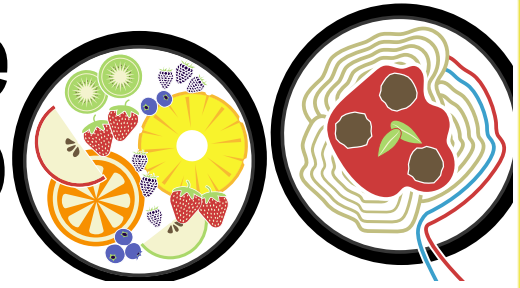
# Modelling group heteroscedasticity in pseudo-bulk single-cell RNA-seq data

Prof. Matt Ritchie

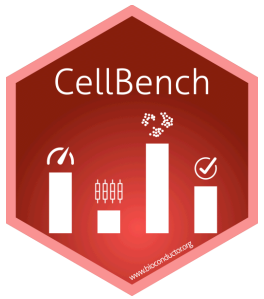
Single-Cell Plus Meeting, BIRS, Banff

4th July 2023

Ritchie Lab



# Biology in the Matrix

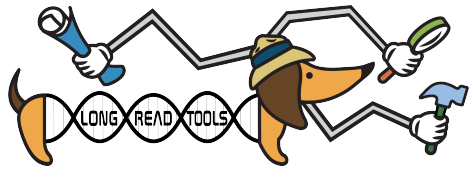





Genomic features

Samples / Cells



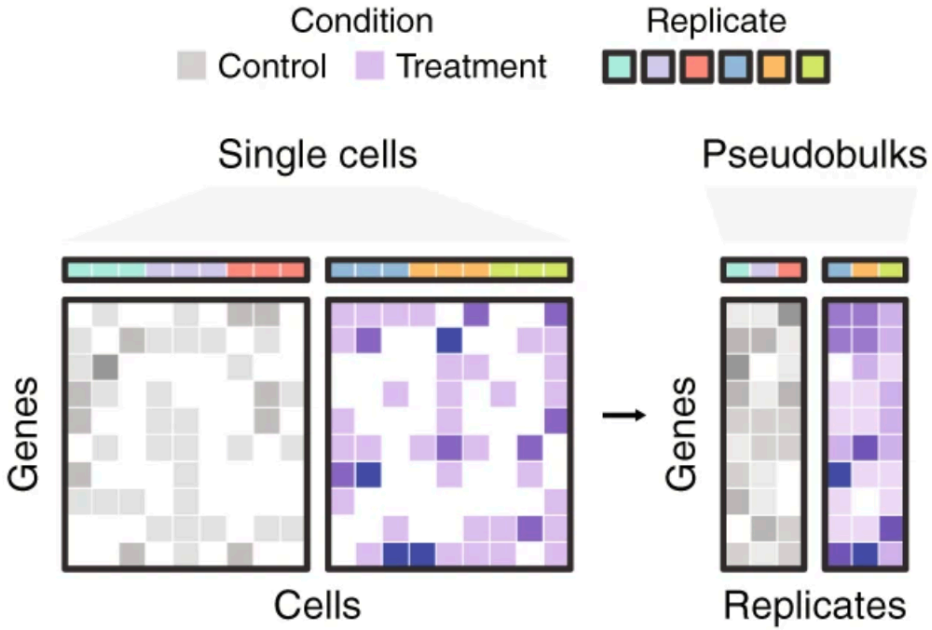
# Multi-group single-cell RNA-seq datasets are becoming more common

**COVID-19 patients**

- 3 Healthy controls (HC)
- 3 Severe samples
- 23 Moderate samples
- 11 Asymptomatic samples

Zhao *et al.* (2021) Signal Transduct Target Ther. 6:342.

