s_i^2 \sigma_{jj}, \\ \text{Covariance:} \quad & \mathsf{Cov}(x_{ij}, x_{ij'}) = s_i^2 \sigma_{jj'}. \end{aligned}$$

# Linear regressions

From the moment conditions, we can write

$$
\begin{aligned}
x_{ij} &= s_i \mu_j + \epsilon_{ij}, \\
(x_{ij} - s_i \mu_j)^2 &= s_i \mu_j + s_i^2 \sigma_{jj} + \eta_{ij}, \\
(x_{ij} - s_i \mu_j)(x_{ij'} - s_i \mu_{j'}) &= s_i^2 \sigma_{jj'} + \xi_{ijj'},
\end{aligned}
$$

where $\mathbb{E}(\epsilon_{ij}) = 0$, $\mathbb{E}(\eta_{ij}) = 0$ and $\mathbb{E}(\xi_{ijj'}) = 0$.

# IRLS Estimation

- **Iteratively reweighted least squares** estimation:

$$\hat{\mu}_j = \min_{\mu} \sum_{i=1}^{n} w_{ij}(x_{ij} - s_i\mu)^2,$$

$$\hat{\sigma}_{jj} = \min_{\sigma} \sum_{i=1}^{n} h_{ij}[(x_{ij} - s_i\hat{\mu}_j)^2 - s_i\hat{\mu}_j - s_i^2\sigma]^2,$$

$$\hat{\sigma}_{jj'} = \min_{\sigma} \sum_{i=1}^{n} g_{ijj'}[(x_{ij} - s_i\hat{\mu}_j)(x_{ij'} - s_i\hat{\mu}_{j'}) - s_i^2\sigma]^2.$$

# Test for independence

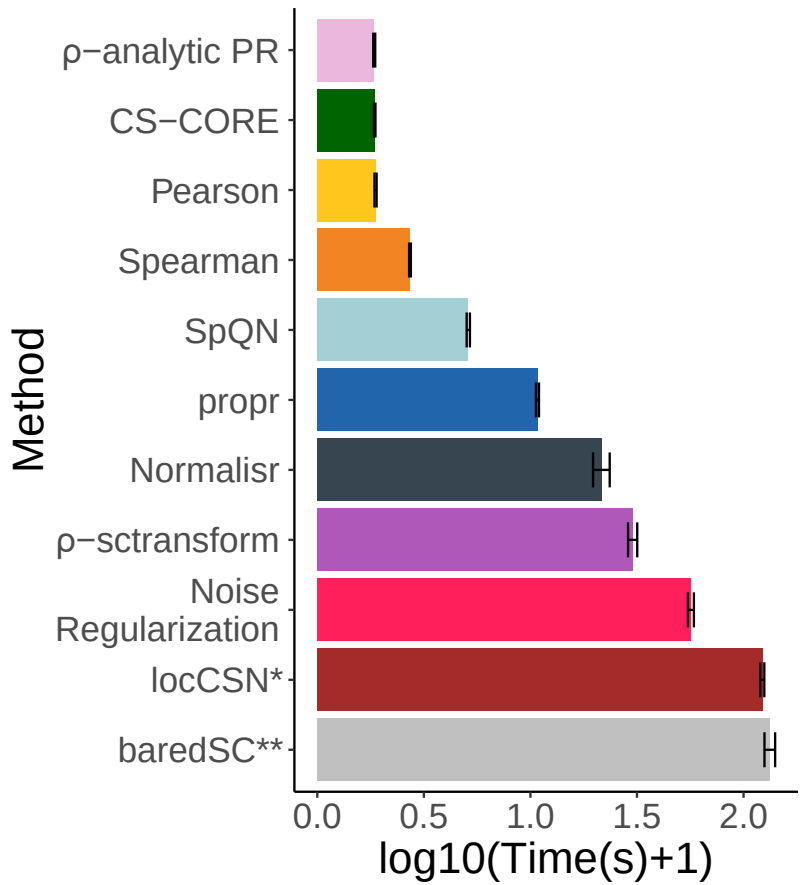
- $H_0$: $\underbrace{Z_j \text{ and } Z_{j'}}_{\substack{\text{underlying expression} \\ \text{levels from genes } j, j'}}$ are independent.
- We define the test statistic

