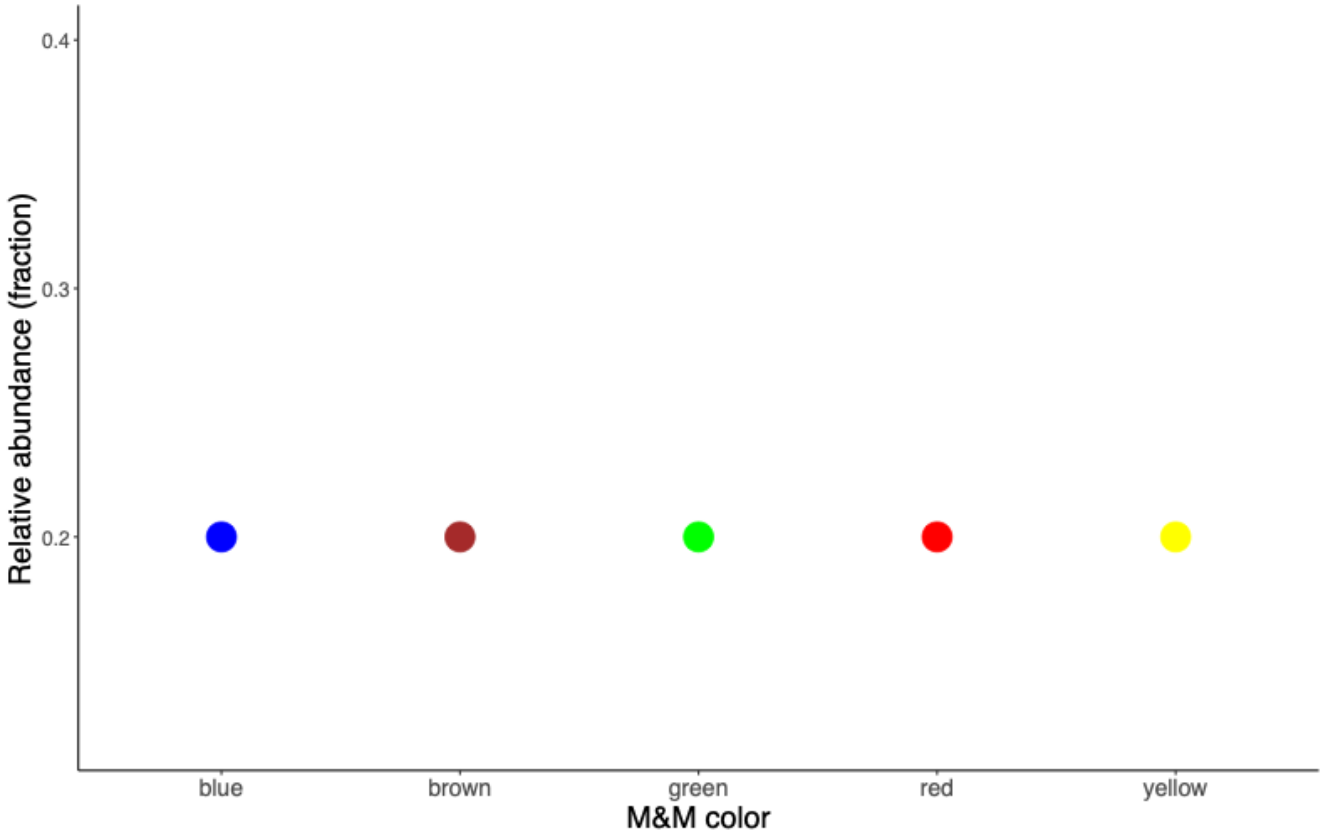
# Assessing changes in cell composition in single-cell data

Koen Van den Berge
Joint work with Alemu Takele Assefa & Bie Verbist
Johnson & Johnson Innovative Medicine
September 27, 2024
Wiesbaden, Germany

Johnson&Johnson
Innovative Medicine

# Relative abundance of M&M's

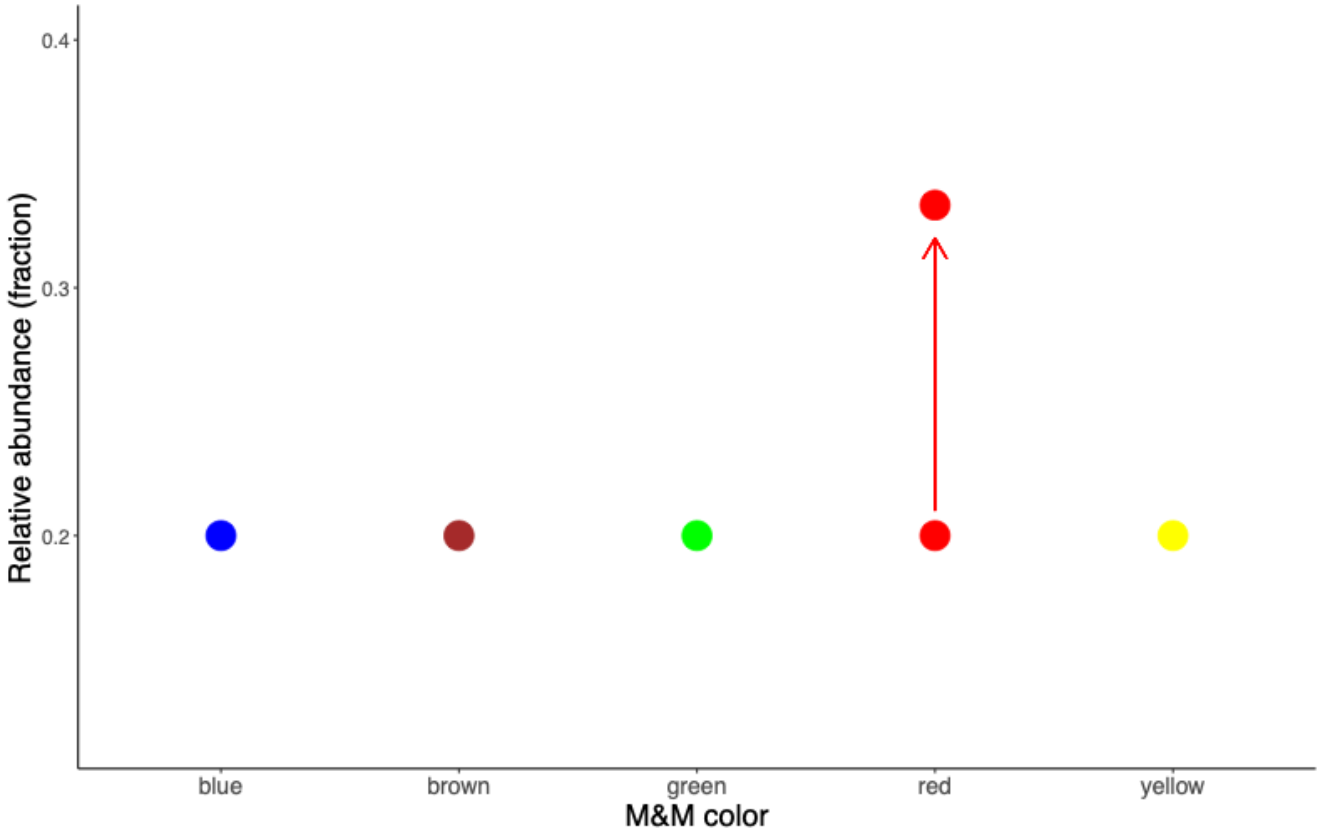



HEALTHY condition:

One bag of M&M's, with 5 colors.

