A statistical approach for modelling differential distributions in single-cell transcriptomic data

Malindrie Dharmaratne[1], Ameya Kulkarni[2], Atefeh Taherian Fard[1], Ernst Wolvetang[1], Nir Barzilai[2], Jessica C Mar[1]

m.dharmaratne@uq.net.au
@MalDharm2

(1) Australian institute for Bioengineering and Nanotechnology, University of Queensland
(2) Department of Endocrinology, Albert Einstein College of Medicine, Bronx, NY

Introduction

- Single-cell RNA sequencing (scRNA-seq) allows the sequencing of the whole transcriptome at the resolution of a single cell.
- Single-cell data can be driven by outliers, over-dispersion and dropouts, resulting in multiple expression modes.
- Most existing tools focus on the effects of change in mean expression, assuming all the genes in the transcriptome follow a single distribution.
- We propose a statistical framework for identifying distributional shapes of transcriptomic data.
- The UMI counts for each gene are modelled using the error distributions Poisson (P), Negative Binomial (NB), Zero Inflated Poisson (ZIP) and Zero inflated negative binomial (ZINB).

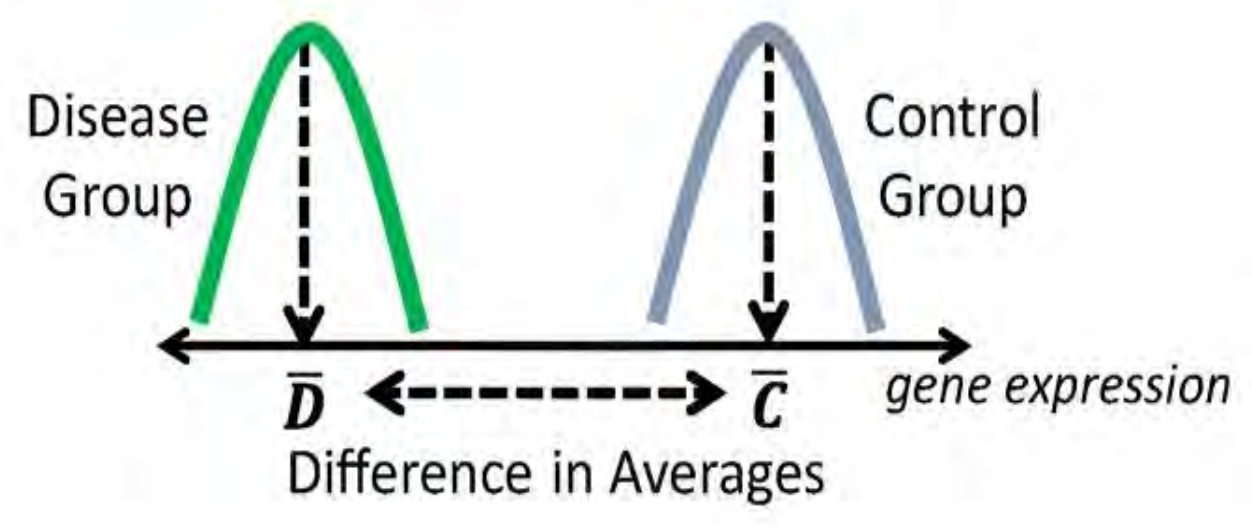


Figure 1: Regulatory information can be derived by looking beyond change in average effects in gene expression values [1].