$$T_{jj'} = \frac{\sum_i s_i^2 (x_{ij} - s_i \mu_j)(x_{ij'} - s_i \mu_{j'}) g_{ijj'}}{\sqrt{\sum_i s_i^4 (s_i \mu_j + s_i^2 \sigma_{jj})(s_i \mu_{j'} + s_i^2 \sigma_{j'j'}) g_{ijj'}^2}}.$$

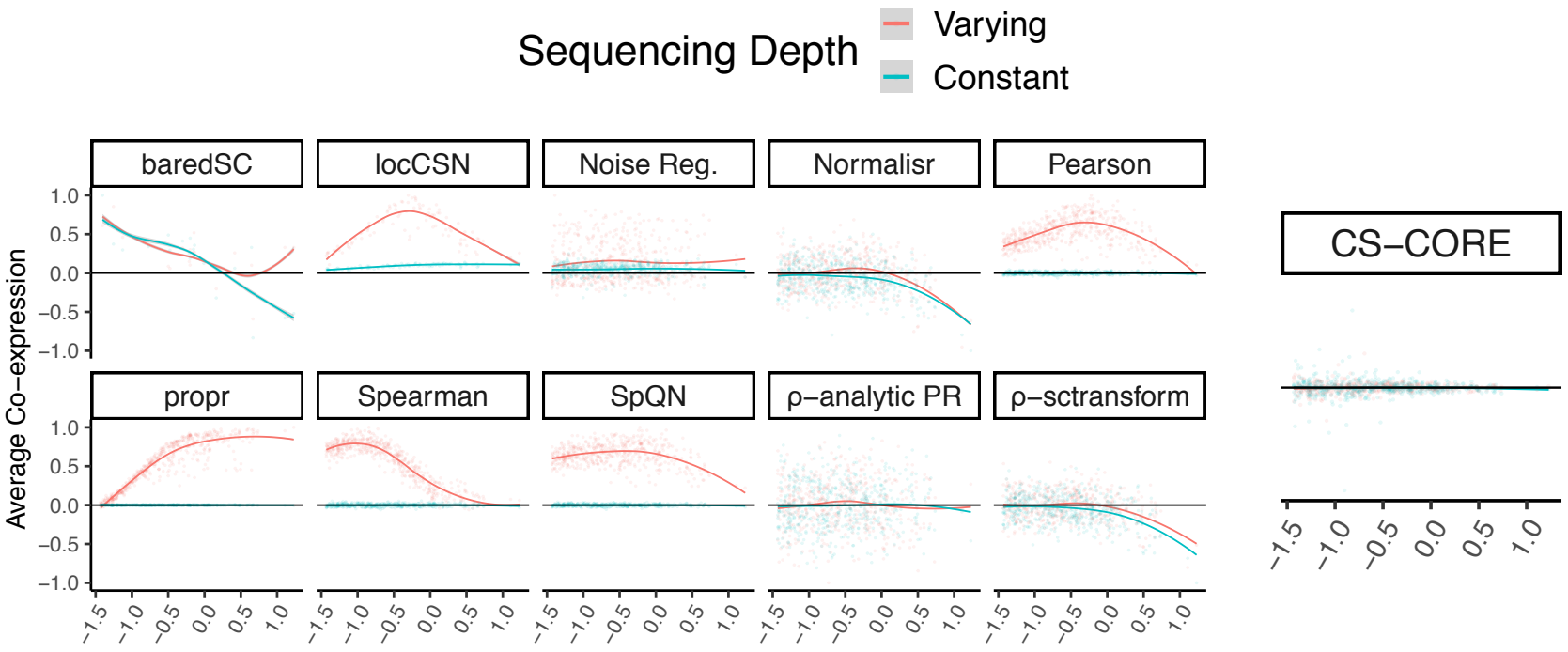
- Under $H_0$, $T_{jj'}$ is asymptotically $\mathcal{N}(0, 1)$.

