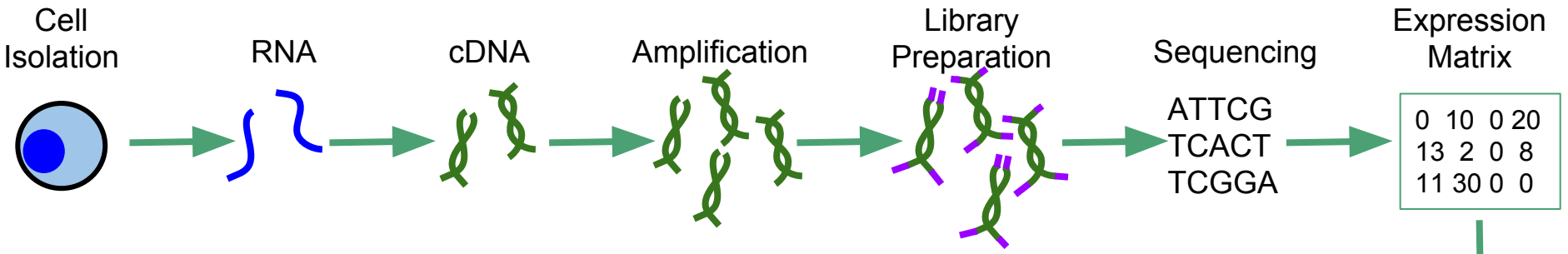
Understanding Nothing: Zeros in scRNASeq

Tallulah Andrews, 27 Sept 2016

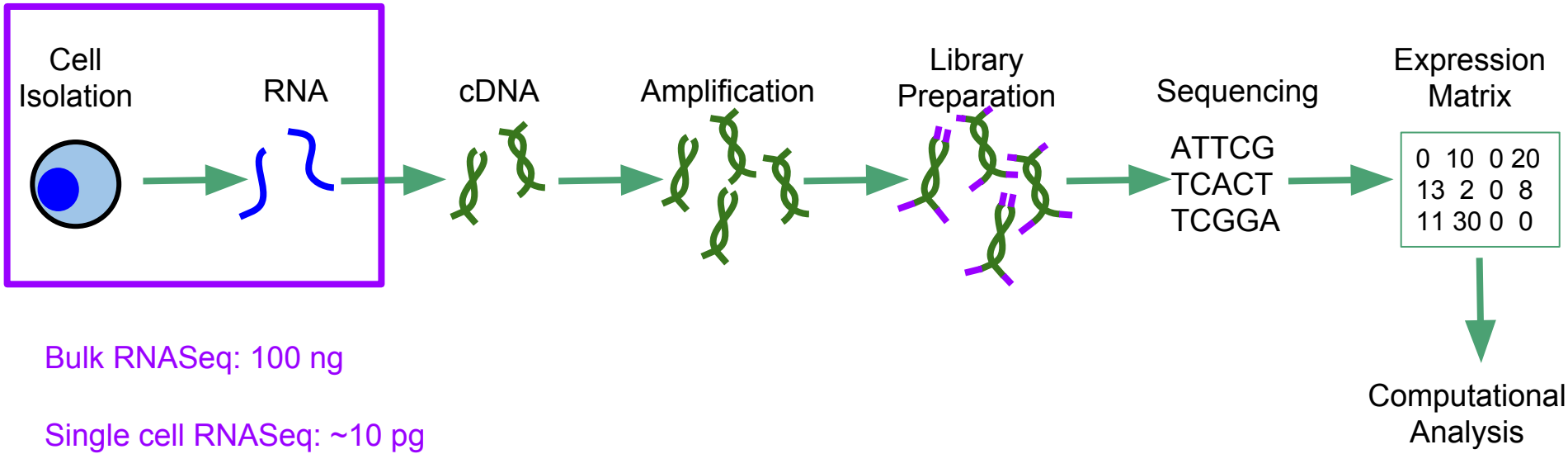
Single-cell vs bulk RNASeq



Enables:

- Unbiased cell-type identification/tissue composition
- Elucidation of cell-fate decisions & development
- Detection of heterogeneity of cellular responses
- Investigation of stochastic gene expression