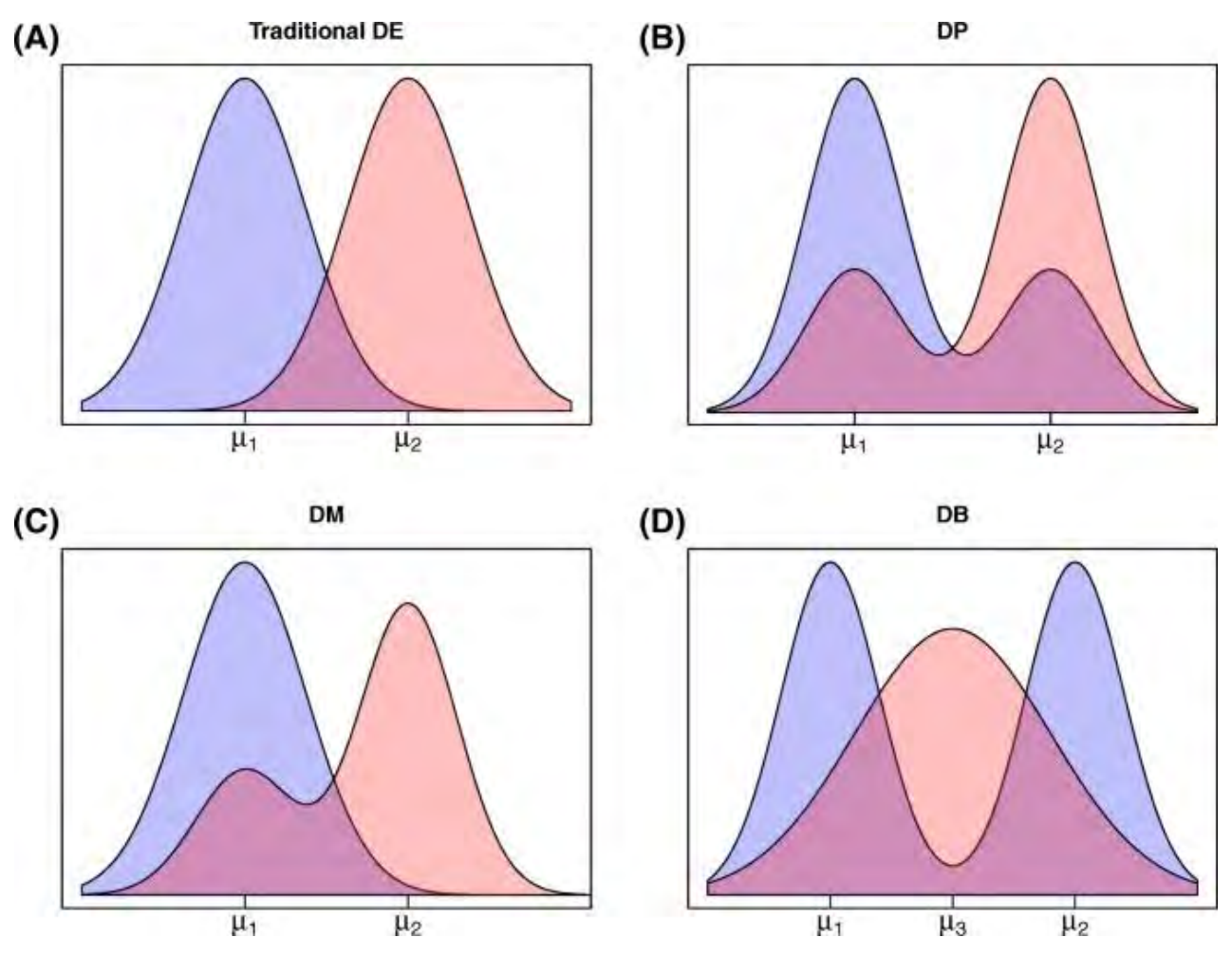
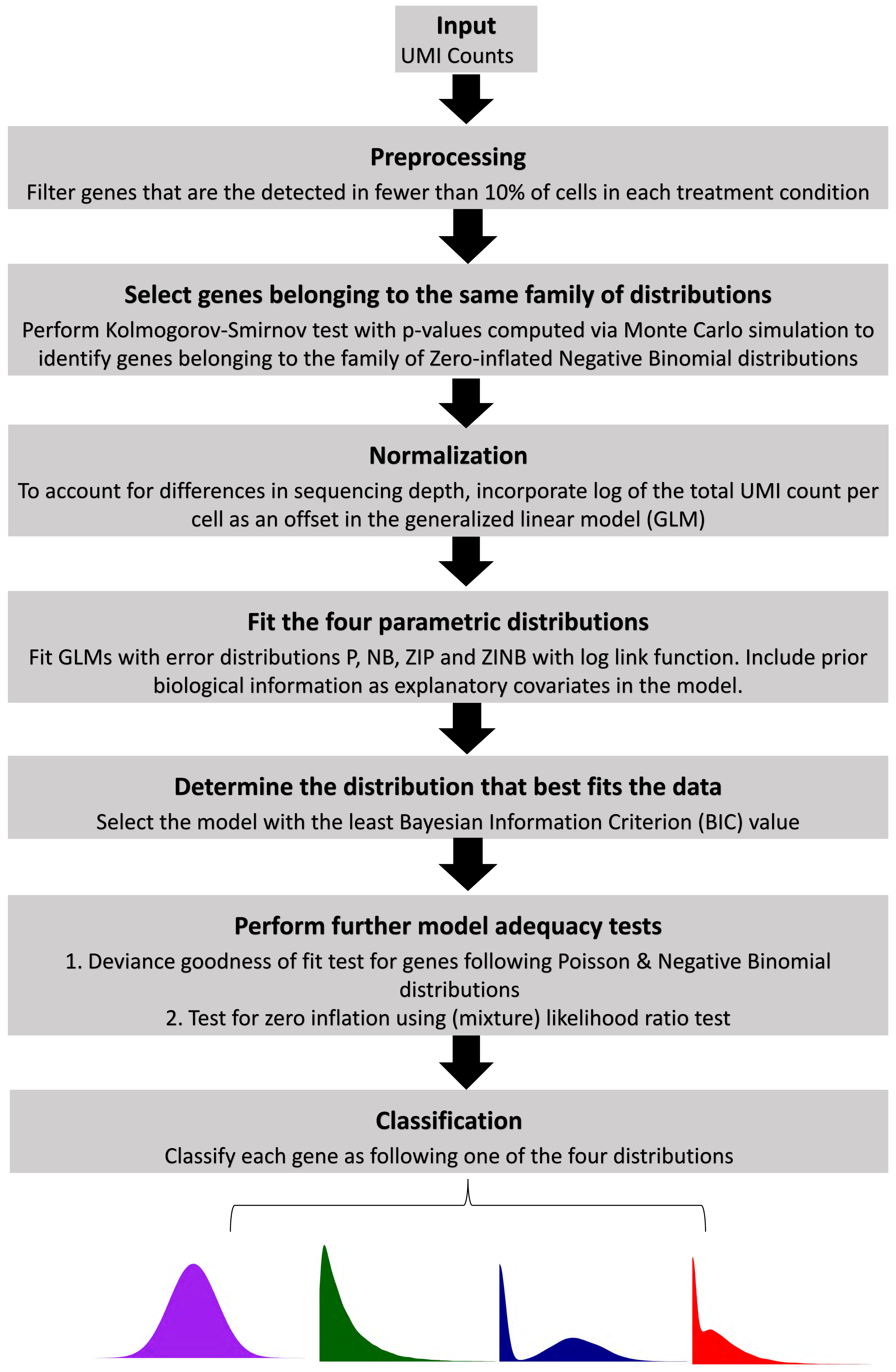


