noisyR: enhancing biological signal in sequencing datasets by characterizing random technical noise

Ilias Moutsopoulos, Lukas Maischak, Elze Lauzikaite, Sergio Vasquez Urbina, Eleanor Williams, Hajk-Georg Drost, Irina Mohorianu*

Ilias Moutsopoulos – im383@cam.ac.uk

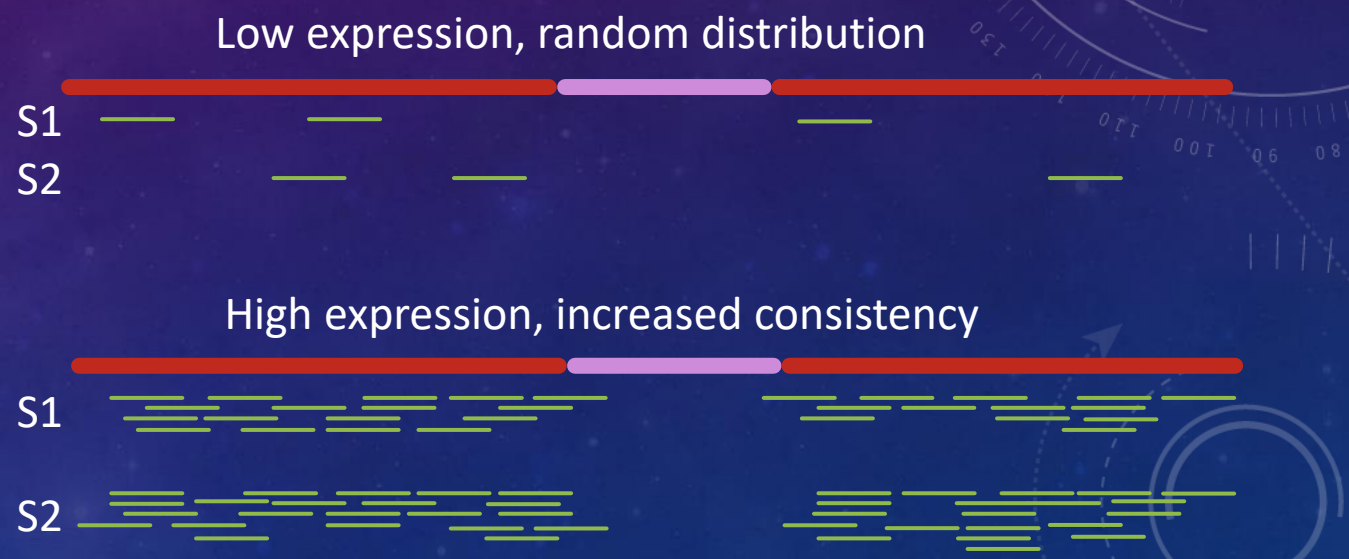
Research Assistant – Bioinformatician

Core Bioinformatics group (Mohorianu group)

Wellcome-MRC Cambridge Stem Cell Institute

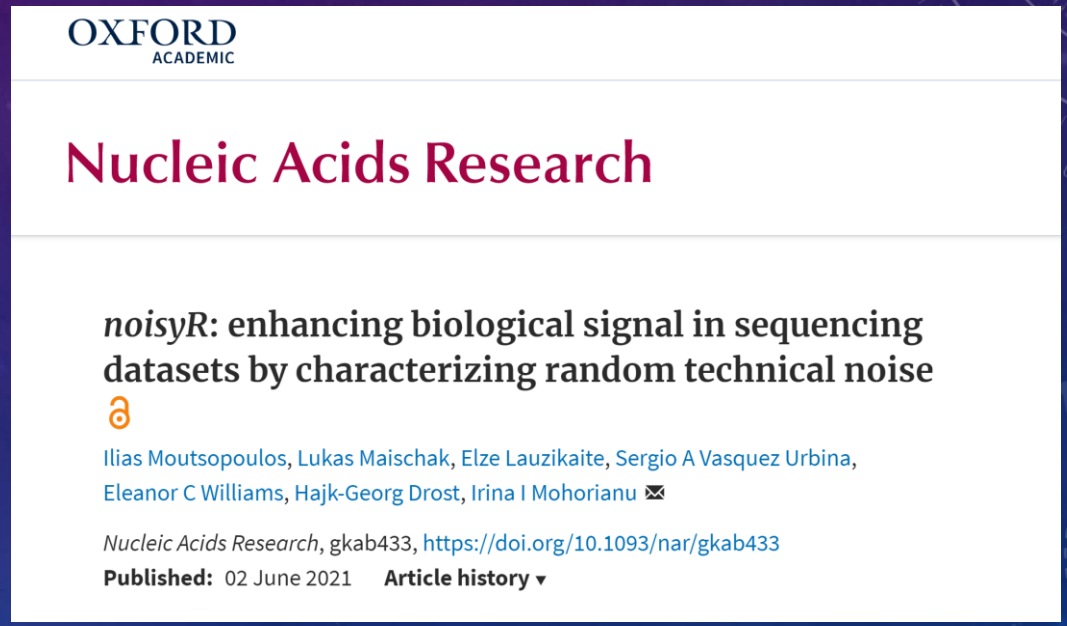
Noise is an intrinsic feature of biological systems, which is amplified by the process of sequencing