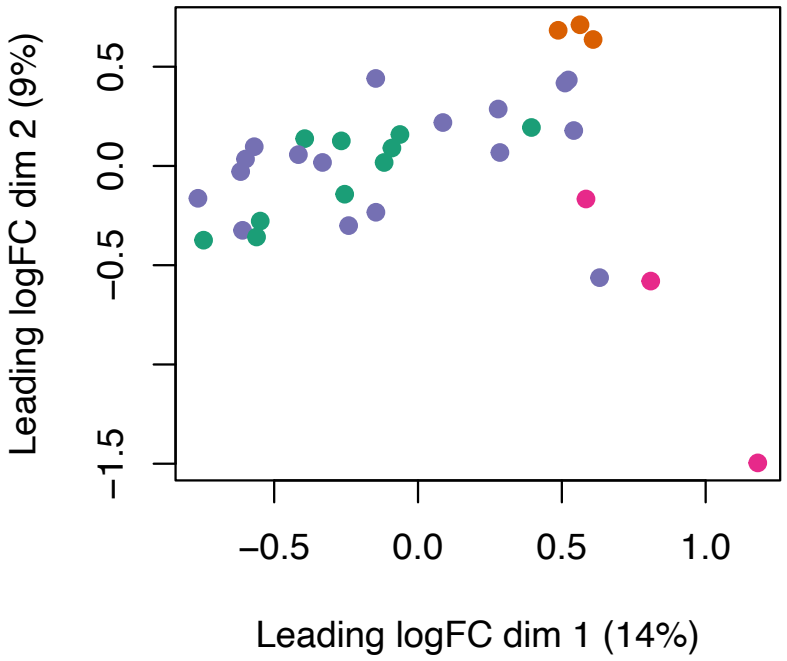
**Pseudo-bulk + Favourite DE tool**



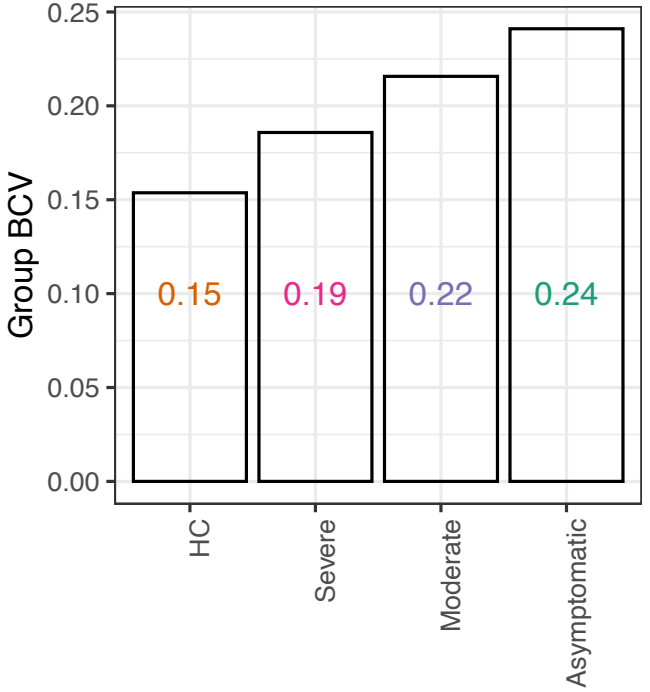

Squair *et al.* (2021) Nat Commun. 12:5692. (Fig 2a)
Crowell *et al.* (2020) Nat Commun. 11:6077.

# Group-specific variation is frequently observed in pseudo-bulk scRNA-seq data

MDS plot
$CD56^{dim}$ $CD16^{+}$ NK cells




Biological Coefficient
of Variation (BCV) by group

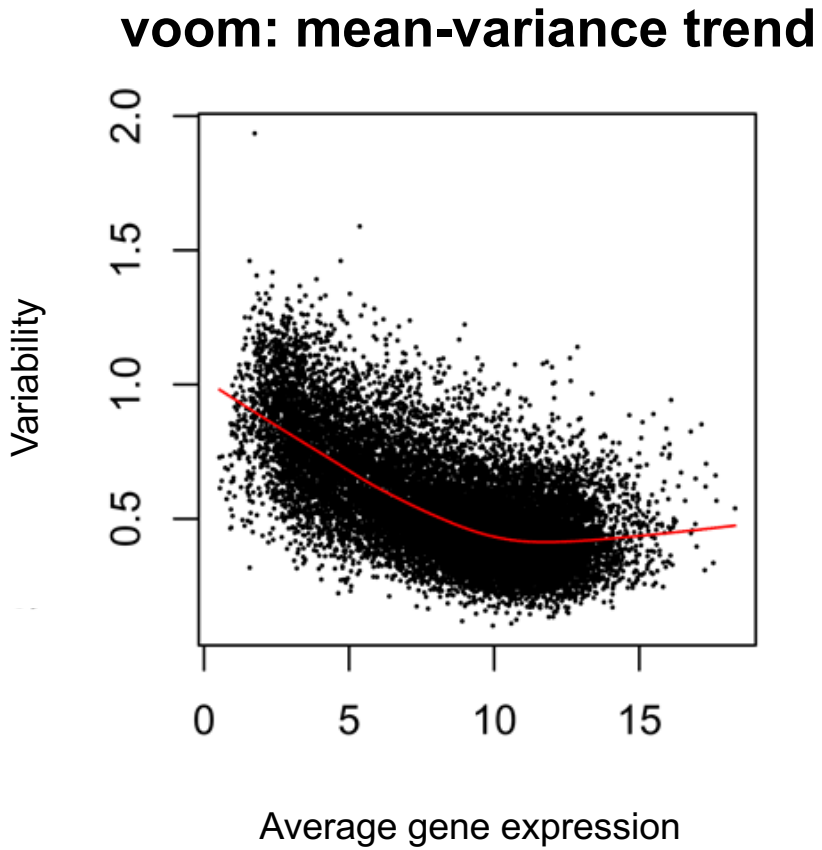



Can we better model the heteroscedasticity present in our data ?