# CS-CORE is fast

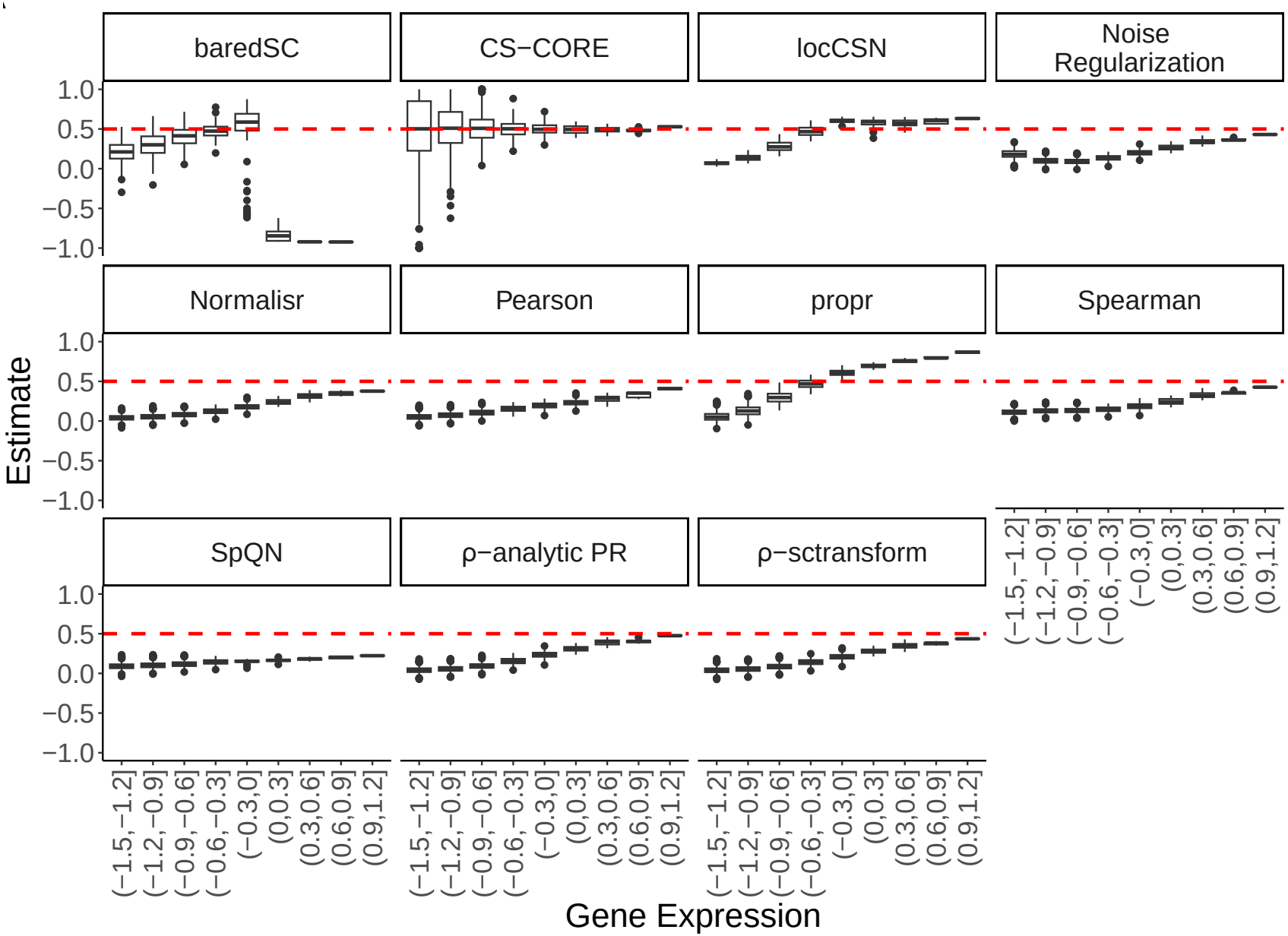
- CS-CORE takes 10s to estimate a co-expression network with 500 genes.