Single-cell vs bulk RNASeq



Bulk RNASeq: 100 ng

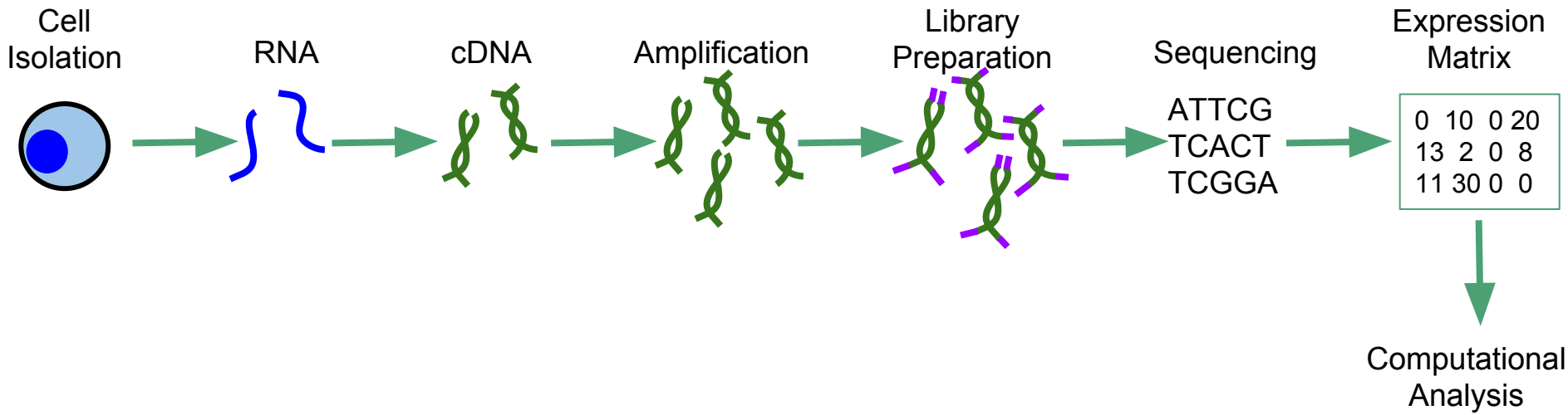
Single cell RNASeq: ~10 pg

Zeros Dominate scRNASeq

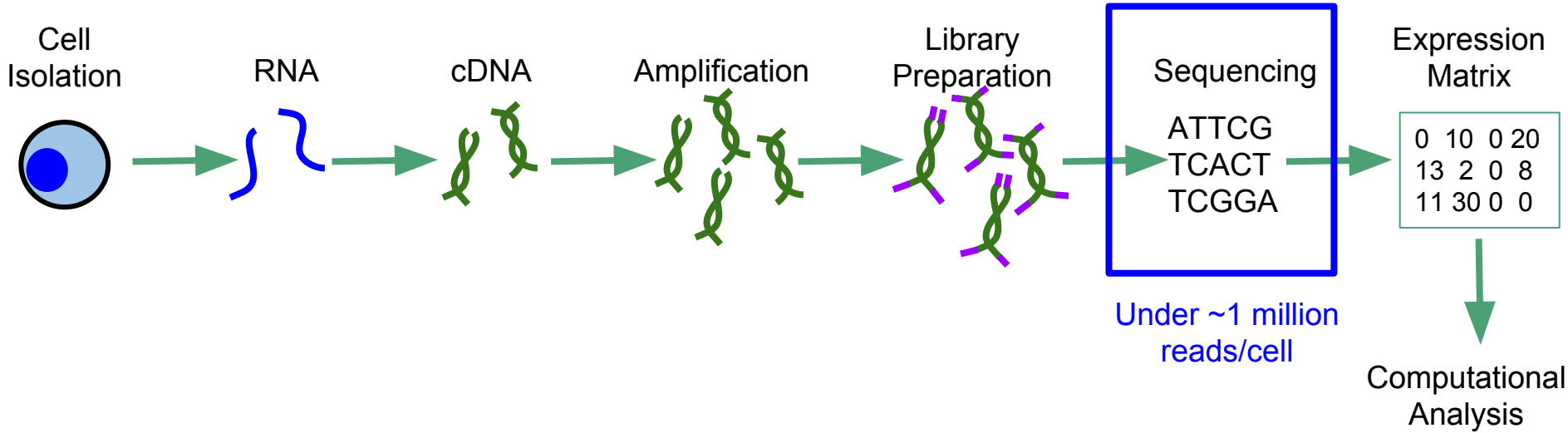
Dataset	Type	No. Cells	No. Genes	**Prop Zero**
Buettner	mouse ESCs	279	17,231	**51.2%**
Shalek	mouse bone marrow	324	12,474	**66.4%**
Deng	mouse embryo	255	17,406	**50.2%**
Usoskin	mouse neuron	530	15,585	**72.5%**
Kirschner	mouse ESCs	2,448	23,729	**62.5%**
Linnarsson	mouse brain	2,542	17,867	**76.9%**
Pollen	human neural	301	19,624	**60.3%**
Zhong	mouse embryo	49	20,558	**38.0%**