- Sequencing technologies are constantly improving and increasing in throughput
- Low-level expression variations are an intrinsic characteristic of next generation sequencing
- At low abundance, read localisation will inevitably vary



The *noisyR* paper presents a newly developed noise filter for next generation sequencing data

- Consistent reduction of random background noise is still challenging
- Meaningful biological signal can be masked by the presence of noise
- Different downstream analysis methods often lead to incompatible results and conclusions



OXFORD ACADEMIC

Nucleic Acids Research

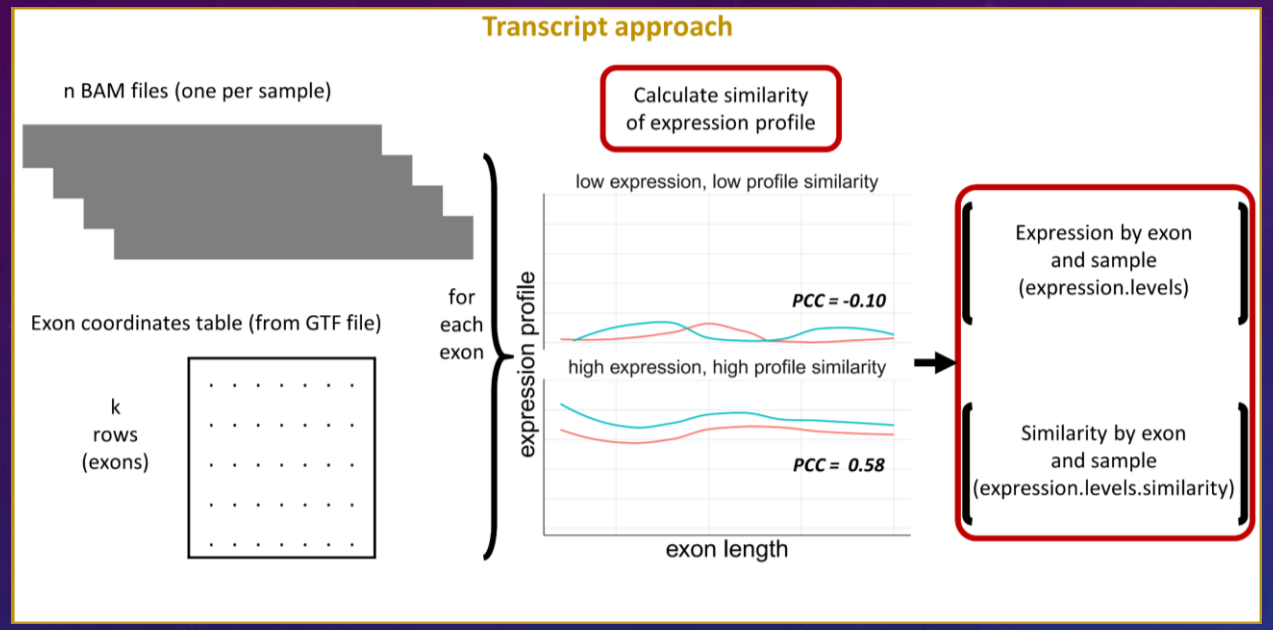
noisyR: enhancing biological signal in sequencing datasets by characterizing random technical noise