# CS-CORE is not confounded by sequencing depths



# CS-CORE is not biased by measurement noises



# Systematic evaluations of CS-CORE

## AIzheimer's disease (AD) and COVID-19 scRNA-seq data

| Data sets | Lau et al. [2020] | Mathys et al. [2019] | Morabito et al. [2021] | Wilk et al. [2020] | Unterman et al. [2022] |
|---|---|---|---|---|---|
| Tissue | Brain | Brain | Brain | PBMC | PBMC |
| Disease | AD | AD | AD | COVID-19 | COVID-19 |
| #cells/nucleus | 169,500 | 70,634 | 61,472 | 44,721 | 153,554 |
| #cell types | 6 | 8 | 7 | 13 | 29 |
| Median seq depth | 2,600 | 1,474 | 6,382 | 1,946 | 3,618 |
| #samples | 21 | 48 | 18 | 14 | 31 |

# Biologically interpretable

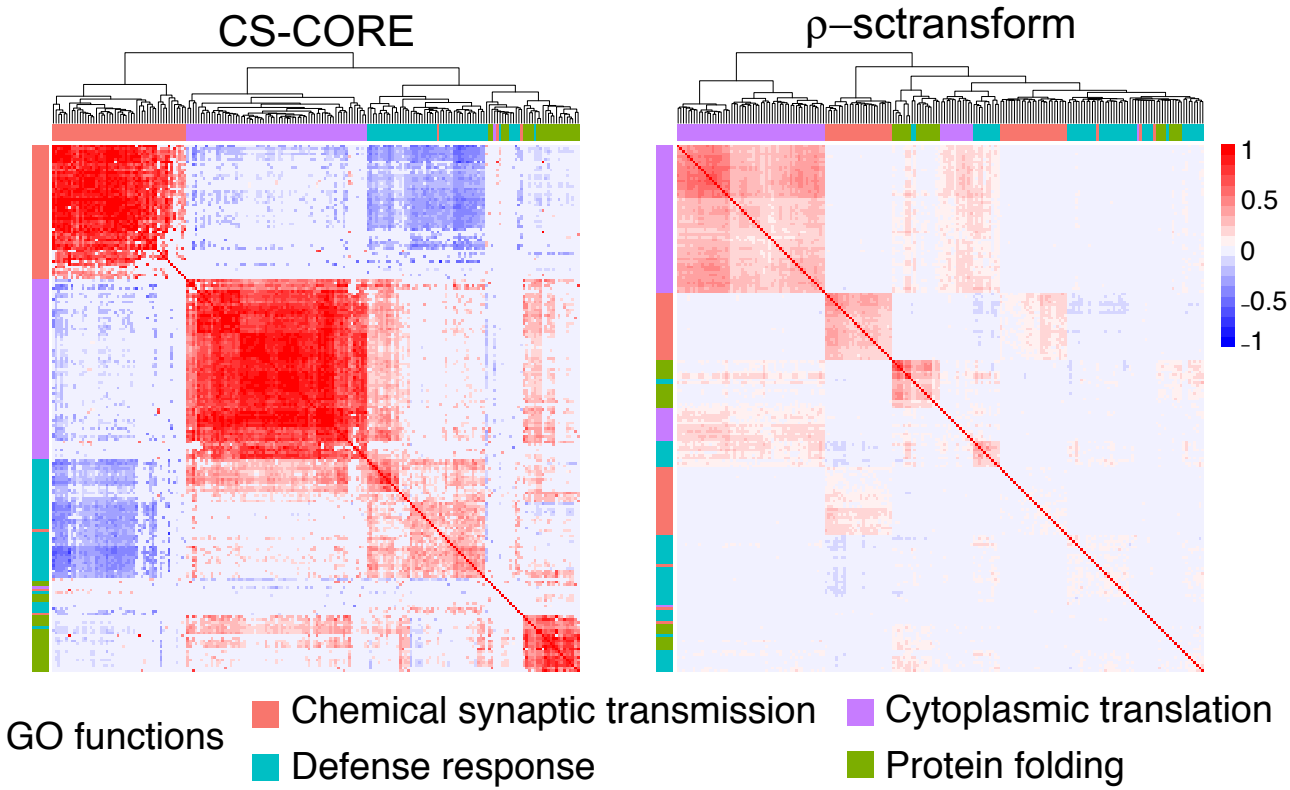



Figure: CS-CORE accurately grouped genes by biological functions in microglia using single cell data from Lau et al. [2020].