*Cells with > 2,000 detected genes
**Genes seen in >3 cells

Source of Zeros

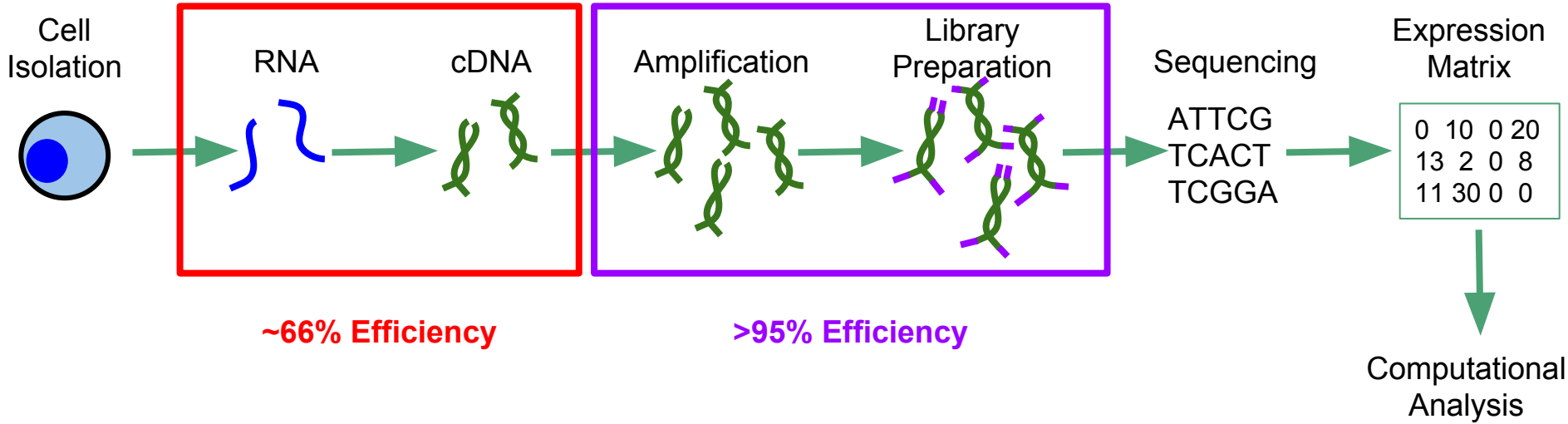


Source of Zeros



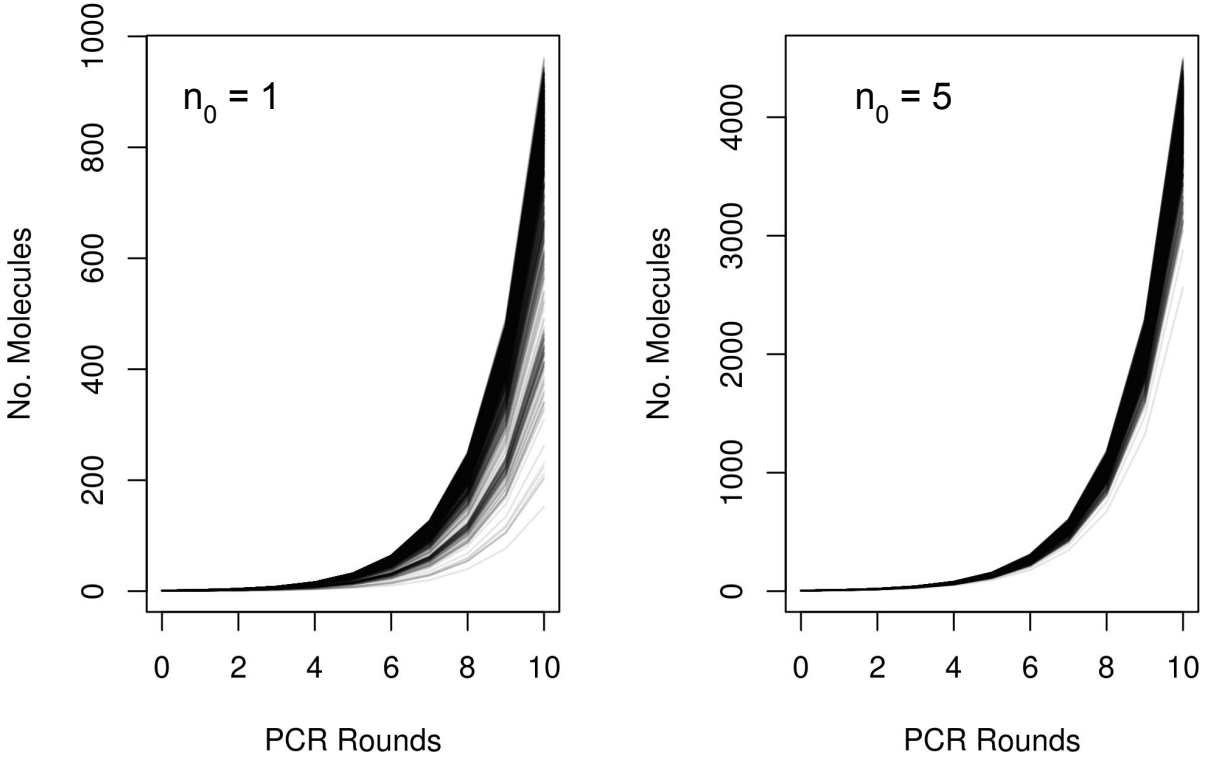
Svensson et al. (2016)

Source of Zeros

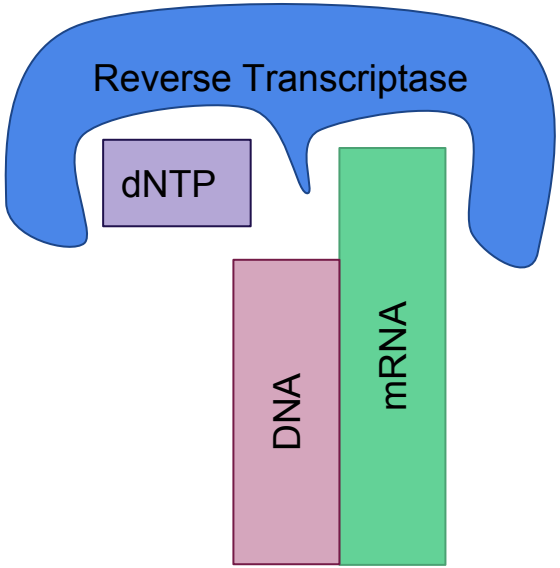


Reiter et al. (2011) & Bengtsson et al. (2008)

RT failure propagates downstream



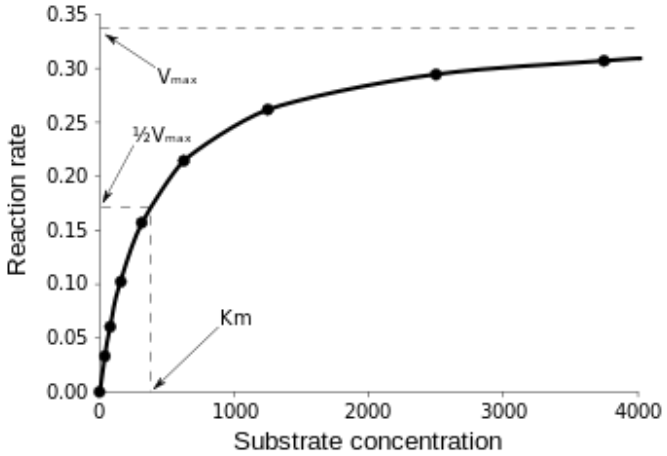
Reverse Transcription = Michaelis-Menten



To model probability:

$V_{max} = 1$

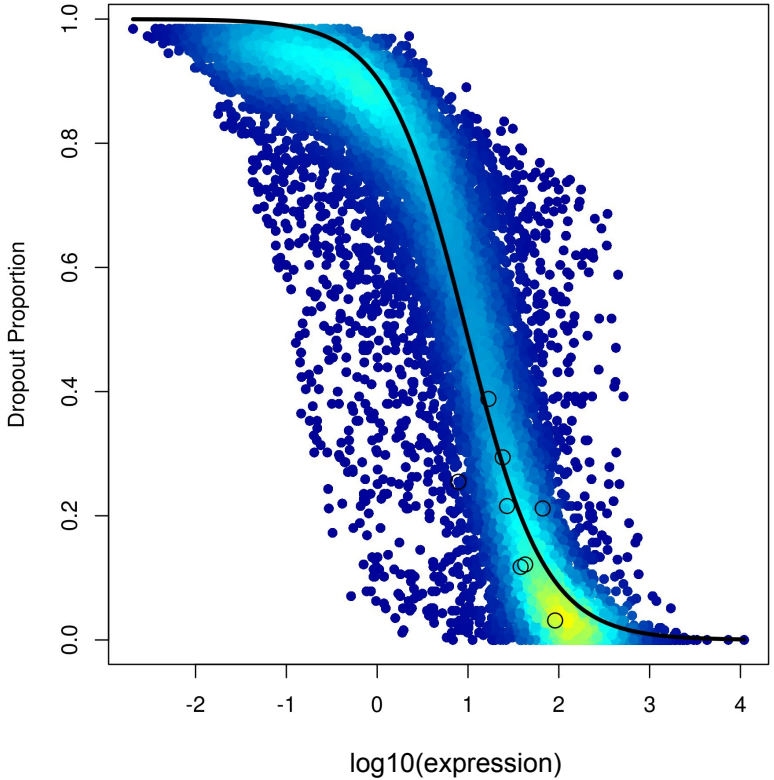
Detection probability $= \dfrac{V_{max}[S]}{(K_M + [S])}$