Each color equally abundant at 20%.

# Relative abundance of M&M's




HEALTHY condition:

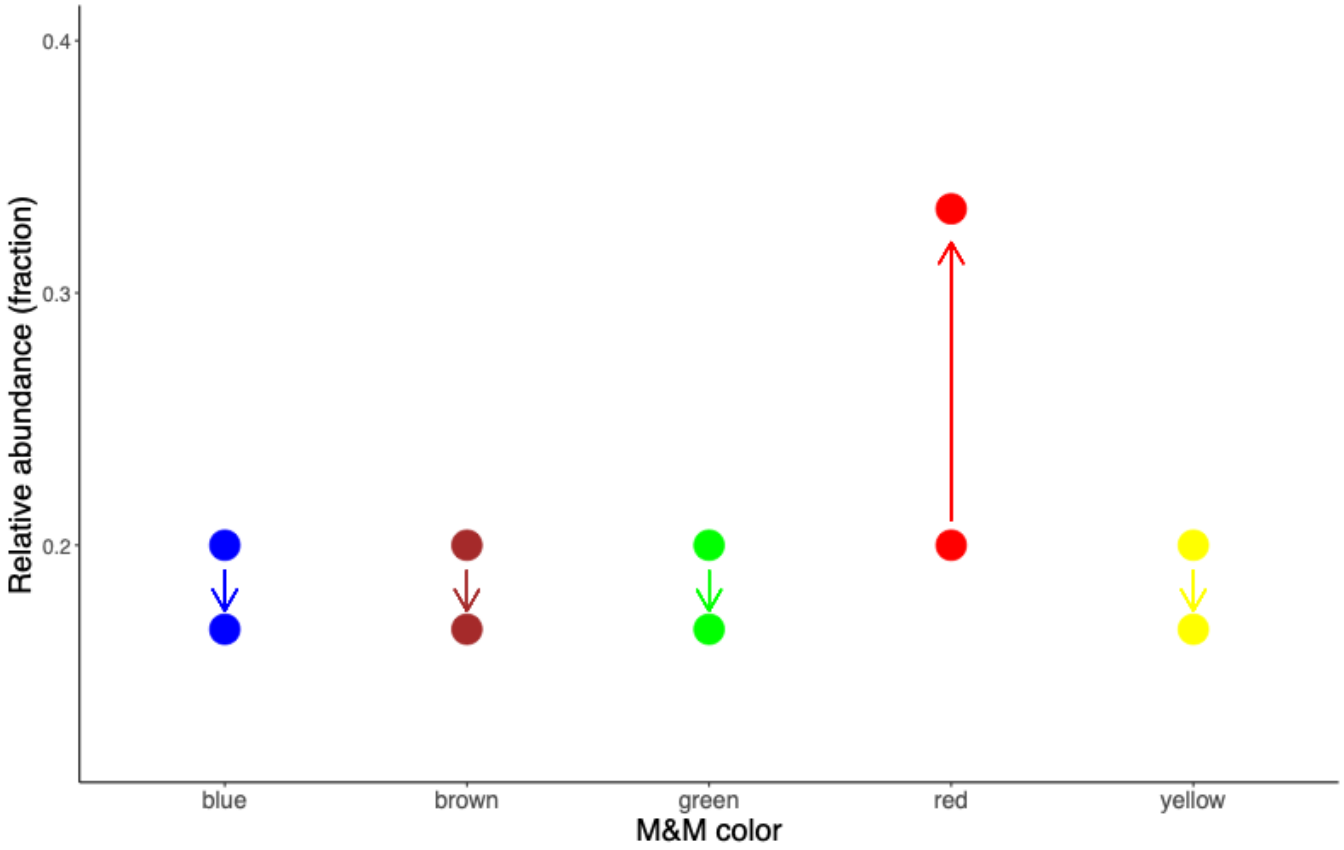
One bag of M&M's, with 5 colors.

Each color equally abundant at 20%.

DISEASED condition:

Red M&M's increase in abundance (20% to 33%).

# Relative abundance of M&M's




HEALTHY condition:

One bag of M&M's, with 5 colors.

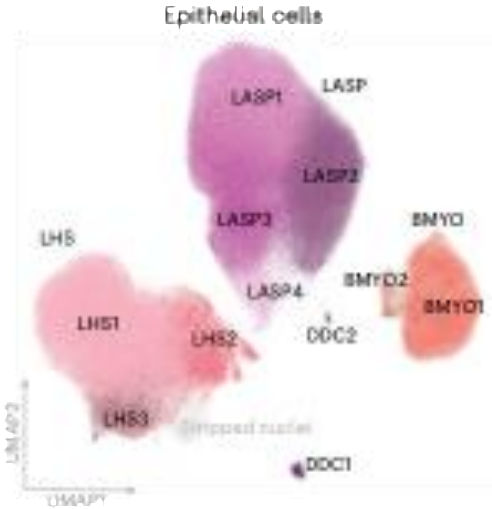
Each color equally abundant at 20%.

DISEASED condition:

Red M&M's increase in abundance (20% to 33%).

We're constrained to 100%, therefore other colors must decrease in relative abundance.

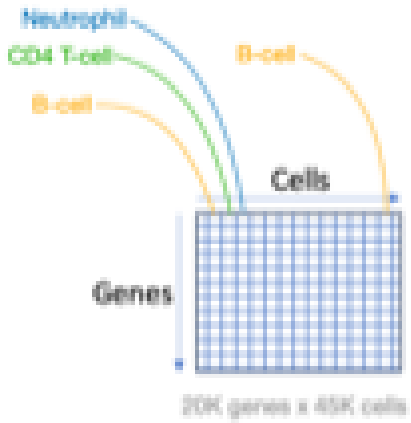
# A composition of single cell types



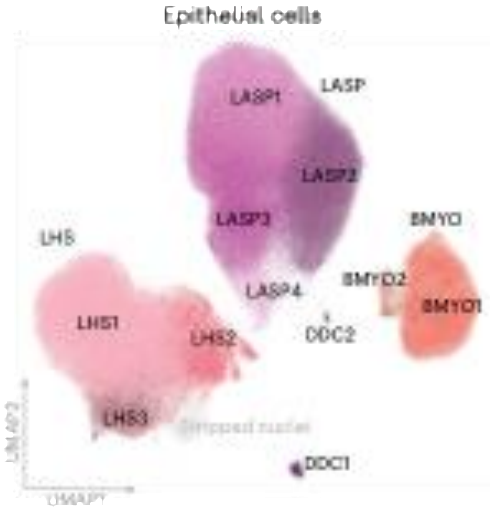

In single-cell RNA-seq data, the M&M colors are cell types.
The same principle applies: only relative information available (compositional data).