# Strategy 1: Model group-specific mean-variance trends using *voomByGroup*

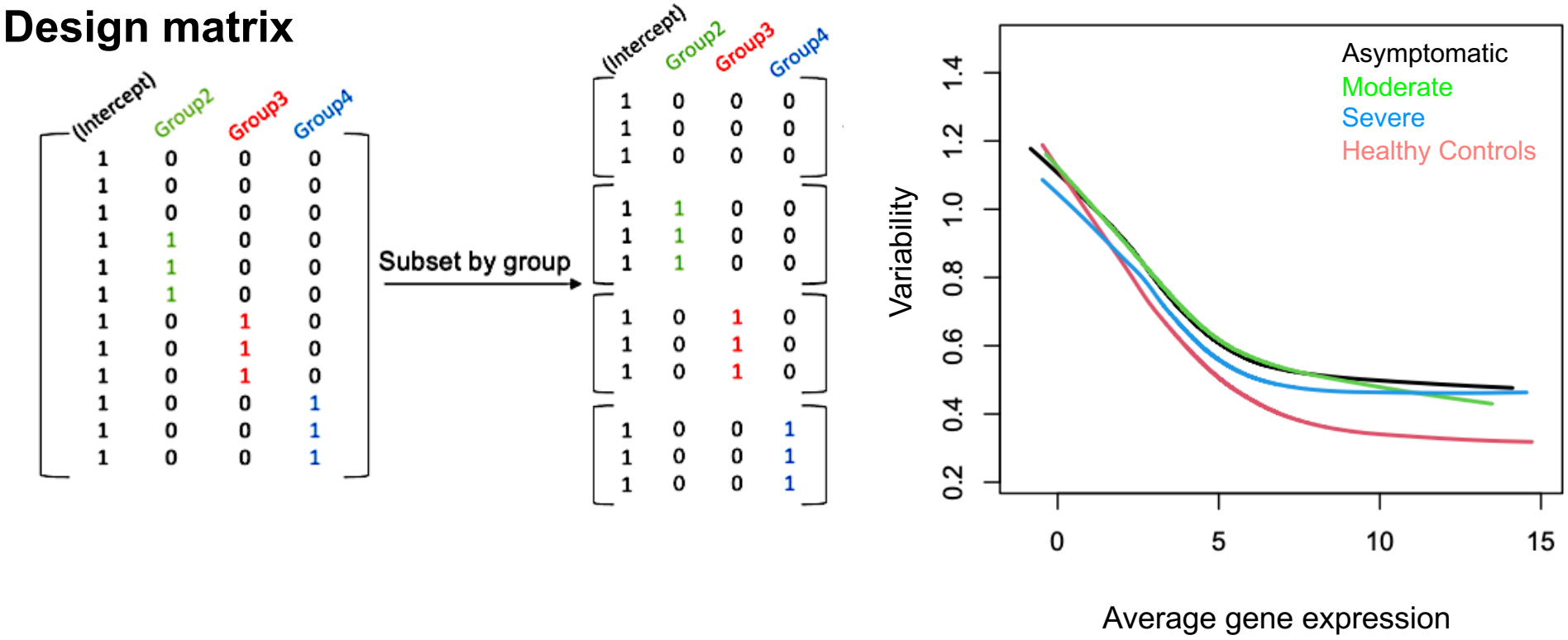
**Design matrix**

| (Intercept) | Group2 | Group3 | Group4 |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |



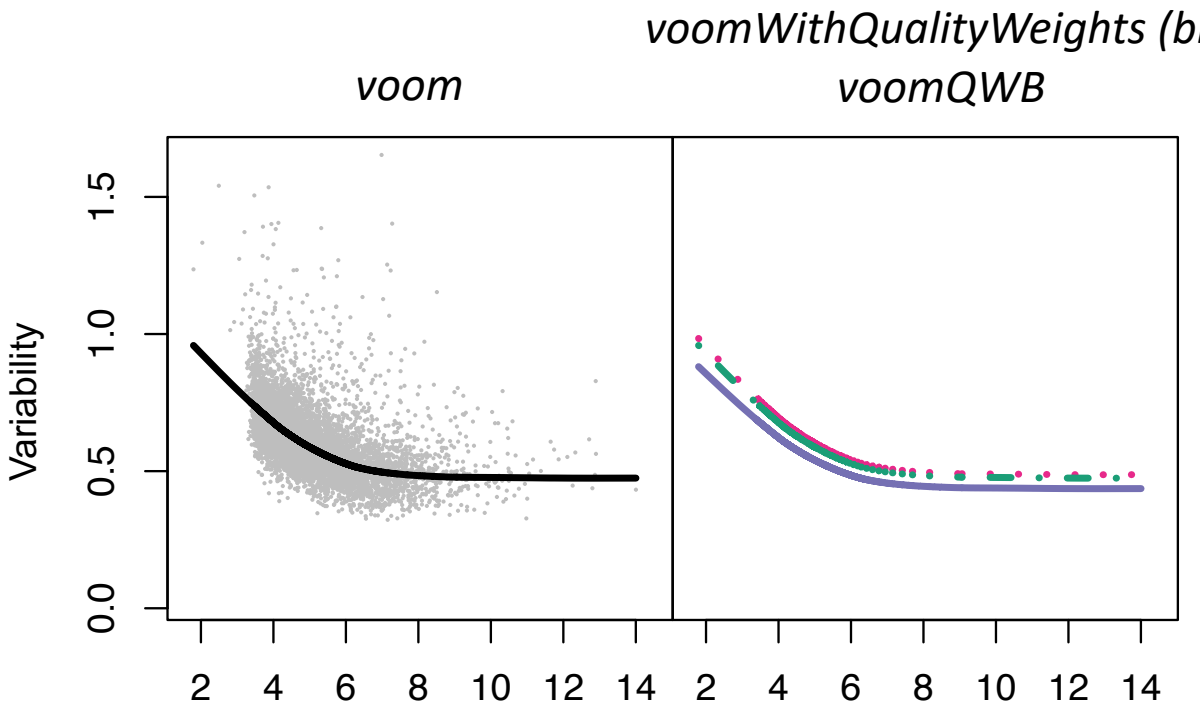

Law *et al.* (2014) Genome Biol. 15(2):R29.