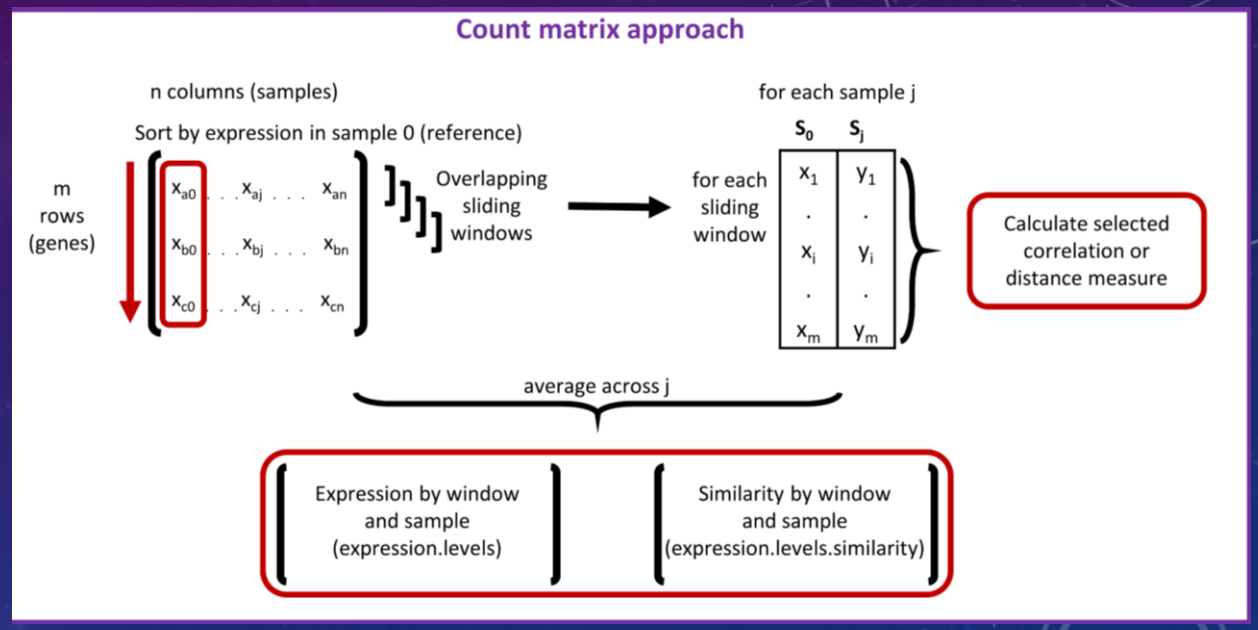
Ilias Moutsopoulos, Lukas Maischak, Elze Lauzikaite, Sergio A Vasquez Urbina, Eleanor C Williams, Hajk-Georg Drost, Irina I Mohorianu