Starting from the single-cell gene expression count matrix:

1. Each single cell gets assigned a cell type label (M&M color), based on its gene expression.



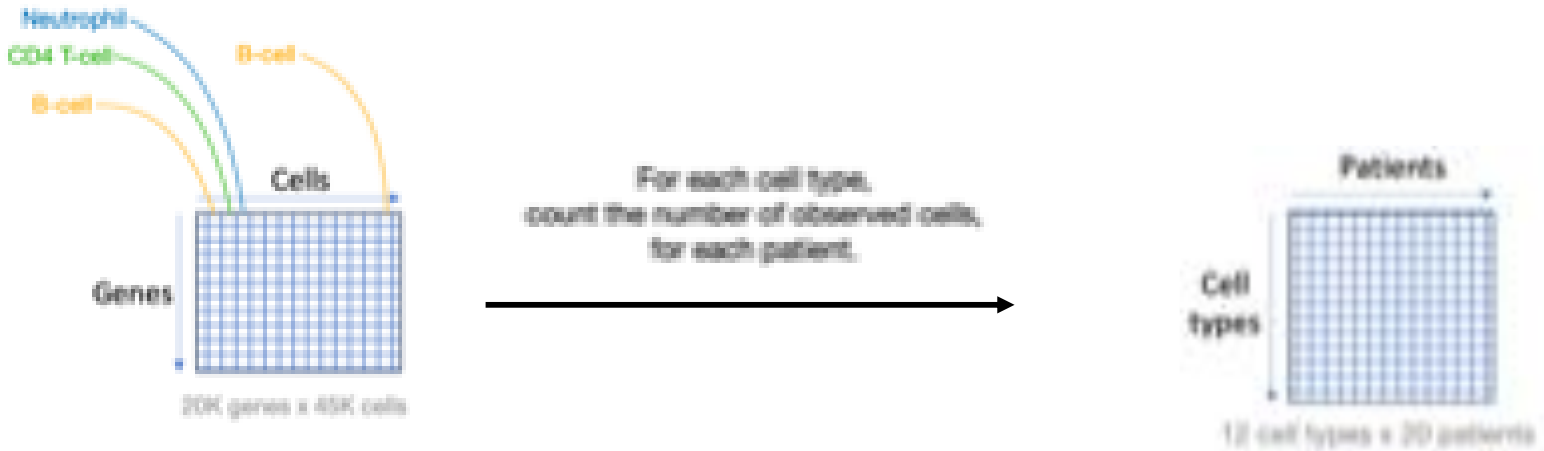
# A composition of single cell types




In single-cell RNA-seq data, the M&M colors are cell types.
The same principle applies: only relative information available (compositional data).

Starting from the single-cell gene expression count matrix:

1. Each single cell gets assigned a cell type label (M&M color), based on its gene expression.
2. One sums the number of cells per patient sample to derive the cell abundance count matrix.



# Compositional data require custom statistical models

In compositional data, we still want to infer upon the latent absolute abundance (actual number of M&Ms in the bag). Two main avenues are possible:

1. Compositional statistical model (e.g., Dirichlet-Multinomial).
2. Compositional transformations (e.g., centered or additive log-ratio).

Let $Y_{ip}$ denote the cell type counts for cell population $p$ in sample $i$. The centered-log-ratio (CLR) transformation is

$$Z_{ip} = log\left[\frac{Y_{ip}}{\widetilde{Y_i}}\right] = log\left[\frac{Y_{ip}}{(\prod_i Y_{ip})^{1/P}}\right],$$

With $\widetilde{Y_i}$ the geometric mean across cell types for sample $i$, and $P$ the total number of cell populations.

# Compositional data require custom statistical models

In compositional data, we still want to infer upon the latent absolute abundance (actual number of M&Ms in the bag). Two main avenues are possible:

1. Compositional statistical model (e.g., Dirichlet-Multinomial).
2. Compositional transformations (e.g., centered or additive log-ratio).

In this talk, we will benchmark the most popular methods out there for assessing differential cell type composition.

Through identifying shortcomings of existing methods,
we develop new methodology by leveraging building blocks from other methods in the literature.

# An overview of existing methods

### Compositional transformation

- `CLR_lm`
  Linear model post CLR transformation.
- `LinDA`
  Linear model post CLR transformation. Bias correction on effect size.