# Strategy 1: Model group-specific mean-variance trends using *voomByGroup*



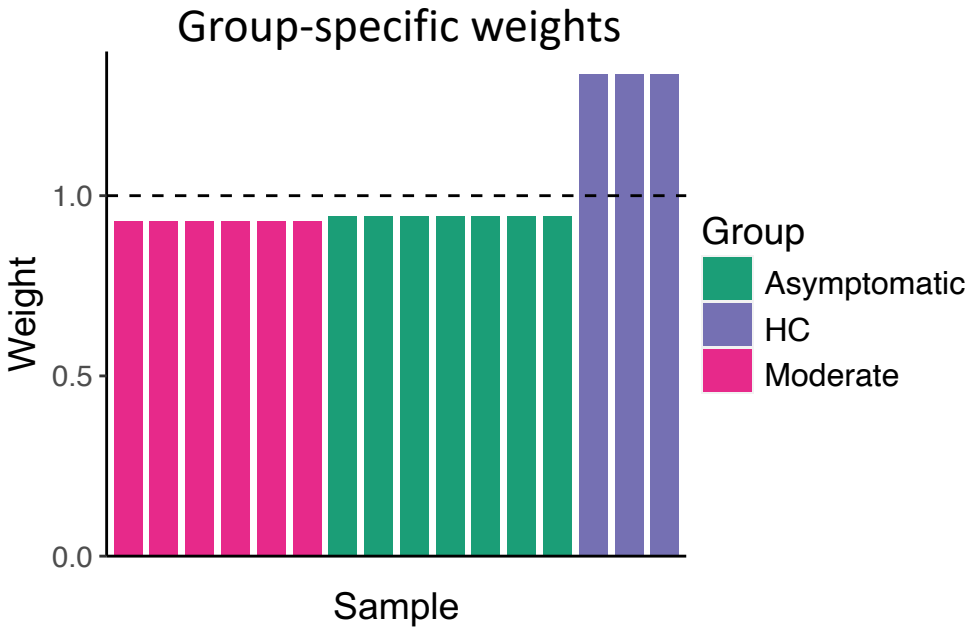

Law *et al.* (2014) Genome Biol. 15(2):R29.

# Strategy 2: Model group-specific variation using *voomWithQualityWeights*




$$w^*_{gj} = w_{gj} / \exp \hat{\gamma}_j$$

Estimated variance for group *j*

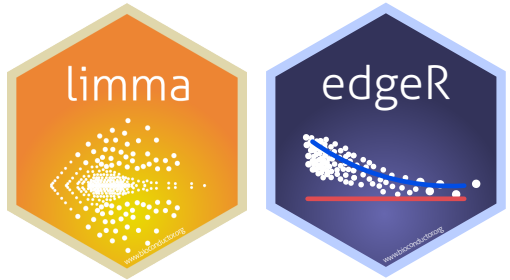




Ritchie *et al.* (2006) BMC Bioinformatics. 7:261
Lui *et al.* (2015) Nucleic Acids Research. 43(15):e97

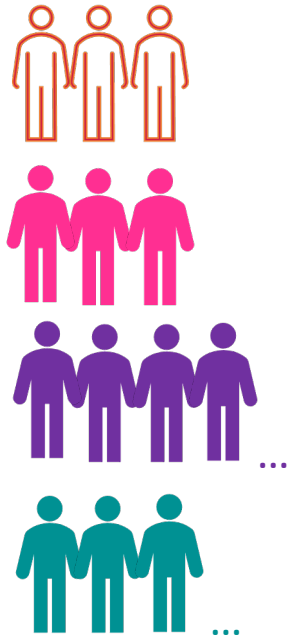
# Compare our new methods to existing pseudo-bulk DE tools