MM vs Other Models

Michaelis-Menten Modelling of Dropouts (M3Drop)

- $P_{dropout} = 1 - [s]/(K+[s])$
- **For Deng: K = 9.5**



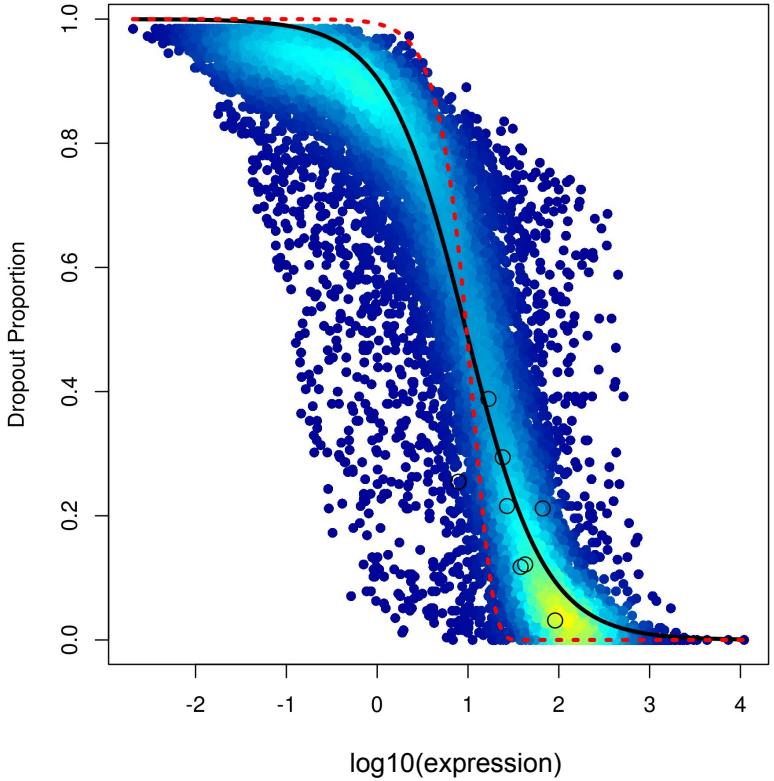
MM vs Other Models

Michaelis-Menten Modelling of Dropouts (M3Drop)

- $P_{dropout} = 1 - [s]/(K+[s])$
- **For Deng: K = 9.5**

Zero Inflated Factor Analysis (ZIFA)

- Dimensionality Reduction for scRNASeq
- $P_{dropout} = e^{-\lambda[s][s]}$
- For Deng: λ = 0.0075



MM vs Other Models

Michaelis-Menten Modelling of Dropouts (M3Drop)

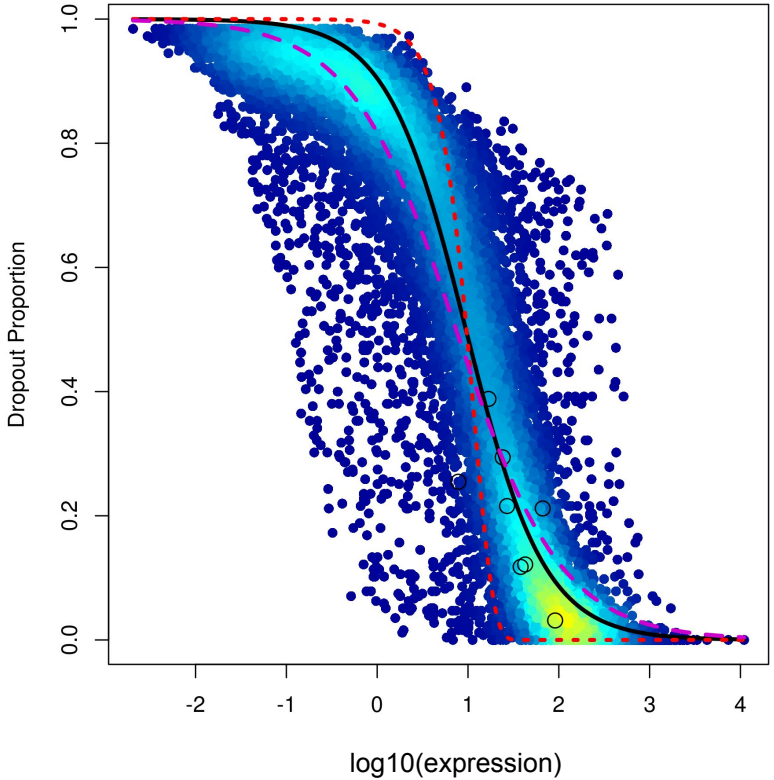
- $P_{dropout} = 1 - [s]/(K+[s])$
- **For Deng: K = 9.5**

Zero Inflated Factor Analysis (ZIFA)

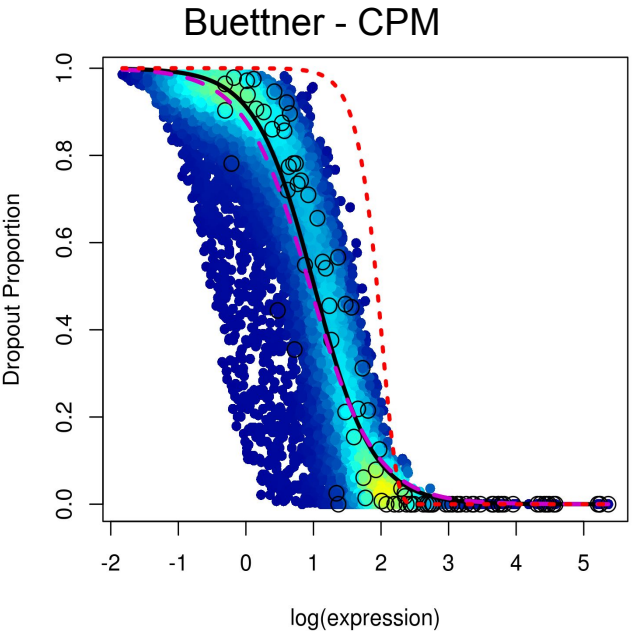
- Dimensionality Reduction for scRNASeq
- $P_{dropout} = e^{-\lambda[s][s]}$
- For Deng: λ = 0.0075