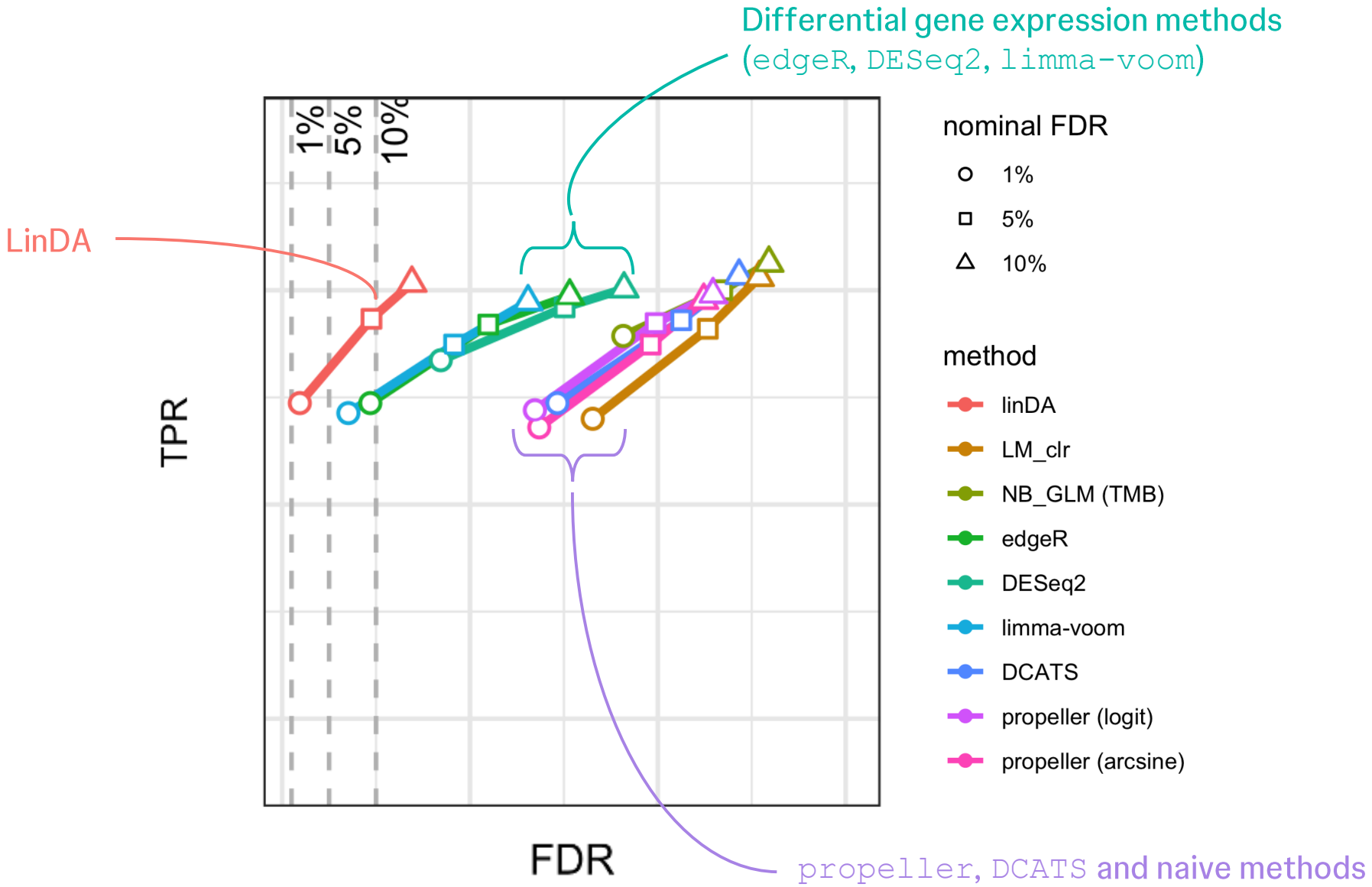
### Non-compositional transformation

- `Limma-voom`
  Log-counts-per-million transformation, weighted linear model and empirical Bayes shrinkage of residual variance.
- `Propeller (logit and arcsin)`
  Linear model post logit or square root arcsine transformation.

### Count model

- `edgeR`
  Negative binomial model, normalization of total count, empirical Bayes shrinkage of dispersion parameter.
- `DESeq2`
  Negative binomial model, normalization of total count, empirical Bayes shrinkage of dispersion parameter.
- `NB GLM`
  Negative binomial model using total count as offset.
- `DCATS`
  Beta-Binomial model with shared dispersion parameter.

# Performances of existing methods




Alemu Takele Assefa

J&J Innovative Medicine

# Best performing method is still suboptimal

**Counts are still heteroscedastic post transformation**

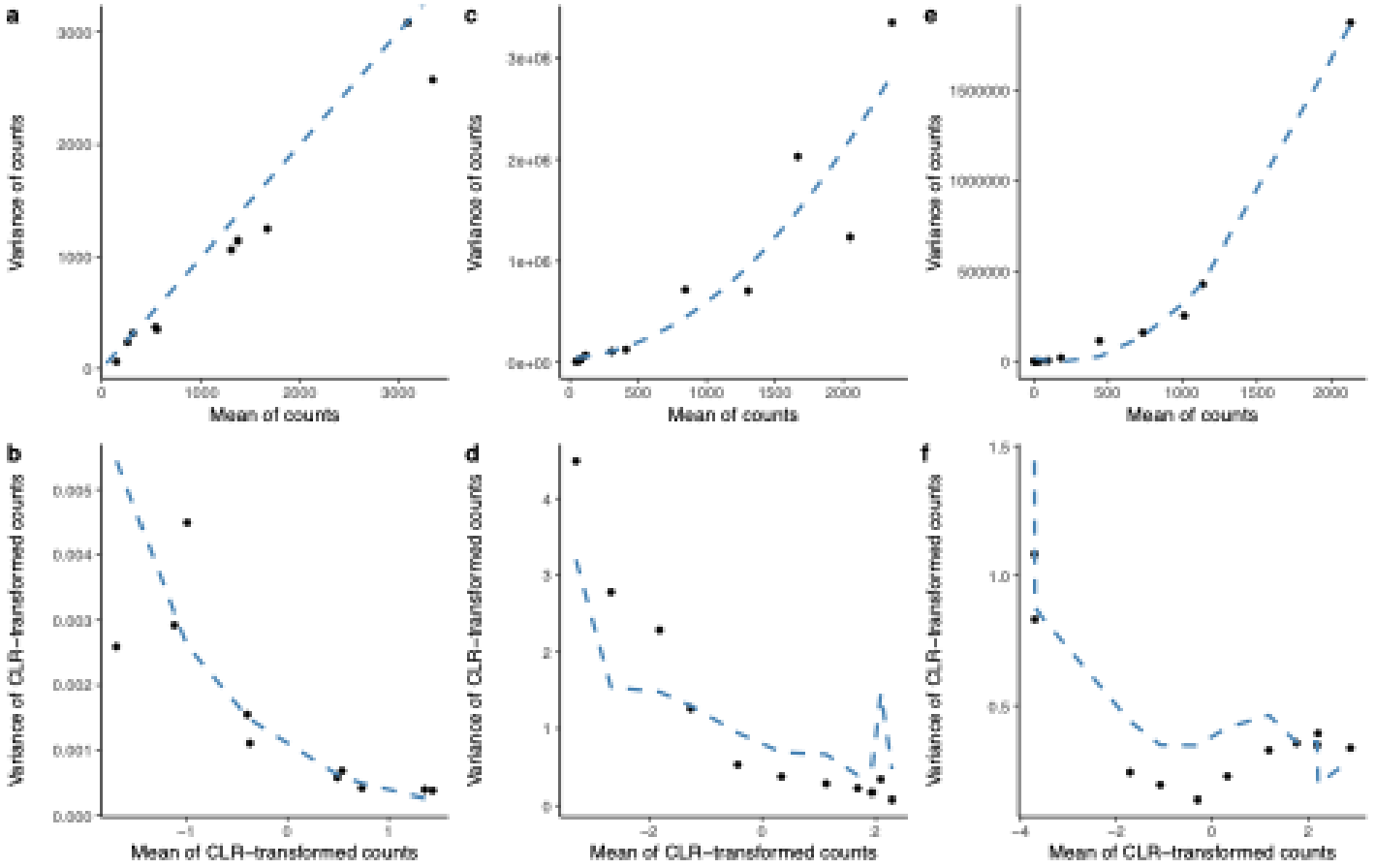
Let $Y_{ip}$ denote the cell type counts for cell population $p$ in sample $i$.

$$\text{CLR: } Z_{ip} = \log \frac{Y_{ip}}{\tilde{Y}_i} \text{ with } \tilde{Y}_i = \left( \prod_{p=1}^{P} Y_{ip} \right)^{1/P}.$$