Simulated Data

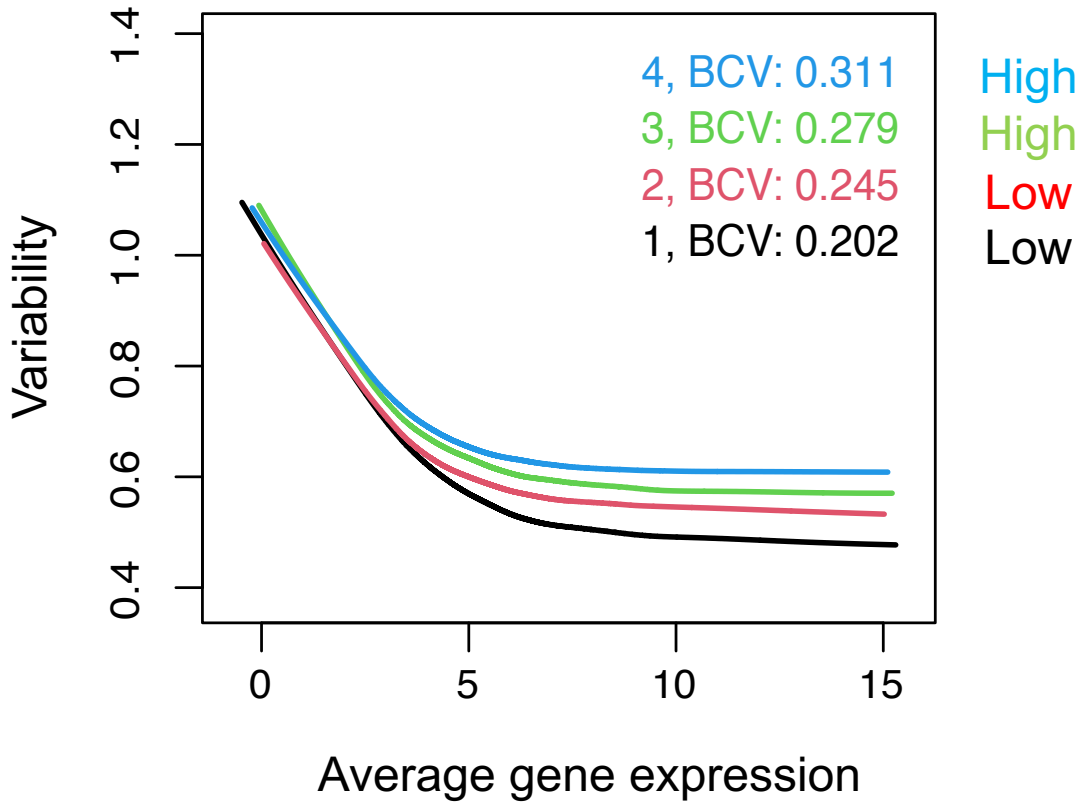



COVID-19 Data

- 3 Healthy controls (HC)
- 3 Severe samples
- 23 Moderate samples
- 11 Asymptomatic samples



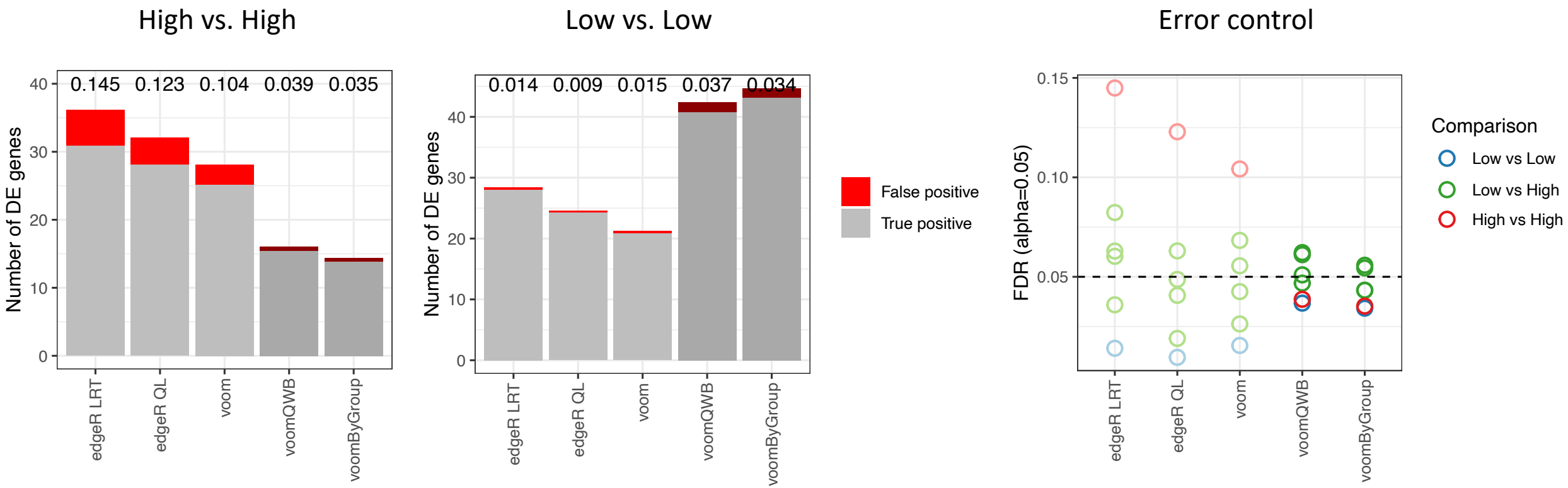
# Simulation 1: Mean-variance trends with distinct levels of *biological variation* in different groups




**Simulation strategy**

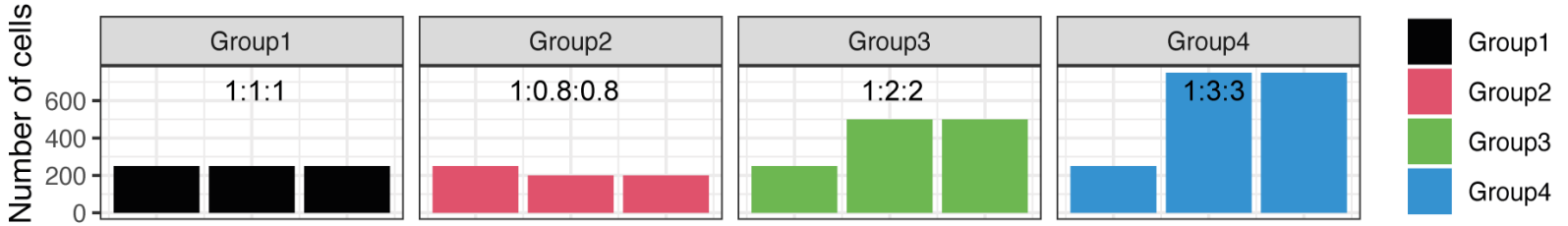
- 4 experimental groups, 2 distinct variability levels ('High' and 'Low')
- 3 replicate samples per group
- Expression levels (counts) simulated for 10,000 genes in 250 cells per sample
- **Ground-truth:** 50 genes up-regulated per group (= 100 true positive DE genes per pair-wise comparison), the rest are equally expressed between groups
- Repeat simulation 50 times and report averaged results