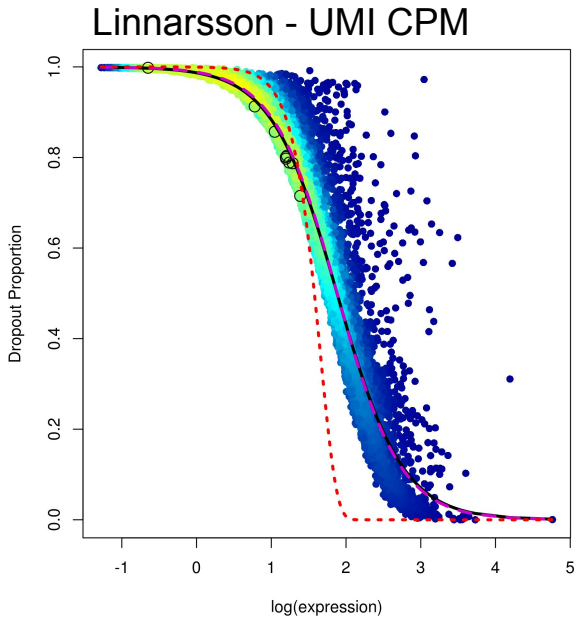
Single Cell Differential Expression (SCDE)

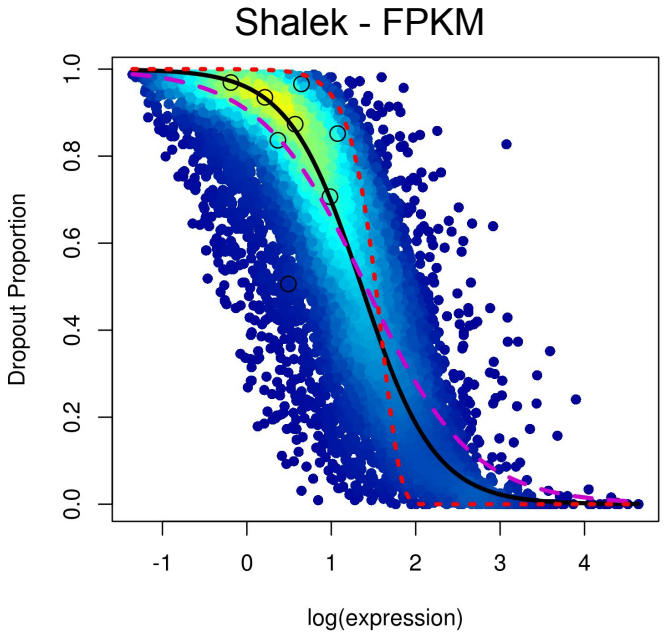
- $P_{dropout} = 1/(1+e^{-(a+b*\log([s]))})$
- For Deng: a = 1.5, b = -0.75



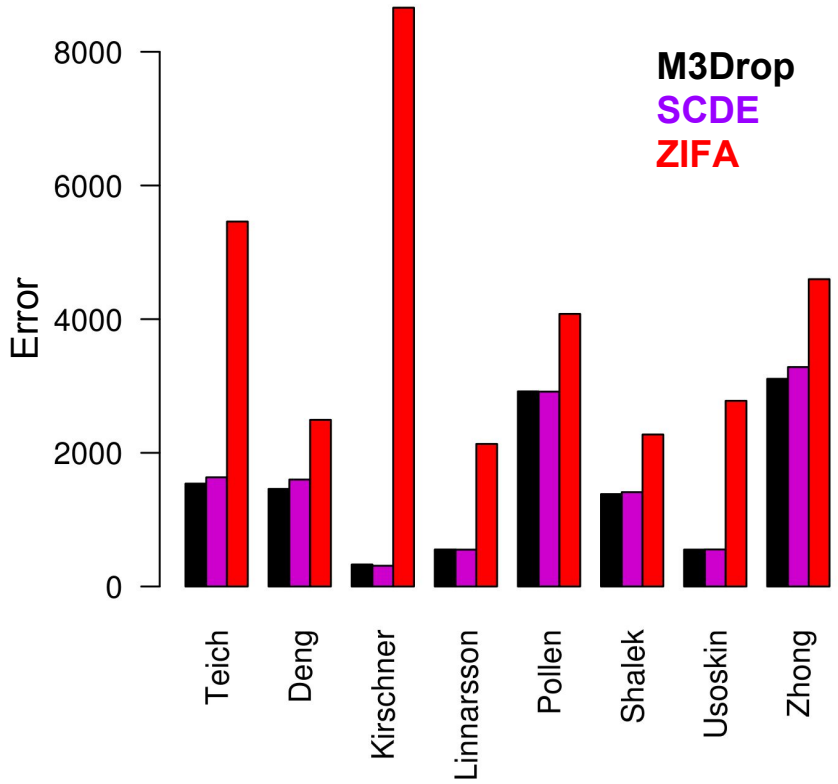
Michaelis-Menten fits diverse datasets.