Figure 2: Possible differential distribution patterns as proposed by Korthauer et. al. [2]. This approach is limited due to its inability to adjust for covariates and only pair-wise comparisons being possible.

Statistical framework



Input
UMI Counts

Preprocessing
Filter genes that are the detected in fewer than 10% of cells in each treatment condition

Select genes belonging to the same family of distributions
Perform Kolmogorov-Smirnov test with p-values computed via Monte Carlo simulation to identify genes belonging to the family of Zero-inflated Negative Binomial distributions

Normalization
To account for differences in sequencing depth, incorporate log of the total UMI count per cell as an offset in the generalized linear model (GLM)

Fit the four parametric distributions
Fit GLMs with error distributions P, NB, ZIP and ZINB with log link function. Include prior biological information as explanatory covariates in the model.

Determine the distribution that best fits the data
Select the model with the least Bayesian Information Criterion (BIC) value

Perform further model adequacy tests
1. Deviance goodness of fit test for genes following Poisson & Negative Binomial distributions
2. Test for zero inflation using (mixture) likelihood ratio test

Classification
Classify each gene as following one of the four distributions

Simulation study

- Simulation performed using "3k PBMCs from a Healthy Donor, v1 Chemistry" from 10x Genomics to learn the most appropriate distribution for each gene.
- Using model classification and parameters estimated from the PBMC data, we simulated P, NB, ZIP, and ZINB genes.