Nucleic Acids Research, gkab433, https://doi.org/10.1093/nar/gkab433

Published: 02 June 2021 **Article history**

Two approaches depending on the input data





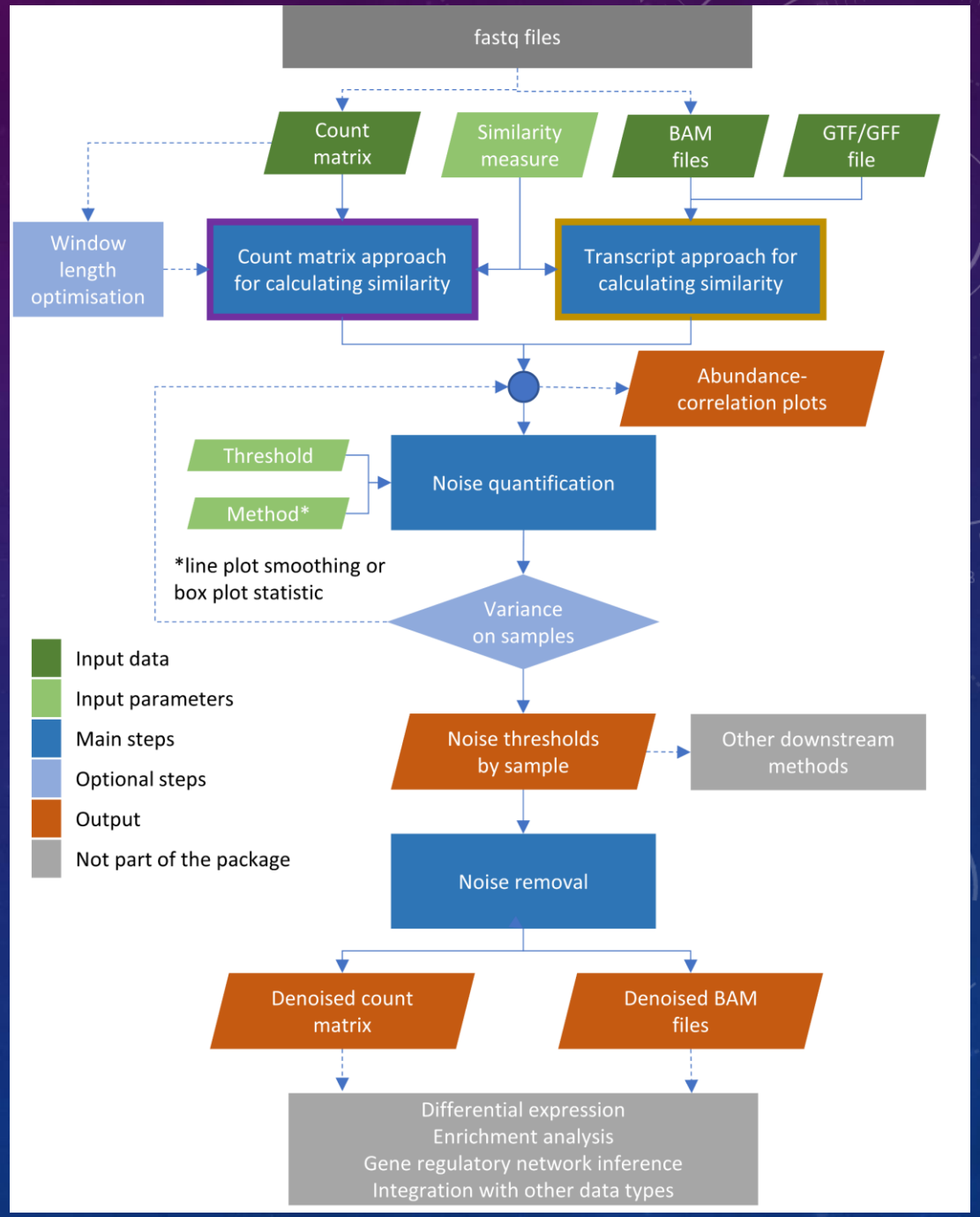
- The transcript approach uses the alignment (BAM) files
  - High computation resources
  - Significantly slower (usually hours)
  - More precise results
- The count matrix approach only uses the total expression
  - Low computation resources
  - Very fast (a few seconds/minutes)
  - Rough results

The software is available as an R package and structured so that little expertise is required to use it

- The ***noisyR*** package is conceptually broken down into 3 main steps:
  - Noise Identification (2 approaches)
  - Noise Quantification
  - Noise Removal
- The "black-box" function *noisyr* can be used to run the pipeline end-to-end
- Parameters can be defaults, user-defined, or optimised at each step, even when running the full pipeline