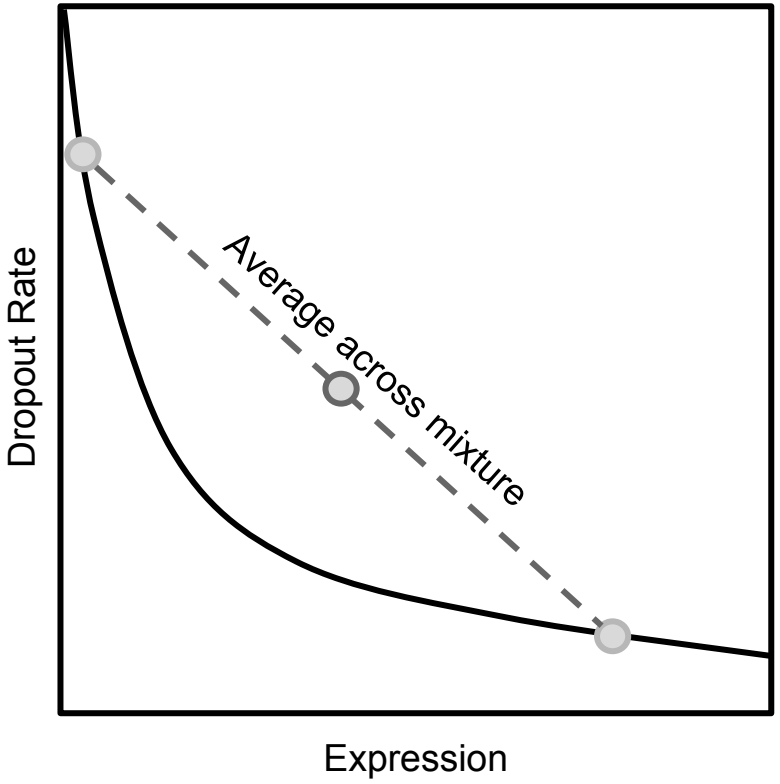


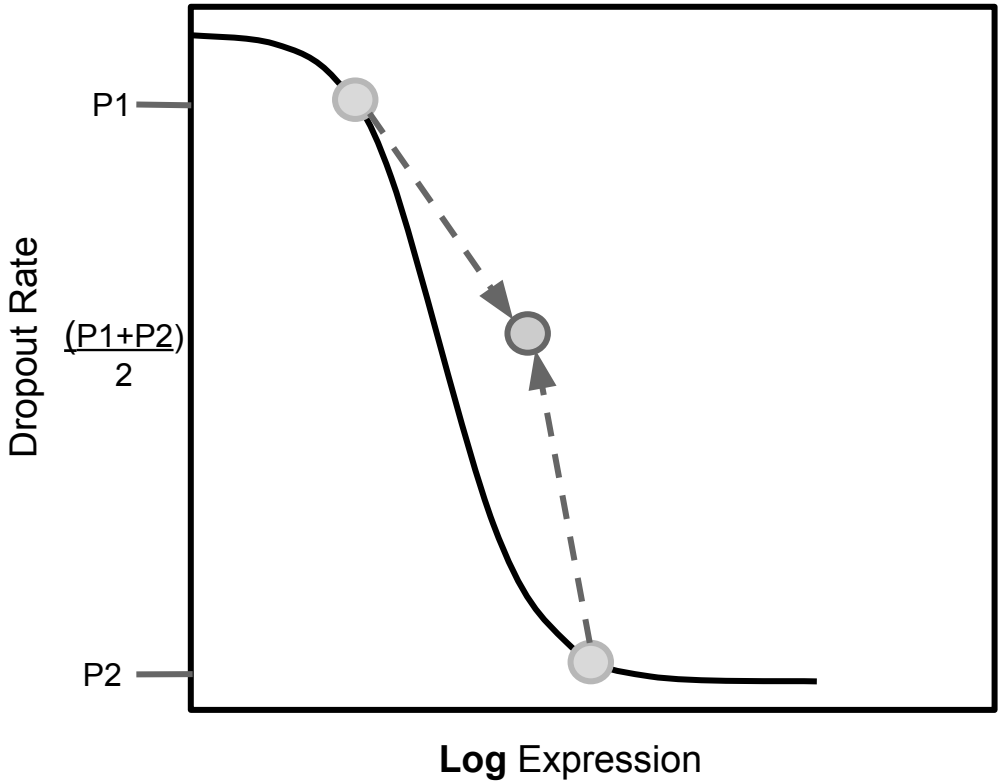


Michaelis-Menten fits diverse datasets.



Differentially Expressed Genes are Outliers



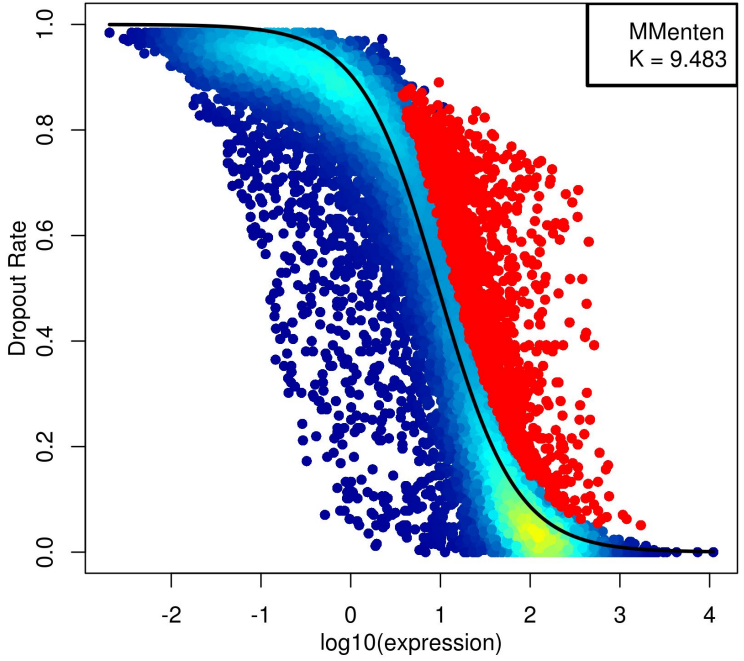


Outlier/DE gene detection

Michaelis-Menten:
$P_{dropout} = 1 - S/(K+S)$

Rearrange to solve for K:
$K = P / (1-P) * S$

1. Calculate K_j for each gene
2. Propagate errors in estimates for S (mean expression) and P (observed dropout rate) to get error for K_j
3. Estimate error of global K_M
4. Test whether K_j is significantly larger than K_M fit across all genes using a Z-test combining errors of (2) & (3)