Table 1: Correct classification rate (averages calculated over 20 replications

Sample Size	True Gene Category			
	Poisson	Negative Binomial	Zero-inflated Poisson	Zero-inflated Negative Binomial
2000	0.90	0.82	0.62	0.59
5000	0.92	0.85	0.74	0.68

Case study

Study I

- We applied the modelling framework to scRNA-seq measurements collected from mice on two tissues, adipose and muscle.

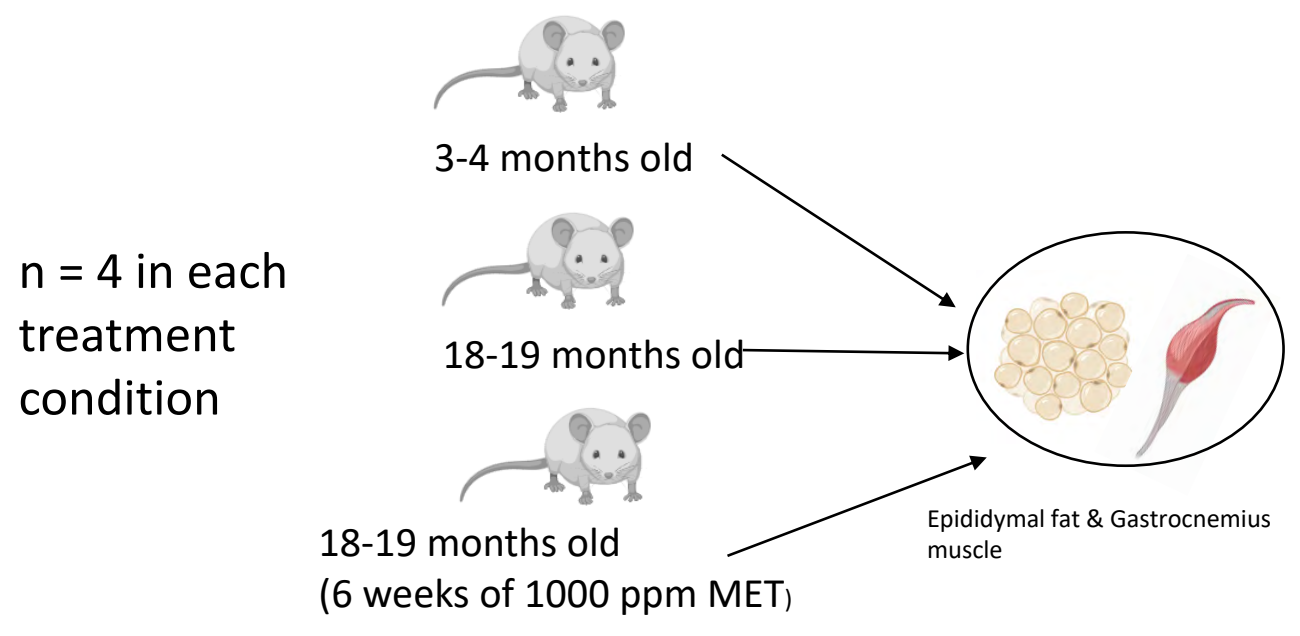

3-4 months old
n = 4 in each treatment condition
18-19 months old
18-19 months old
(6 weeks of 1000 ppm MET)
Epididymal fat & Gastrocnemius muscle