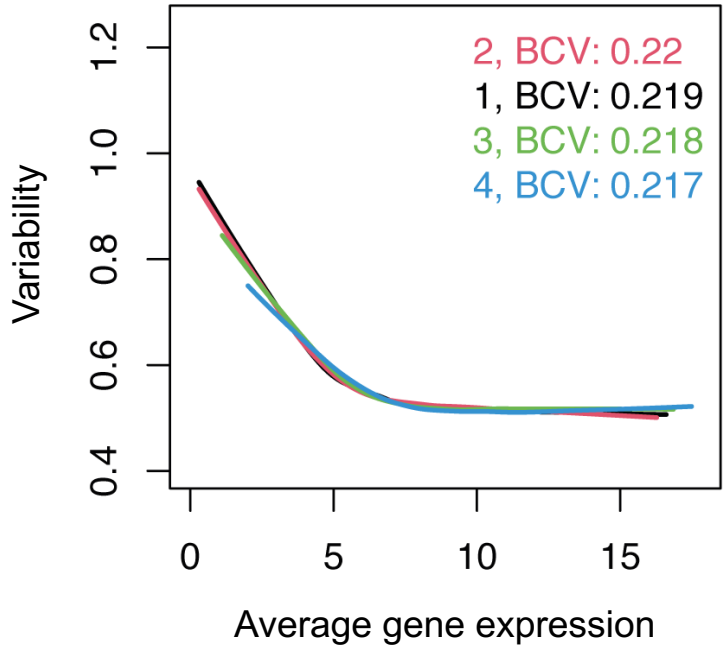
# Group variance methods provide a good balance between power and error control

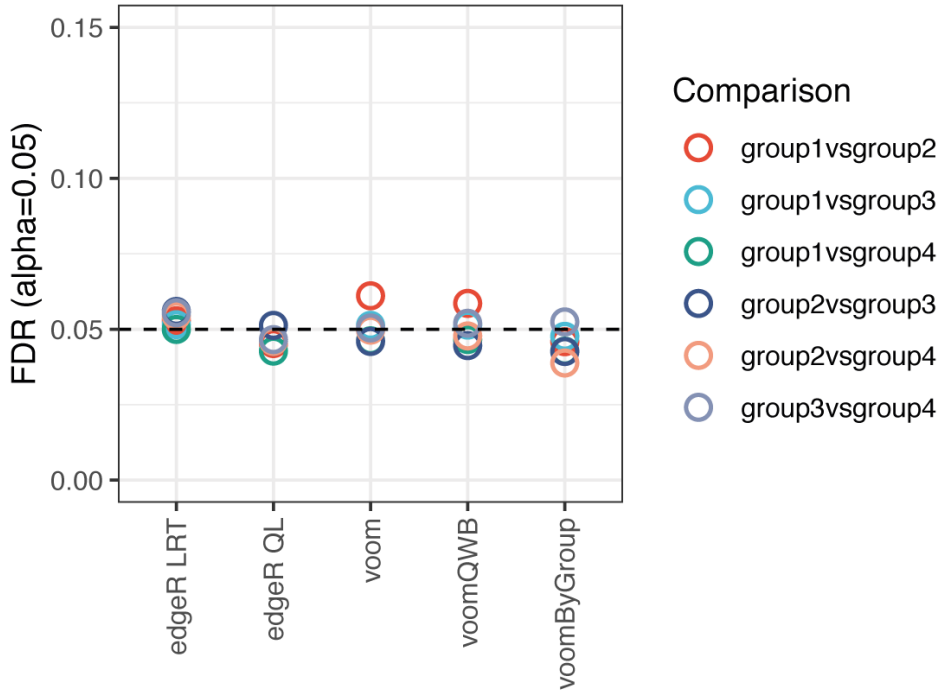



Threshold = 0.05

# Simulation 3: Mean-variance trends estimated with equal *biological variation*, but differences in *technical variation* between groups