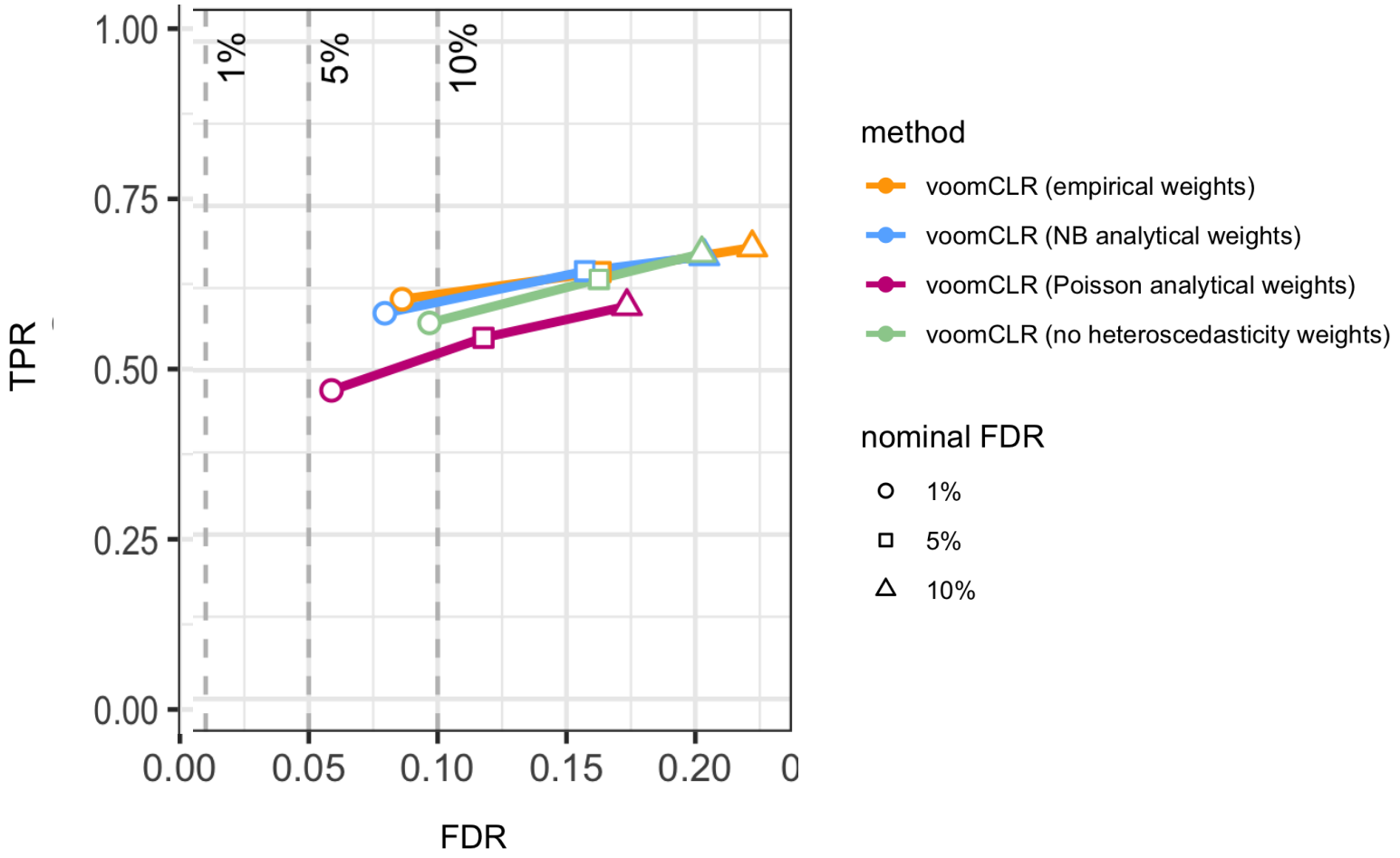
$$\text{If } Y_{ip} \sim Poi(\lambda_{ip}), \quad Var(Z_{ip}) = \left(\tfrac{P-1}{P}\right)^2 \tfrac{1}{\lambda_{ip}}.$$

$$\text{If } Y_{ip} \sim NB(\mu_{ip}, \phi_p), \quad Var(Z_{ip}) = \left(\tfrac{P-1}{P}\right)^2 \left(\tfrac{1}{\mu_{ip}} + \phi_p\right).$$

Mean-variance relationship of counts (top row), and CLR-transformed counts (bottom row).



# Impact of accounting for heteroscedasticitv

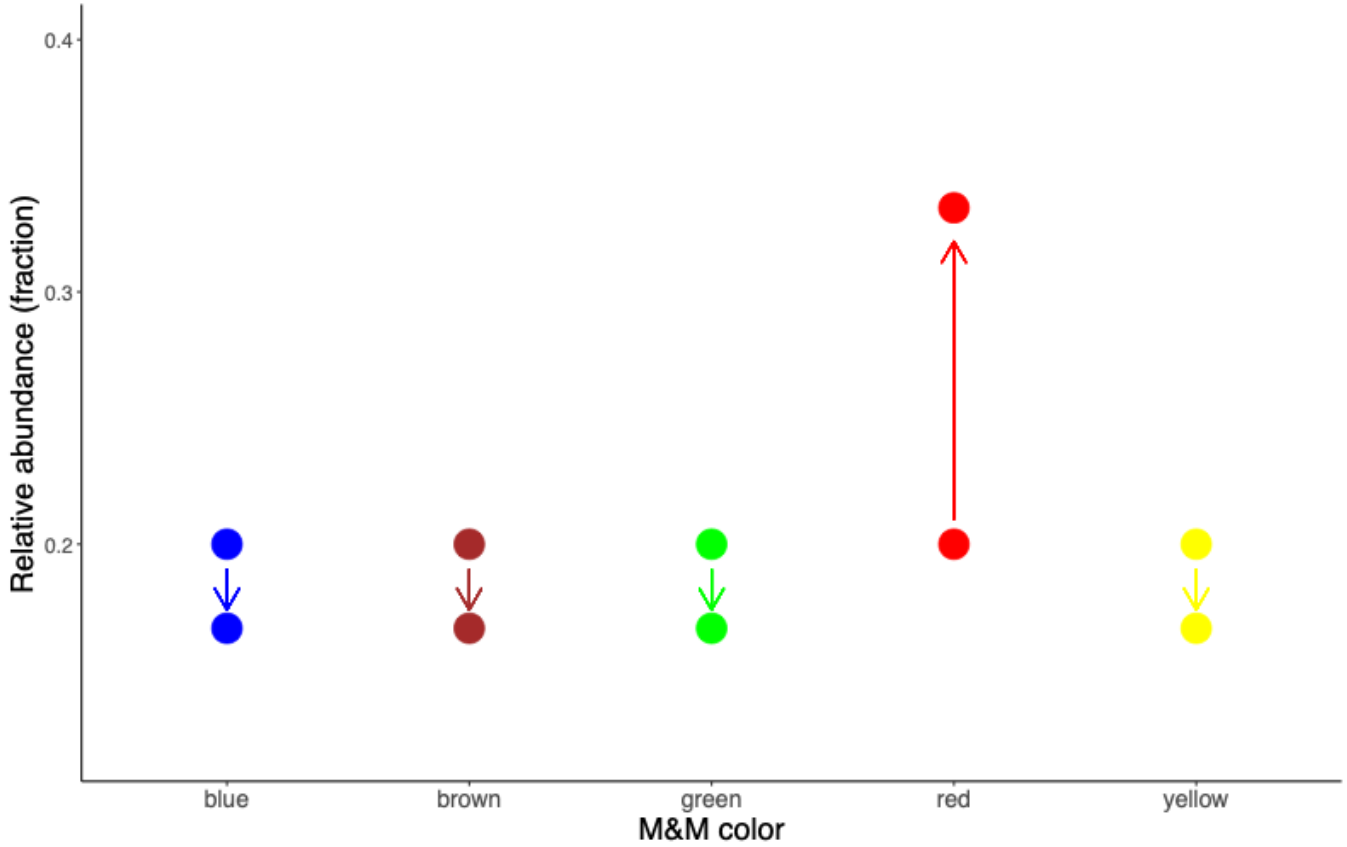



Alemu Takele Assefa

J&J Innovative Medicine

# Remember the M&M example; effect sizes are biased

**Uncertainty in bias correction is not propagated in statistical inference**




Modeling the fractions directly for each cell type independently would lead us to find all colors / cell types are changing.