CRAN: *https://cran.r-project.org/package=noisyr*

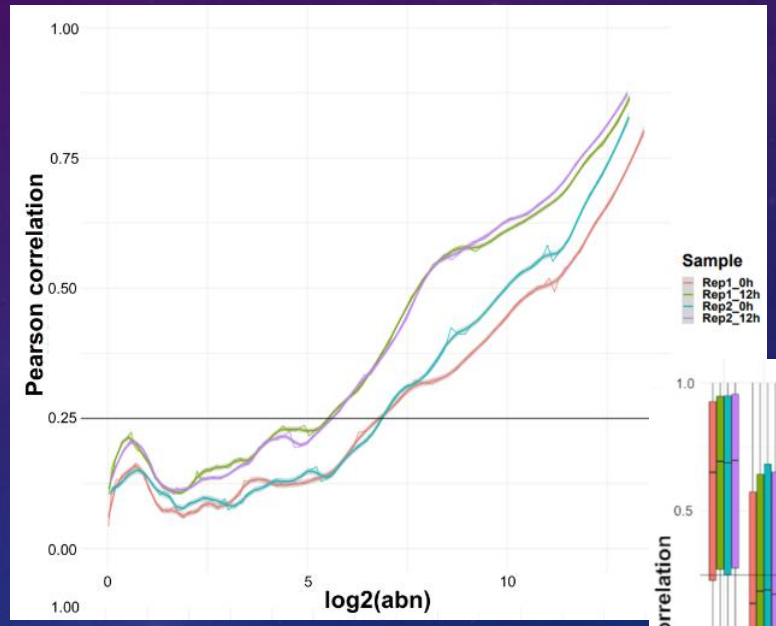
Github: *https://github.com/Core-Bioinformatics/noisyR*

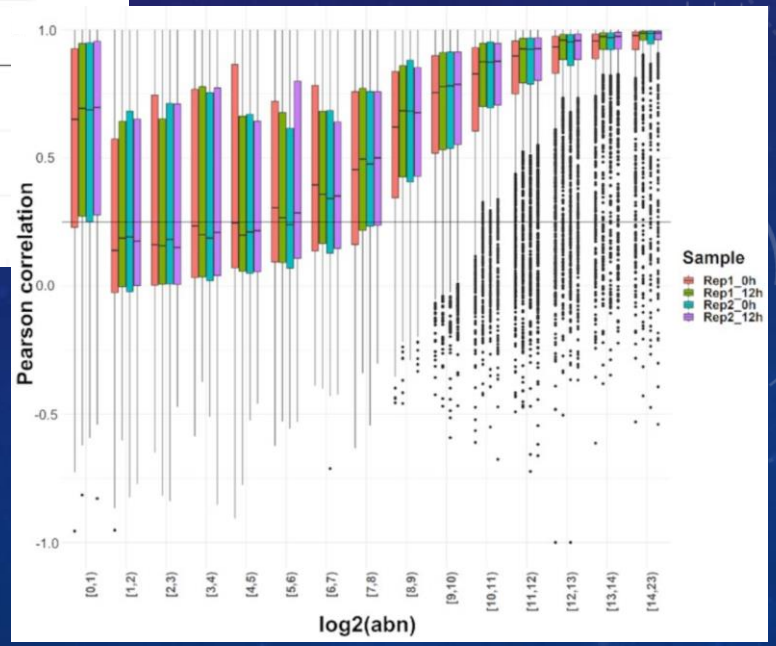


A relationship between abundance and correlation is used to infer a noise threshold

Both approaches produce the same output

- Both use a similarity metric
- They evaluate sample consistency
- Transcript approach calculates per gene, count matrix uses windows