Highly Variable Genes

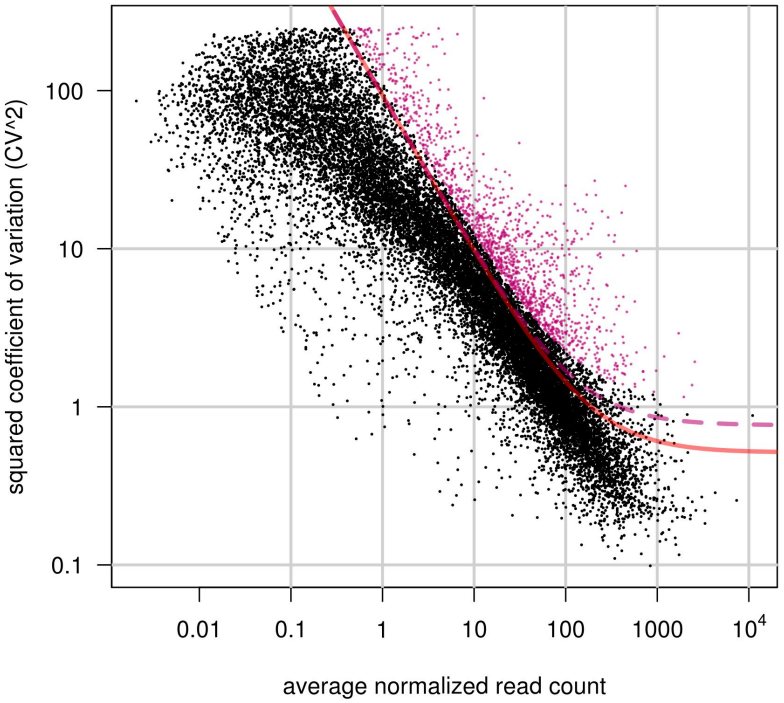
In general:

f(variance) = g(mean)

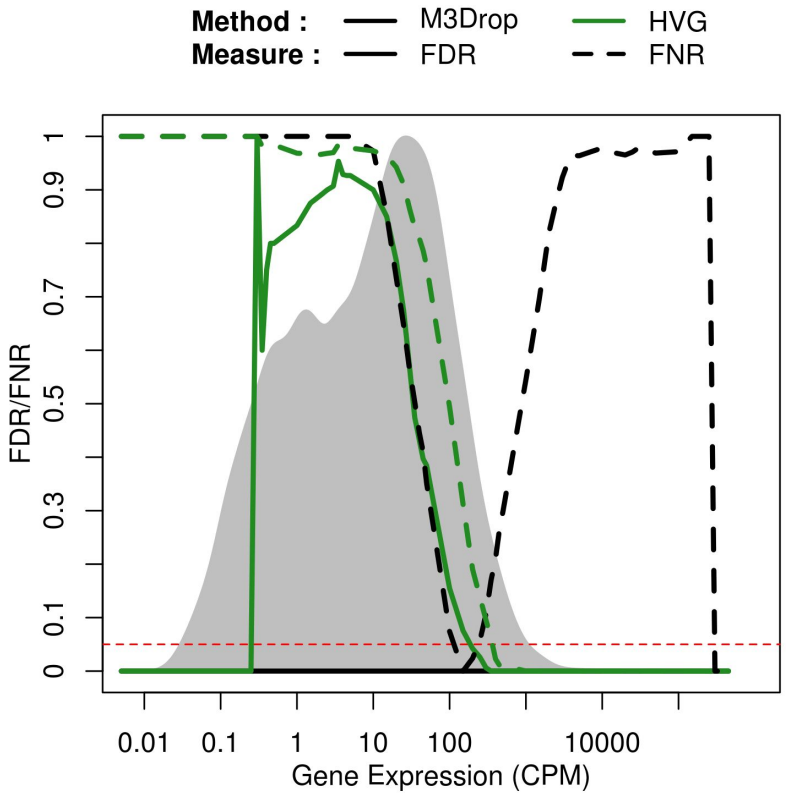
1. Fit a relationship between variance and mean expression
   a. May use all genes or only spike-ins in fitting
2. Identify points above this relationship

Brennecke et al. (2013) :

1. $CV^2 = a_1/\mu + \alpha_0$
2. Significant outliers detected using χ^2-test

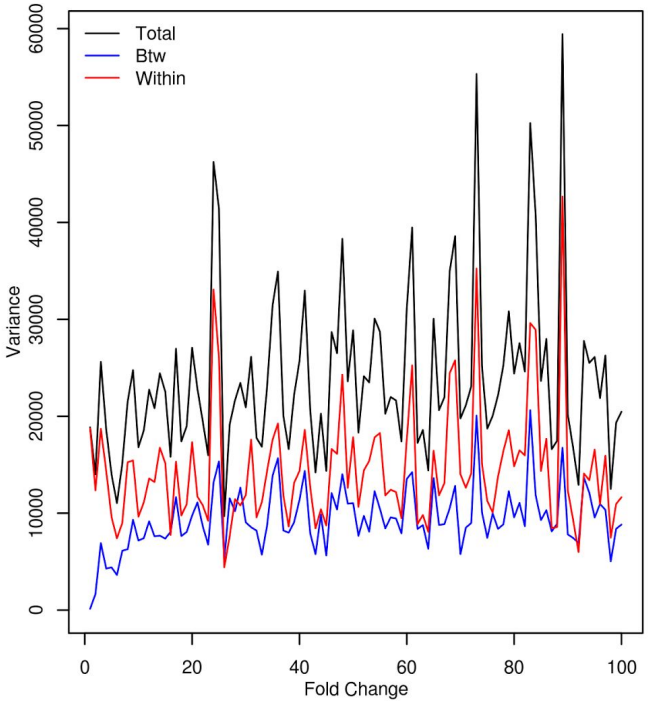


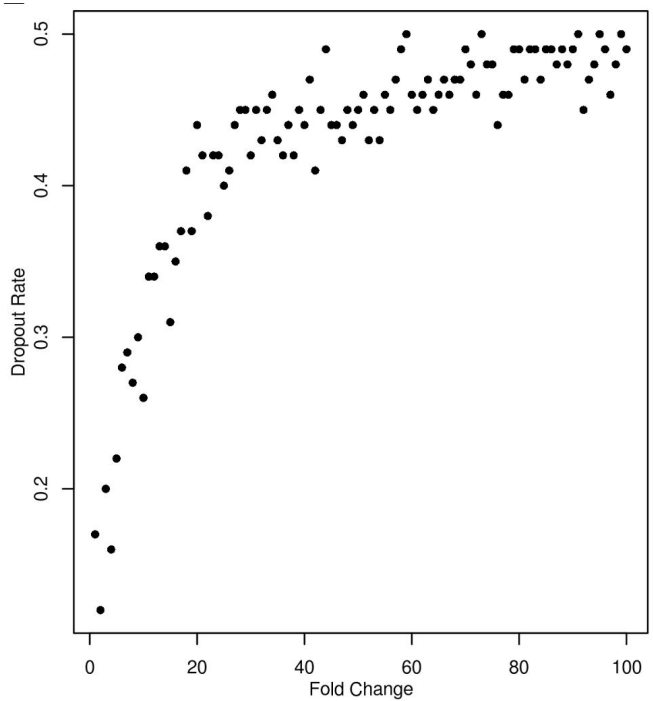
DE Simulations - Dropouts vs Variance.