We should only find the red color.

Effect sizes are biased due to compositionality.

# Best performing method is still suboptimal

**Uncertainty in bias correction is not propagated in statistical inference**

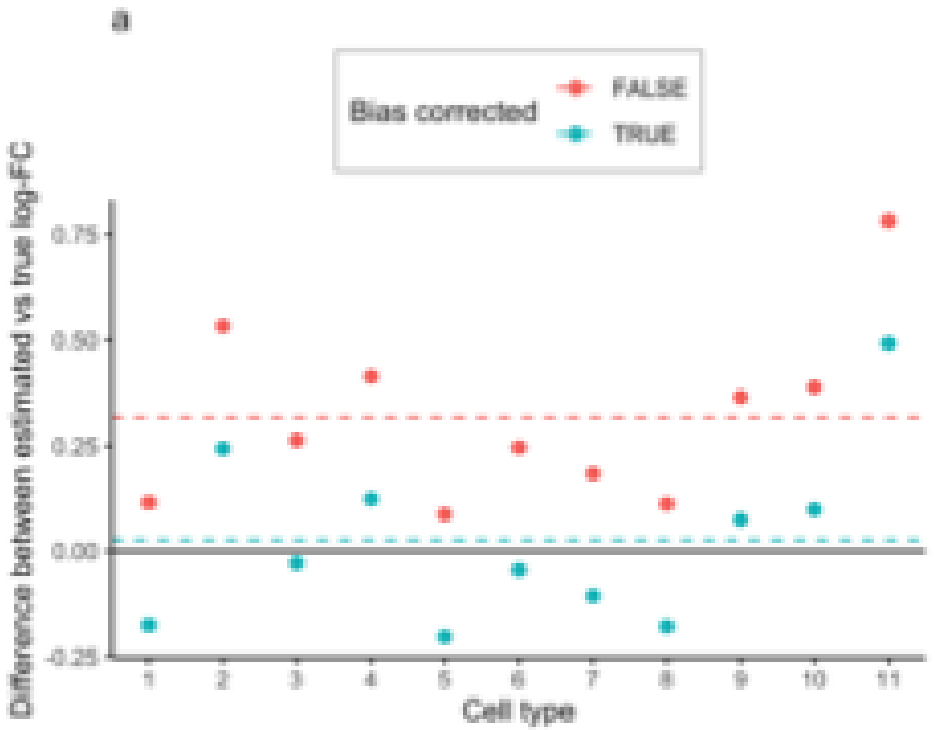
Method | Open access | Published: 14 April 2022

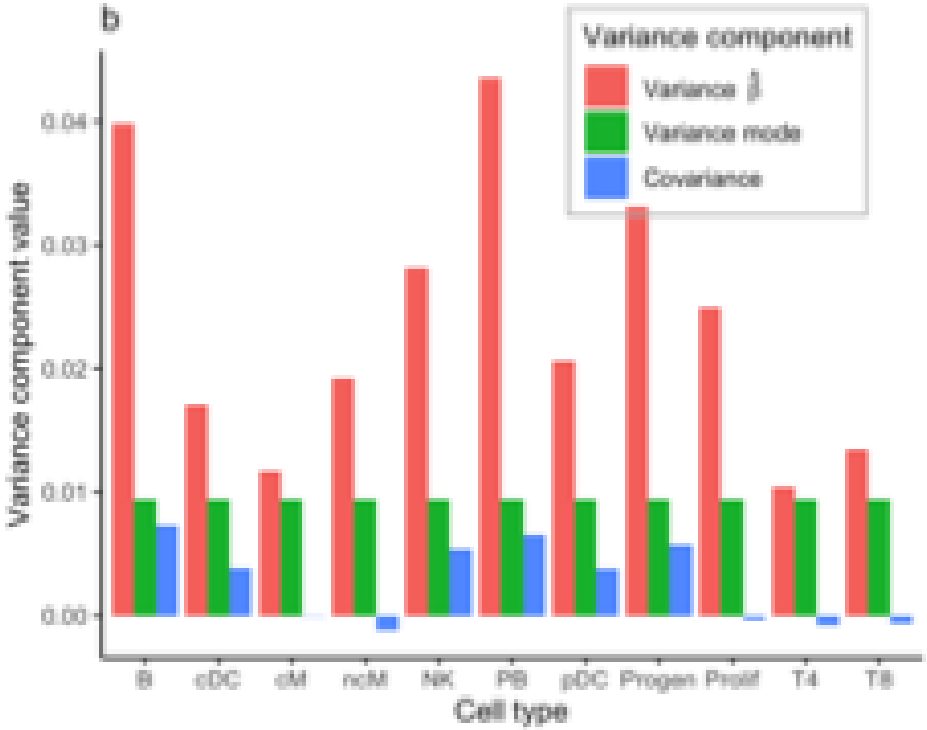
**LinDA: linear models for differential abundance analysis of microbiome compositional data**