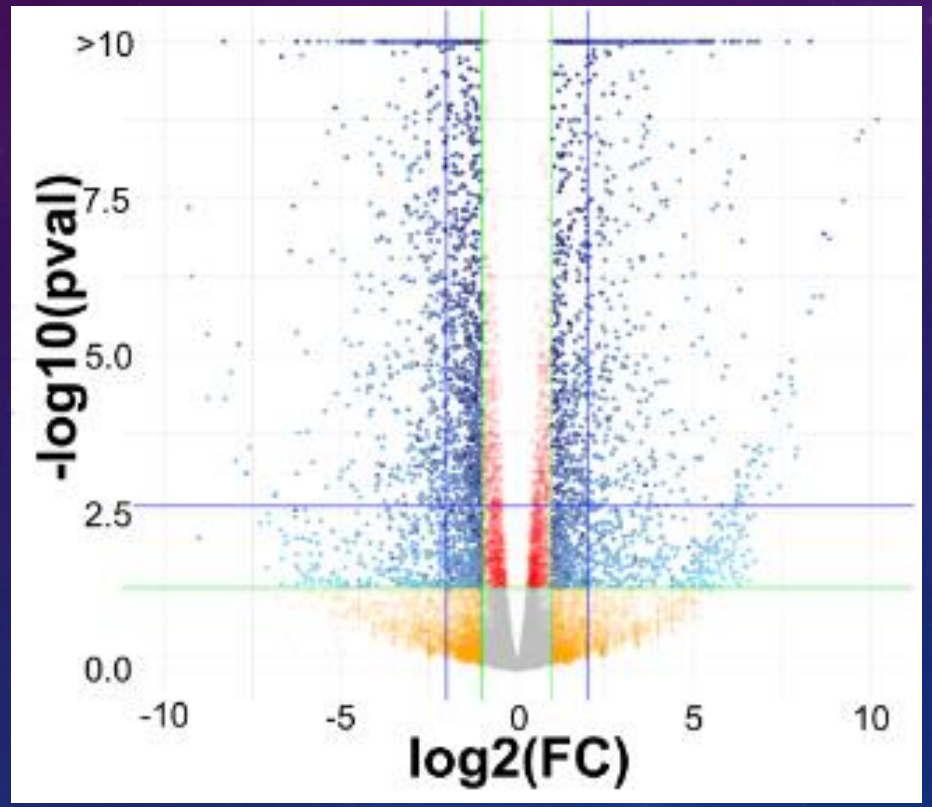


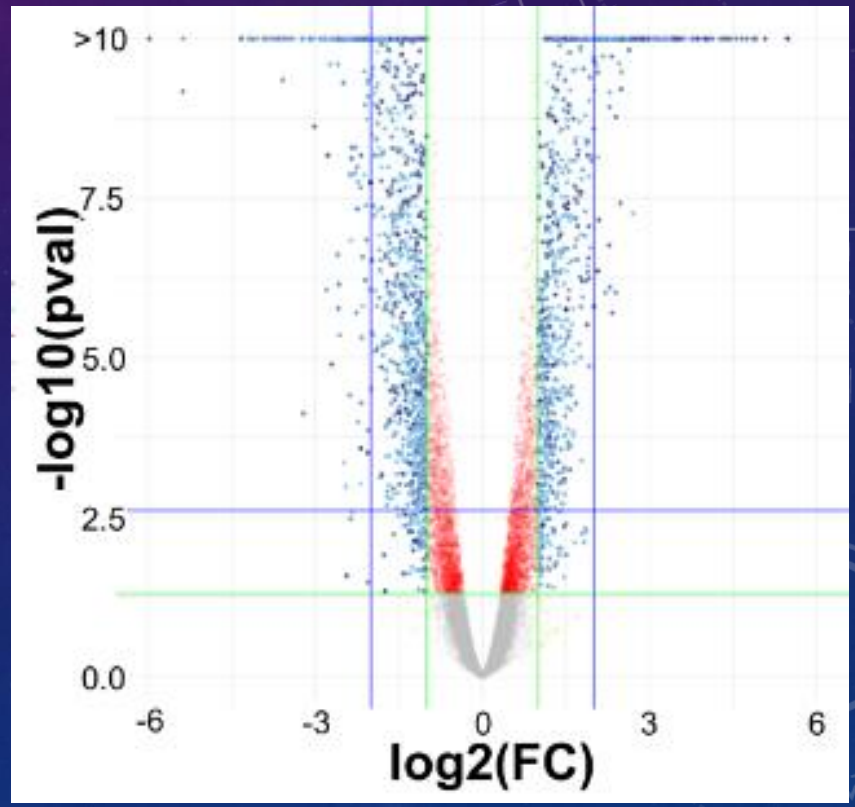


Bulk data from Yang et al (2019)
https://doi.org/10.1016/j.cels.2019.03.012

In bulk mRNA-Seq data, noise removal increases the convergence between analysis methods