Figure 3: Experimental design [3] study I

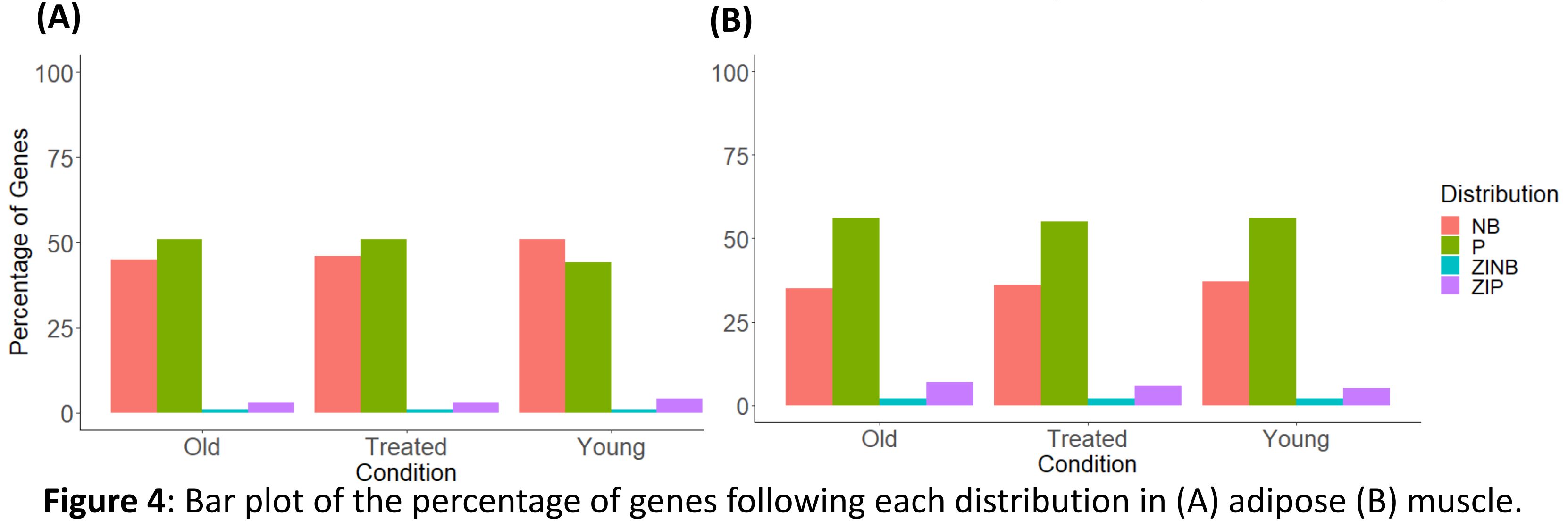


Figure 4: Bar plot of the percentage of genes following each distribution in (A) adipose (B) muscle. Fitted GLM includes cell type and mouse ID as explanatory covariates.

- Among the differentially distributed genes we find the transcription factors (TFs):
- FOXO3 often referred to as the "longevity gene" [4] ;
- RXRA overexpression of which reduces DNA damage accumulation leading to delays in replicative senescence [5] in adipose
- SRF reduction of which leads to premature aging in skeletal muscle [6] ;
- IRF3 a novel inhibitor of cellular senescence and inducer of cell growth inhibition [7] in muscle

Study II

- We applied our framework to publicly available COVID-19 dataset by Wilk et al. [8] , which has over 40,000 cells with multiple donors and ~20 cell-types.
- Through visualisation of some of the differentially distributed genes, we can see that our framework captures subtle changes in gene expression.