Huijuan Zhou, Kejun He, Jun Chen & Xianyang Zhang

*Genome Biology* **23**, Article number: 95 (2022) | Cite this article

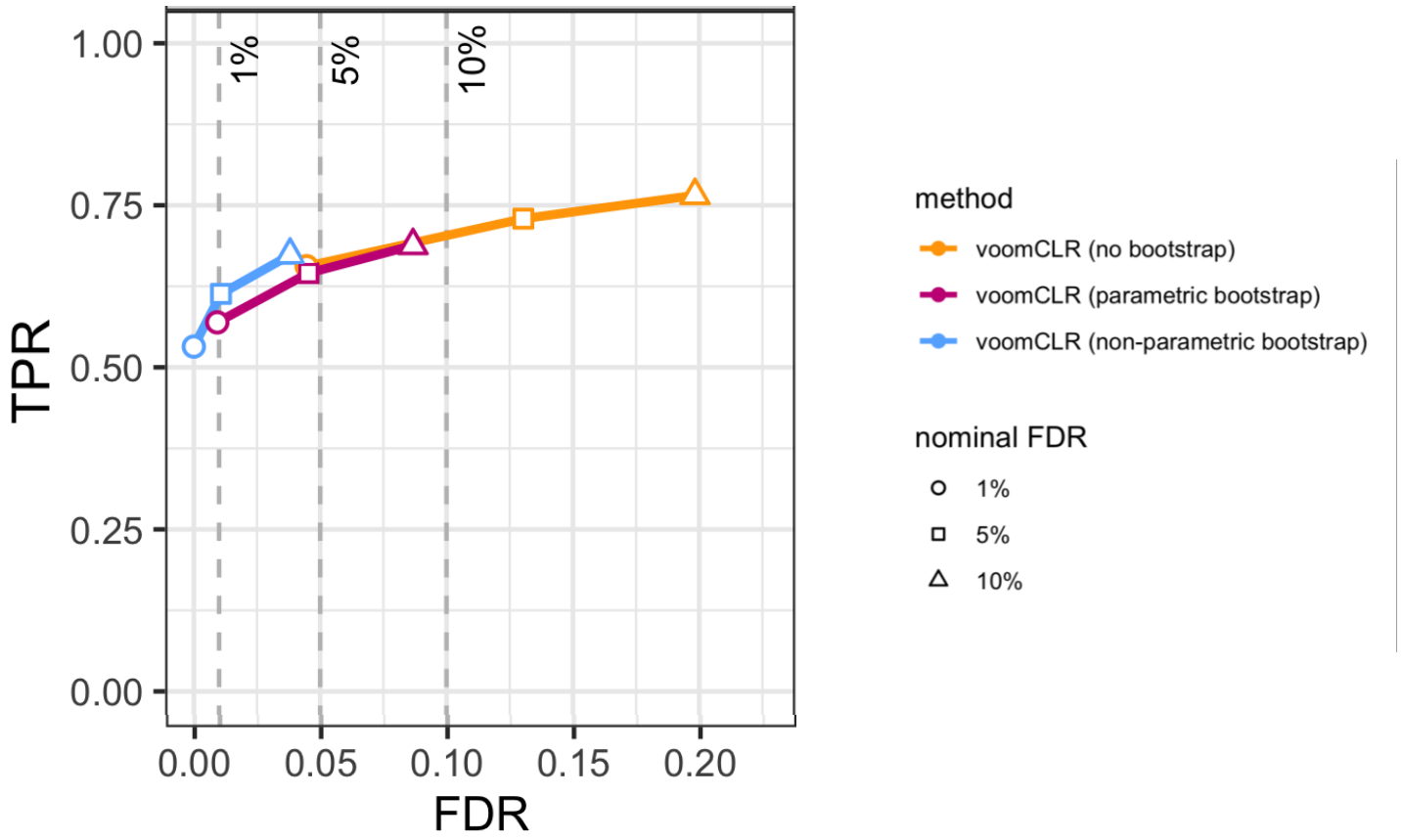
$$Var(\tilde{\beta}_{jp}) = Var(\hat{\beta}_{jp}) + Var(\check{\beta}_j) - 2Cov(\hat{\beta}_{jp}, \check{\beta}_j).$$



# Impact of accounting for bias correction uncertainty

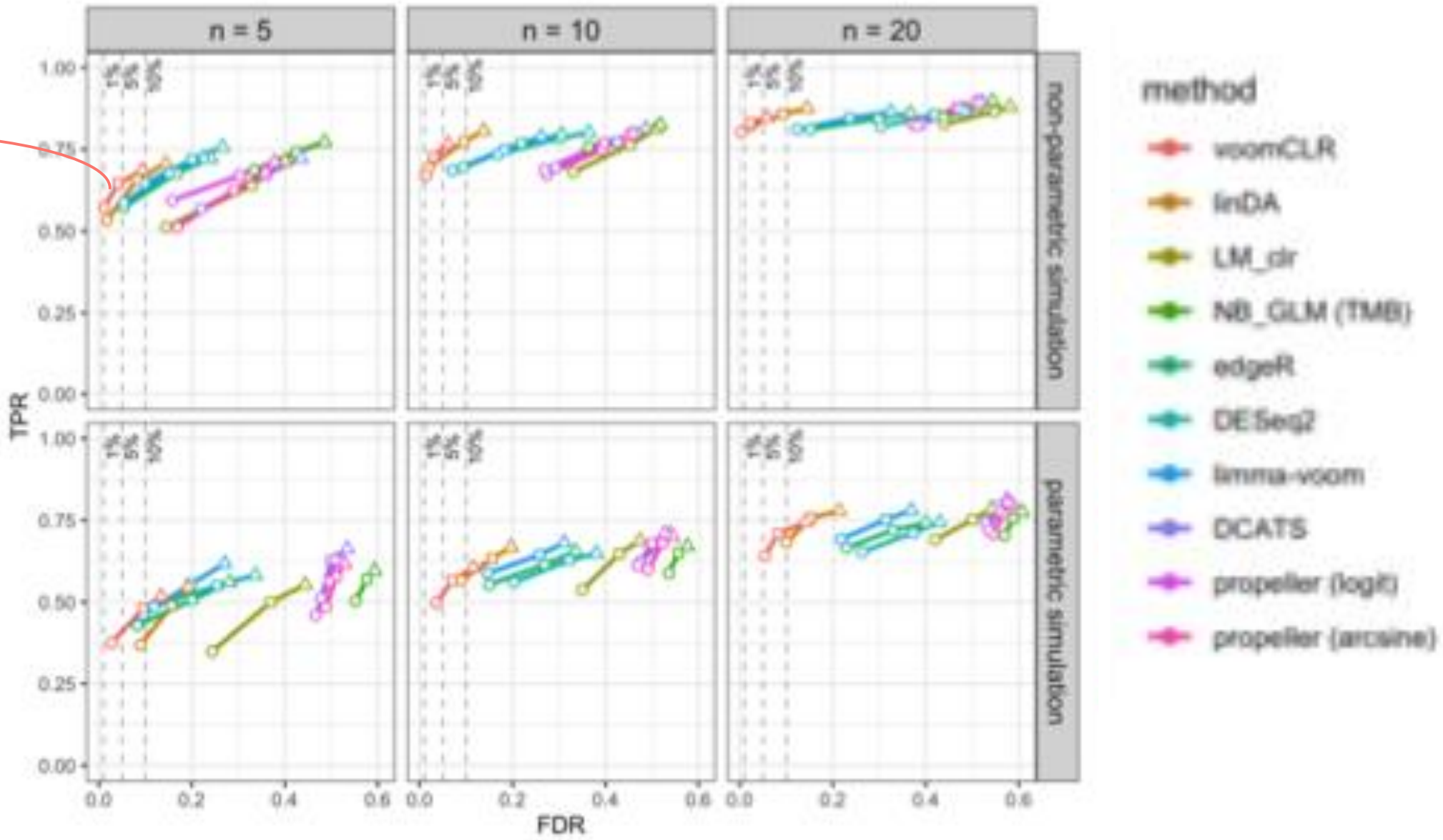
**Bias correction uncertaintypropagation contributes to better false positive control**




Alemu Takele Assefa

J&J Innovative Medicine

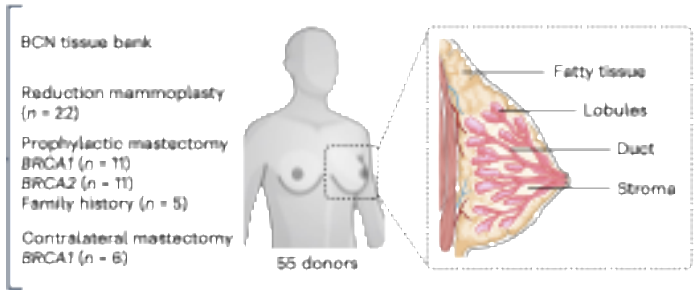
# voomCLR is at least on par and often outperforms other methods