# Reproducible and consistent with known gene pairs

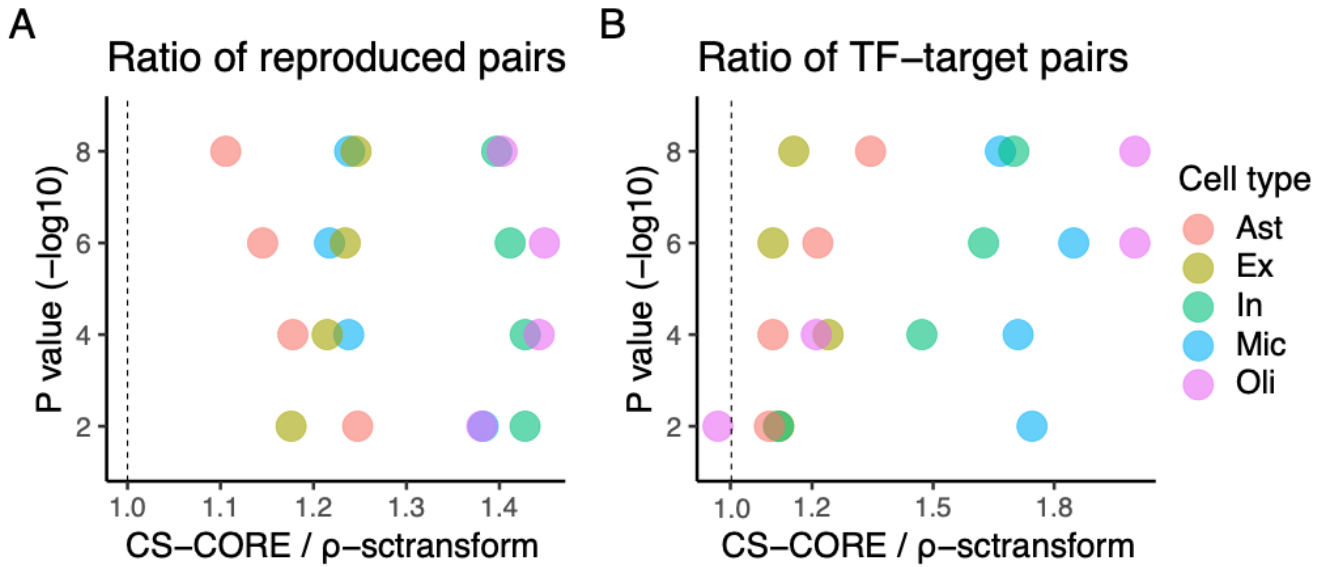
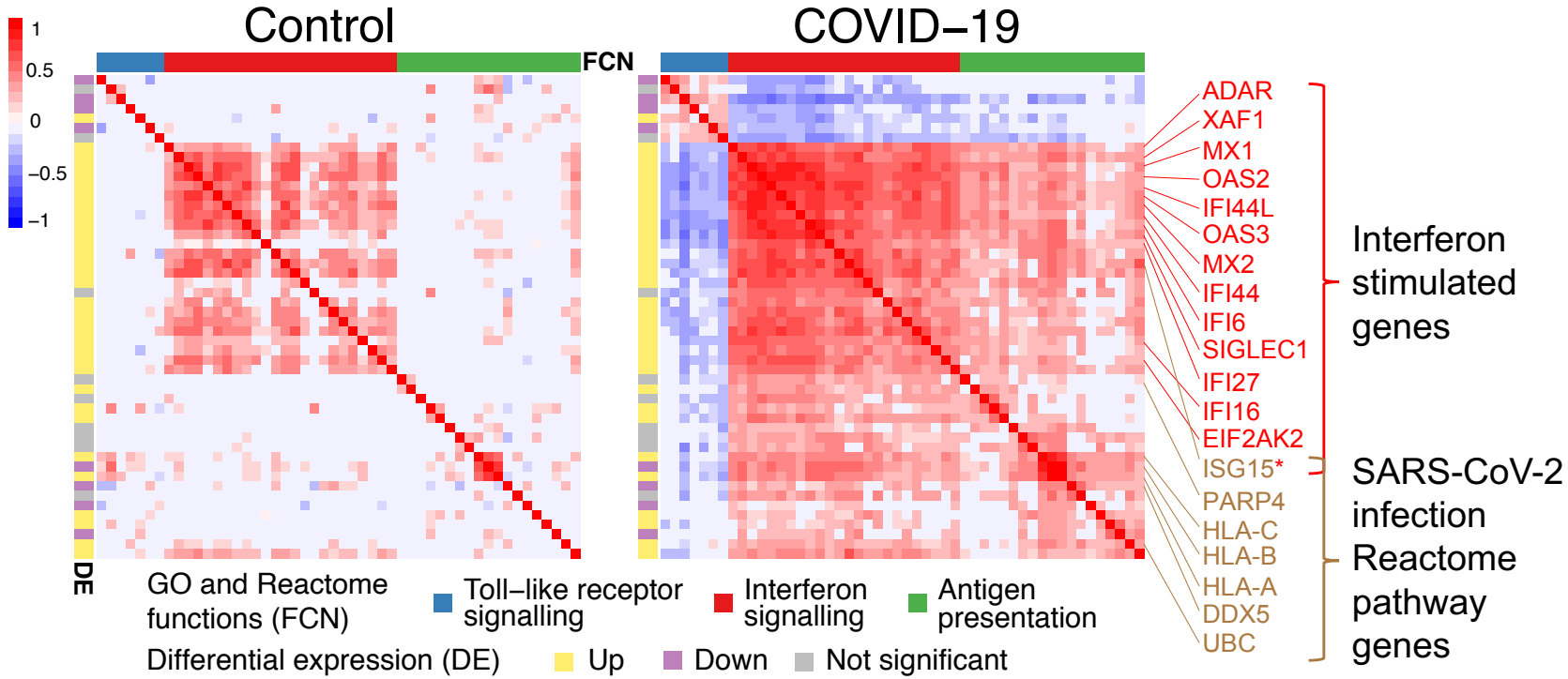



Figure: CS-CORE uncovered co-expressions that are more reproducible and more consistent with known transcription factor (TF)-target pairs using single cell data on brain from Mathys et al. [2019] and Lau et al. [2020].

# CS-CORE identified up-regulated co-expressions in Interferon signaling pathway from COVID-19 blood samples



# Team, paper and software

- CS-CORE makes minimal distribution assumptions, is fast and provides a valid test (also fast).
- **Team**: Chang Su (Emory U.), Zichun Xu, Xinning Shan, Biao Cai, Hongyu Zhao

- **Paper**: Cell-type-specific co-expression inference from single cell RNA-sequencing data, bioRxiv, 2022.