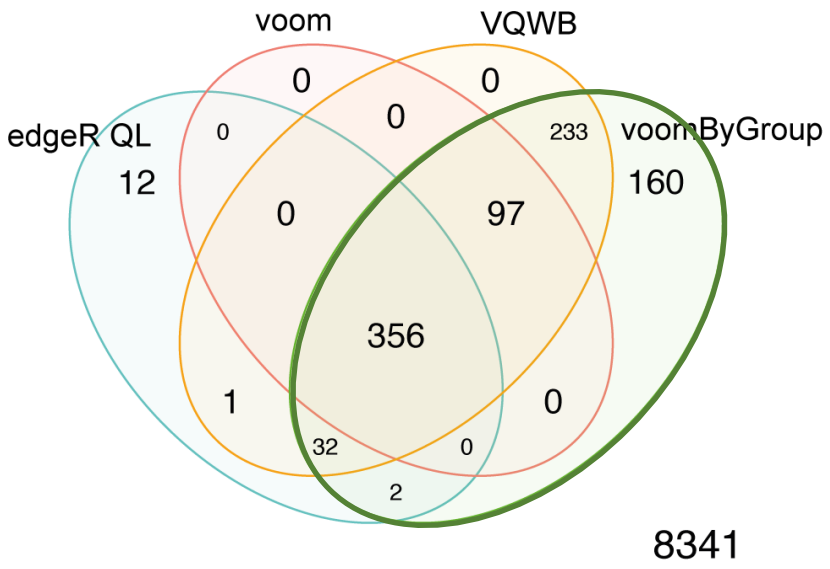



# COVID-19 Data: additional genes recovered by *voomByGroup* are biologically meaningful

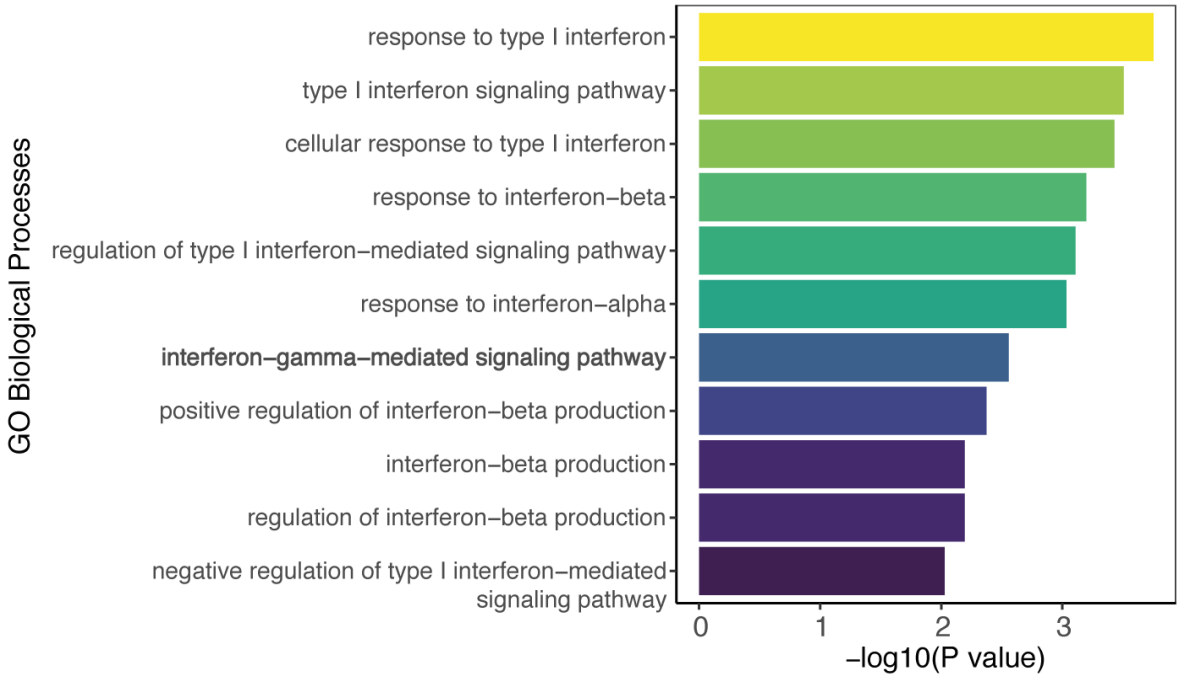
$CD56^{dim}$ $CD16^{+}$ NK cells

DE genes for Asymptomatic vs. HC



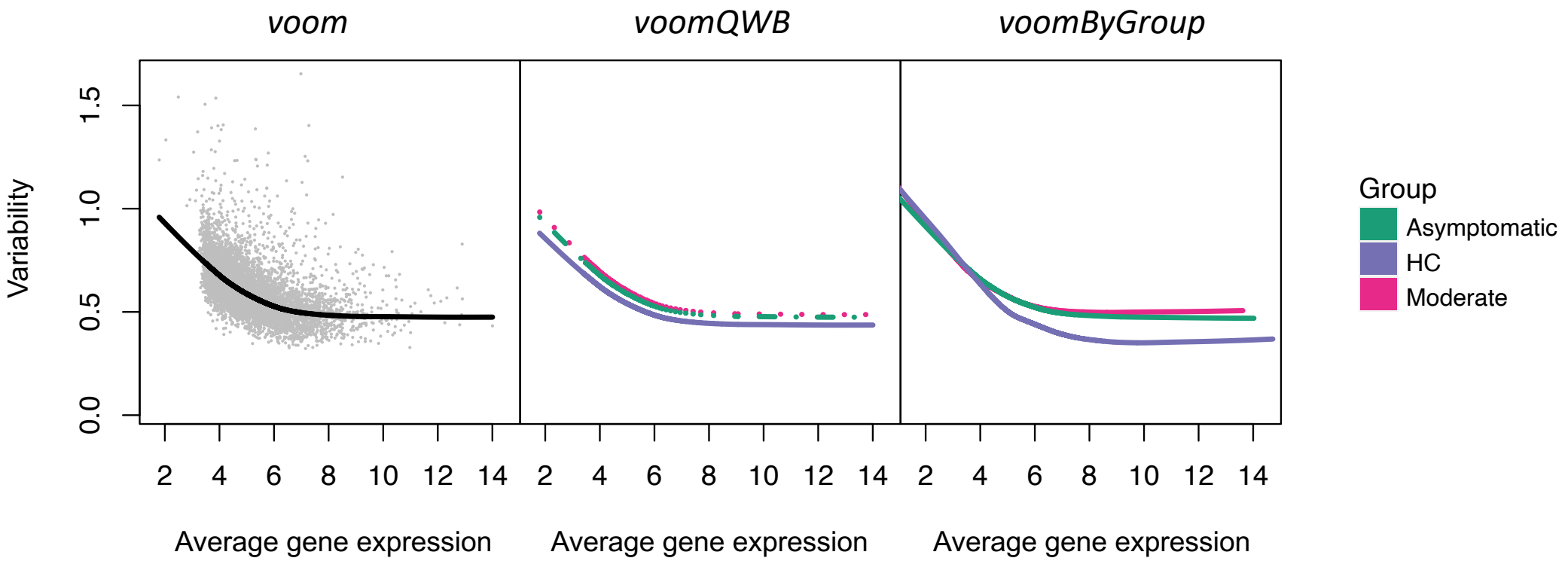

Zhao *et al.* (2021) Signal Transduct Target Ther. 6:342.