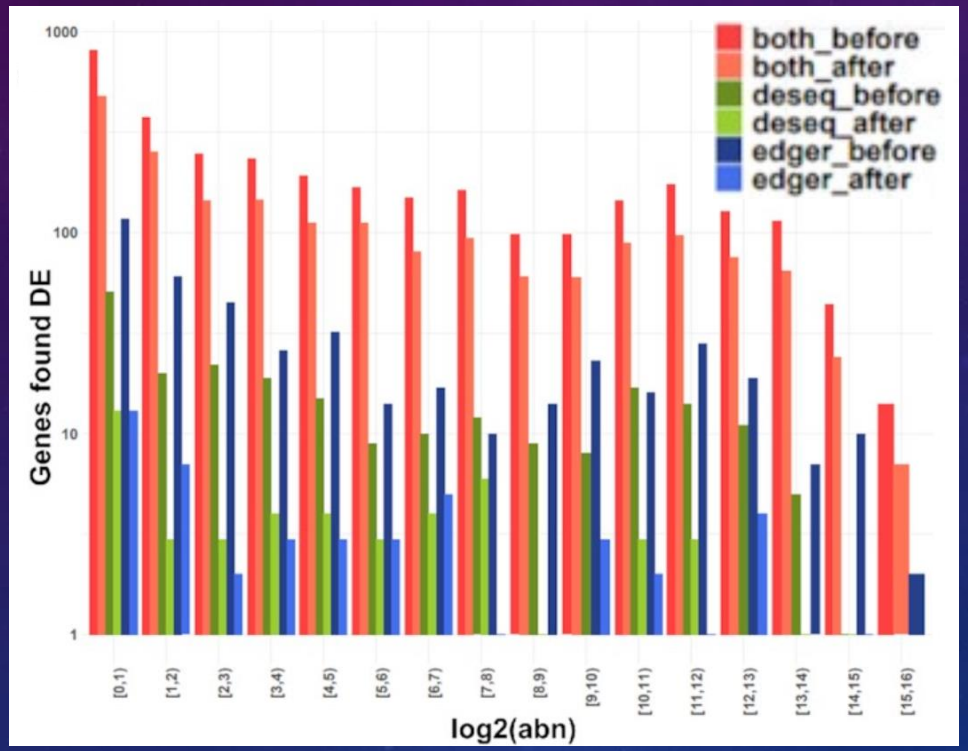
- The number of DE genes is reduced
- Potential false positives are greatly reduced

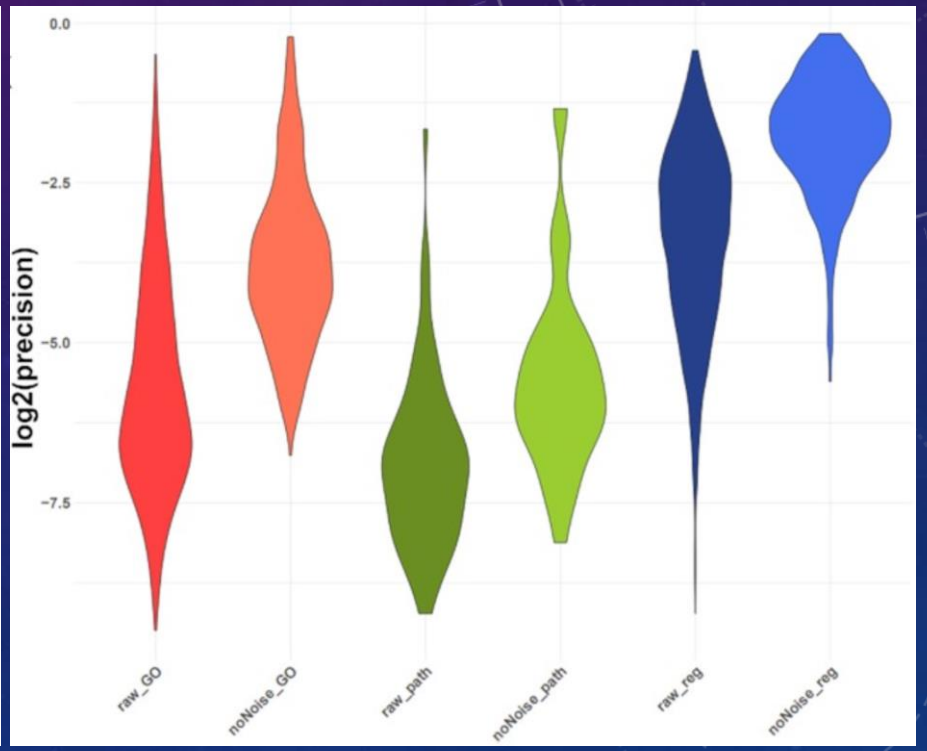




In bulk mRNA-Seq data, noise removal increases the convergence between analysis methods