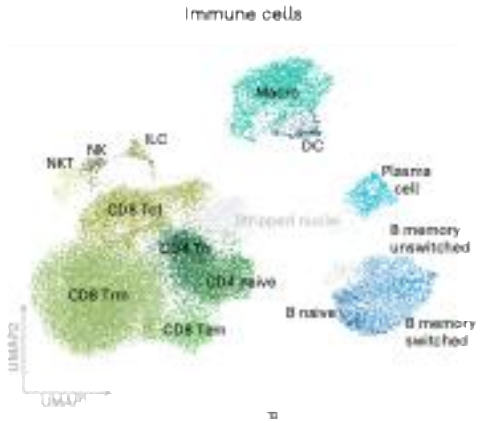


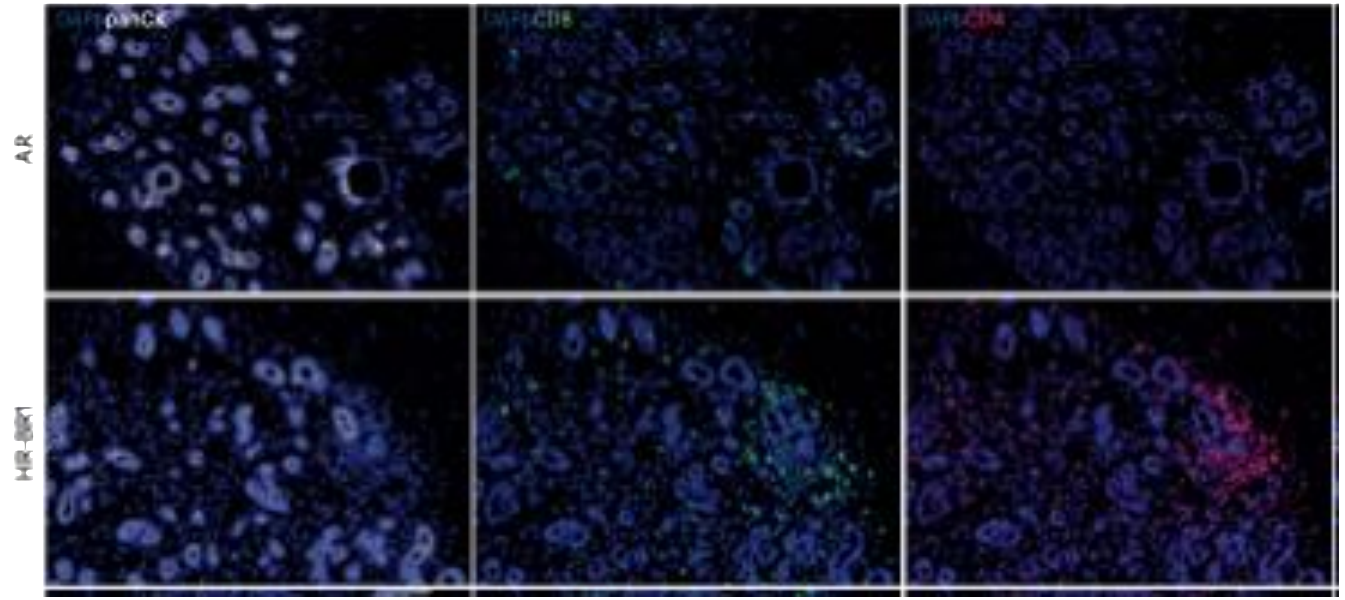

Alemu Takele Assefa

J&J Innovative Medi

# Case study on breast cell atlas



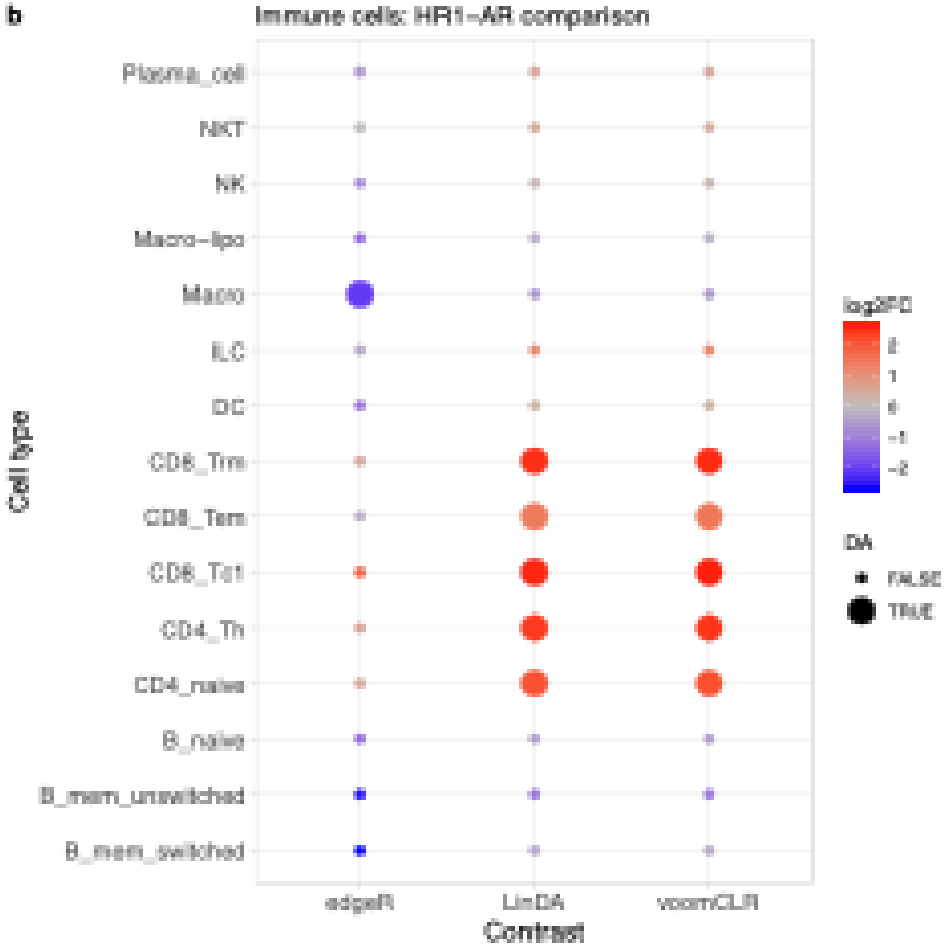




nature genetics

Article

https://doi.org/10.1038/s41588-024-01688-9

**A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast**



# Thank you

Alemu Takele Assefa

Bie Verbist

If you have more questions, please contact:
kvande14@its.jnj.com

Johnson&Johnson
Innovative Medicine