Gene Ontology Analysis



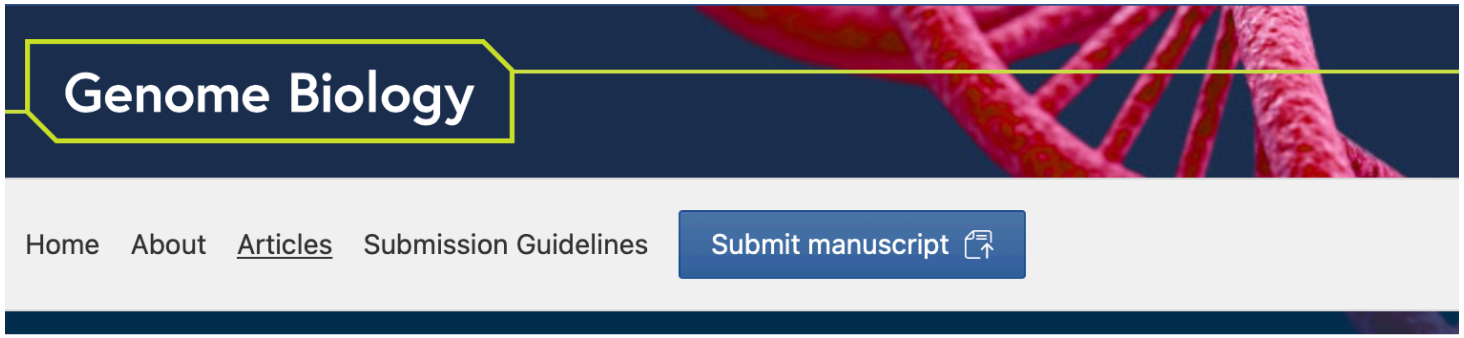

Stephenson *et al.* (2021) Nat Med. 27(5):904-16.

# Summary of new strategies for modelling group heteroscedasticity in pseudo-bulk scRNA-seq data




| Advantages: | Can deal with complicated experimental designs | Flexibility to capture distinct mean-variance trends |
|---|---|---|

# To learn more, refer to our publication 👇




Method | Open Access | Published: 05 May 2023

## Modeling group heteroscedasticity in single-cell RNA-seq pseudo-bulk data

Yue You, Xueyi Dong, Yong Kiat Wee, Mhairi J. Maxwell, Monther Alhamdoosh, Gordon K. Smyth, Peter F. Hickey, Matthew E. Ritchie & Charity W. Law

*Genome Biology* **24**, Article number: 107 (2023) | Cite this article

Paper: https://doi.org/10.1186/s13059-023-02949-2
Code: https://github.com/YOU-k/voomByGroup

# Acknowledgements


Ritchie Lab

**Yue You**
Xueyi Dong
Peter Hickey
Charity Law
Quentin Gouil
Shian Su
Kathleen Zeglinski
Changqing Wang
Mei Du
Oliver Voogd
Jakob Schuster
Lucinda Xiao
Callum Sargeant

WEHI brighter together

Gordon Smyth

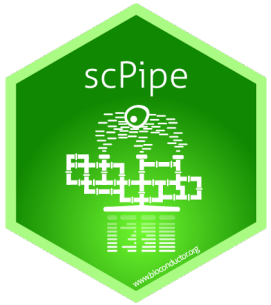
**CSL**
Yong Kiat Wee
Mhairi Maxwell
Monther Alhamdoosh



**Funders:** CZI, CSL, Australian Government National Health and Medical Research Council, The University of Melbourne

# Biology in the Matrix … Revisited

Tian *et al.* (2018) PLoS Comp Biol



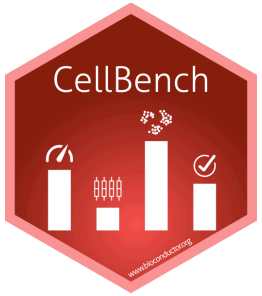

Lee *et al.* (2020) NAR Genom Bioinform


Kariyawasam *et al.* (2021) NAR Genom Bioinform


Tian *et al.* (2021) Genome Biol



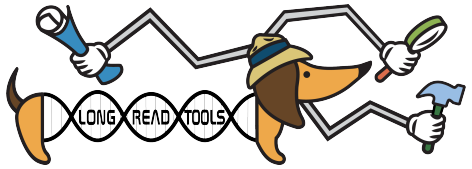
Tian *et al.* (2019) Nat Methods
Su *et al.* (2020) Bioinformatics
You *et al.* (2021) Genome Biol





Gigante, Gouil *et al.* (2019) NAR
Su *et al.* (2021) PLoS Comp Biol

You *et al.* (2023) Genome Biol


Amarasinghe *et al.* Genome Biol 2020
Amarasinghe *et al.* GigaScience 2021