- Analysis methods converge
- Interpretation is closer to the biological truth



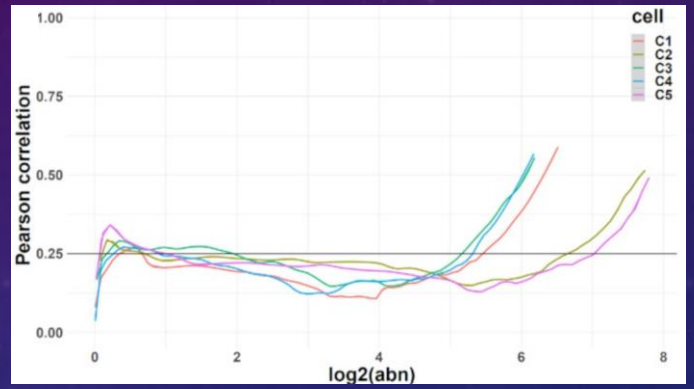


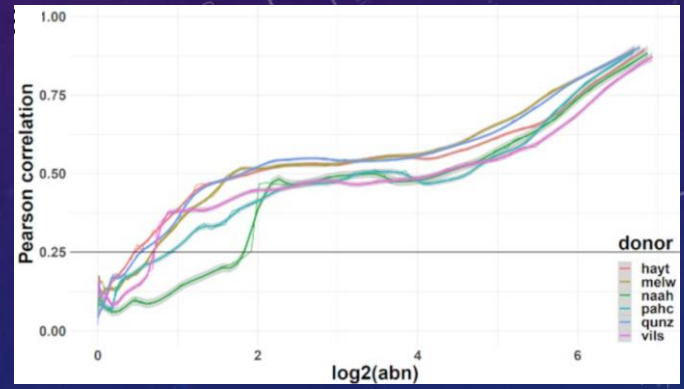
edgeR: McCarthy et al (2012) *https://doi.org/10.1093/nar/gks042*

DESeq2: Love et al (2014) *https://doi.org/10.1186/s13059-014-0550-8*

noisyR is a universal approach that can be applied to other types of sequencing data

- On single-cell (SmartSeq2) data
  - Used pseudo-samples to increase consistency
  - Grouping according to experimental design





- Working on a similar approach for 10x data
- Have illustrated the approach to improve the prediction of micro-RNA targets in plants (Degradome data)
- Will adapt the approach for epigenetic data

Single-cell data from Cuomo et al (2020) *https://doi.org/10.1038/s41467-020-14457-z*

UNIVERSITY OF CAMBRIDGE

wellcome-MRC cambridge stem cell institute

MAX PLANCK GESELLSCHAFT

Thank you for your attention!

Irina Mohorianu

Elze Lauzikaite

Eleanor Williams

Hajk-Georg Drost

Lukas Maischak

Sergio Vasquez Urbina

Alex Calderwood

Yu Zhang

Arash Shahsavari

Andi Munteanu

This work was funded by the Wellcome-MRC Cambridge Stem Cell Institute

approach	method	similarity.thresh	abn.thresh.min	abn.thresh.mean	abn.thresh.coef.var	abn.thresh.max
Density_based_ fixed_threshold	No_ normalisation	N/A	299.00	299.00	0.00	299.00
Density_based_ fixed_threshold	RPM_ normalisation	N/A	298.00	298.00	0.00	298.00
Density_based_ fixed_threshold	Quantile_ normalisation	N/A	296.00	296.00	0.00	296.00
Line_plot	No_smoothing	0.25	17.10	59.06	0.58	127.25
Line_plot	loess10_smoothing	0.25	16.84	58.66	0.60	133.97
Line_plot	loess25_smoothing	0.25	16.84	58.93	0.62	137.50
Line_plot	loess50_smoothing	0.25	19.13	61.51	0.71	165.21
Boxplot	Median	0.25	18.38	60.13	0.57	128.00
Boxplot	IQR	0.25	18.38	61.41	0.57	137.19
Boxplot	Quant5	0.25	18.38	61.41	0.57	137.19

Window length optimization

- Calculate similarity matrix for a range of windows
- Average correlation of each window across samples
- Jensen-Shannon divergence with uniform distribution
- T-test to determine stability

Determining the noise threshold

- Calculate similarity matrix
- Use a range of similarity thresholds and methods
- Calculate the coefficient of variation across samples
- Pick the combination with the lowest variation