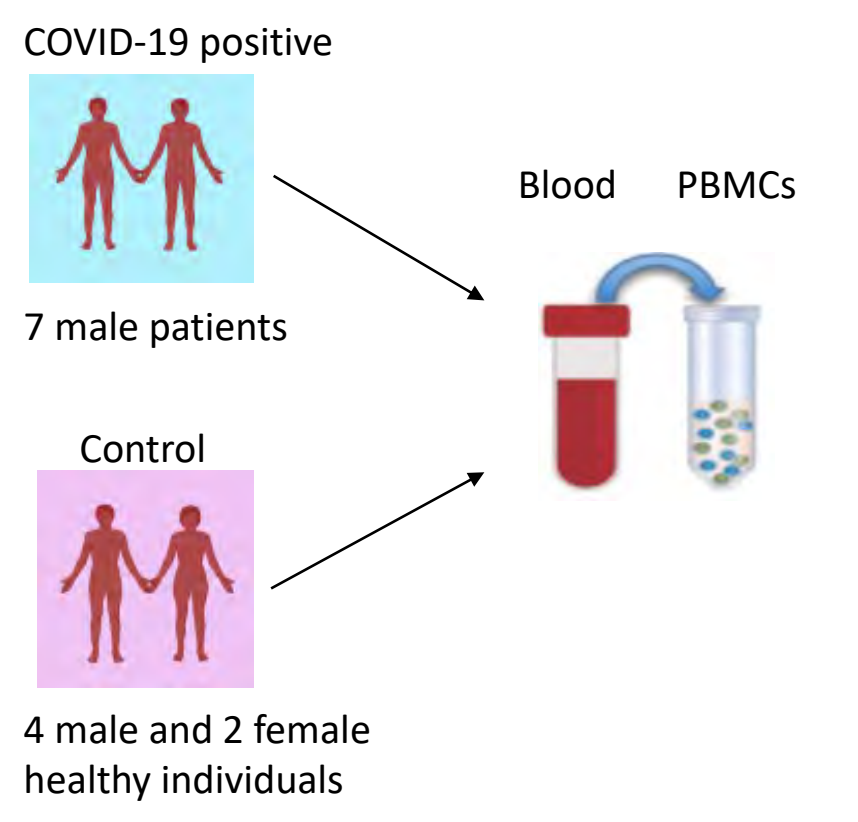


Figure 5: Experimental design [8] study II

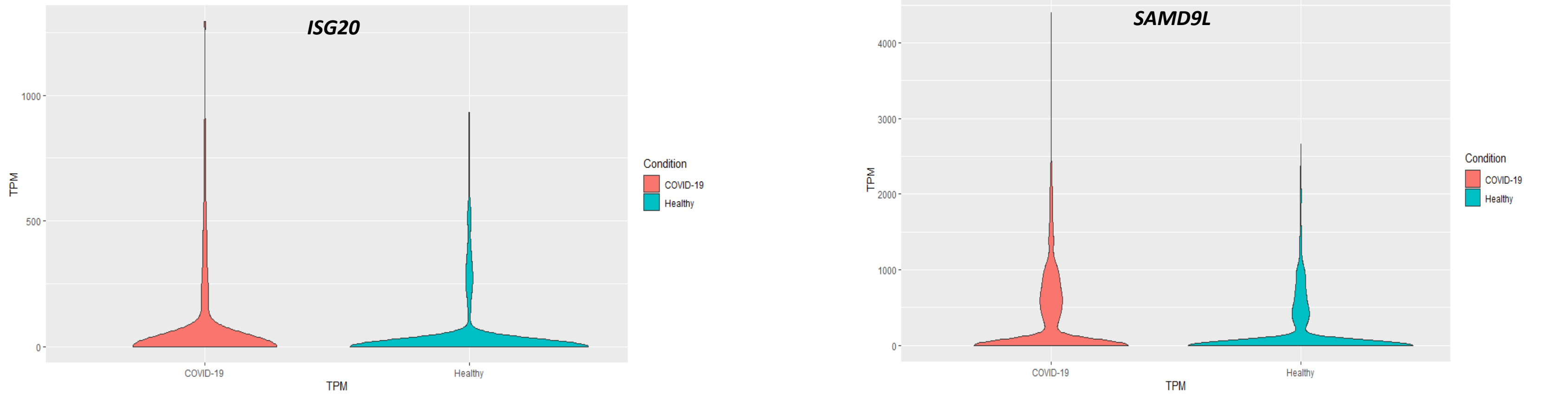


Figure 6: Violin plots of differentially distributed interferon genes

Conclusion

In summary here we present a novel statistical framework that can;

- identify and classify genes according to their shape of gene expression distribution;
- handle excess zeros in scRNA-seq data;
- adjust for covariates (e.g. batch effects, cell-types etc.);
- compare multiple groups of treatment conditions.

References

[1] Mar, J.C., The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond. Biophys Rev, 2019. 11(1): p. 89-94.
[2] Korthauer, K.D., Chu, L., Newton, M.A. *et al*. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol **17**, 222 (2016).
[3] Kulkarni, A.S., et al. Single Cell RNA-Seq reveals the gerotherapeutic effects of metformin in a cell-type-specific manner in mouse muscle and adipose SVF. Manuscript in Preparation.
[4] Flachsbart, F., et al., *Identification and characterization of two functional variants in the human longevity gene FOXO3.* Nature Communications, 2017. **8**(1): p. 2063.
[5] Ma, X., et al., *The nuclear receptor RXRA controls cellular senescence by regulating calcium signaling.* Aging Cell, 2018. **17**(6): p. e12831.
[6] Lahoute, C., et al., *Premature aging in skeletal muscle lacking serum response factor.* PLoS One, 2008. **3**(12): p. e3910.
[7] Kim, T.K., et al., *Interferon regulatory factor 3 activates p53-dependent cell growth inhibition.* Cancer Lett, 2006. **242**(2): p. 215-21.
[8] Wilk, A.J., Rustagi, A., Zhao, N.Q. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat Med 26, 1070–1076 (2020).