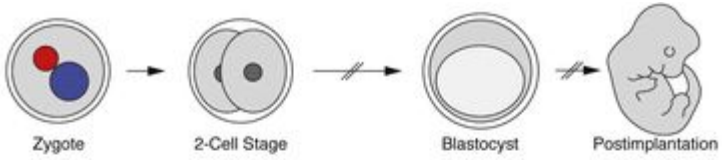
DE Simulations - Dropouts vs Variance.

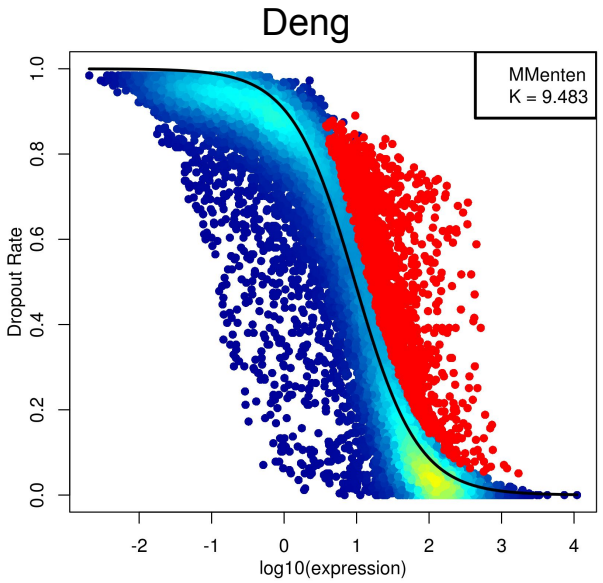
μ = 100, n = 100

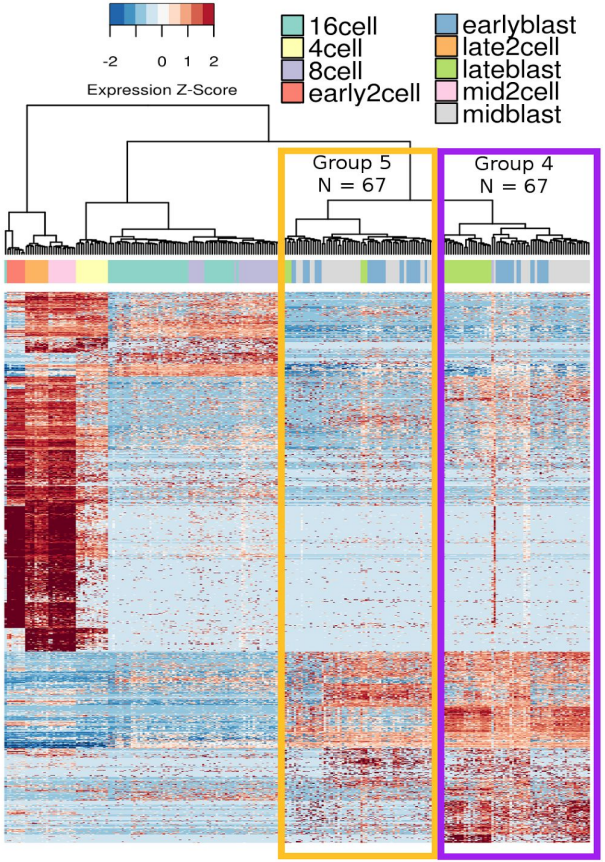




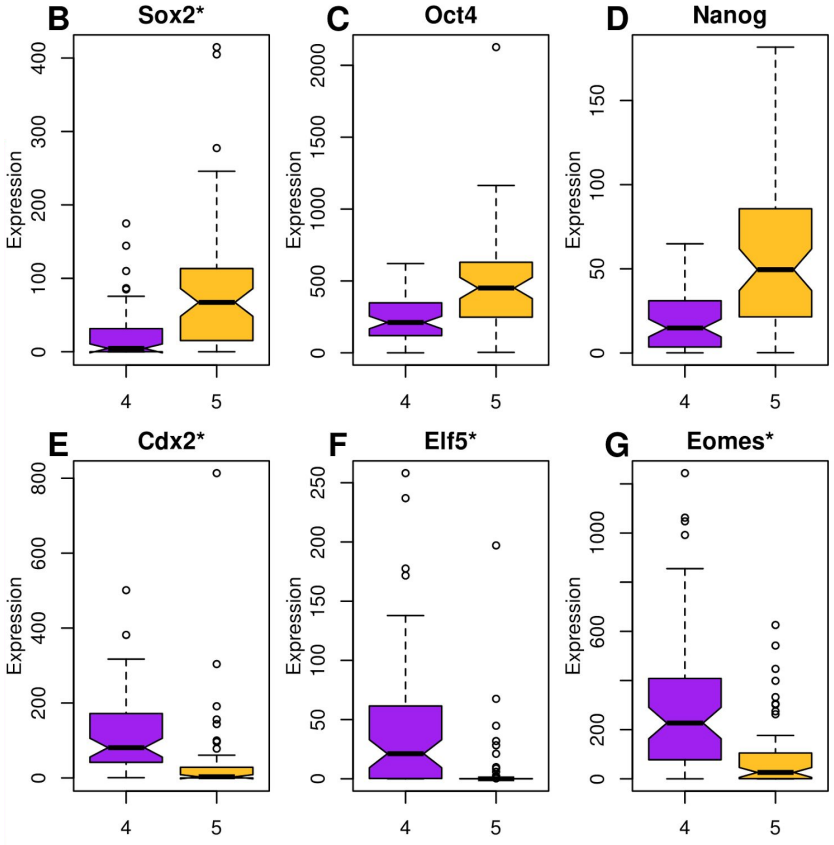
Applying M3Drop to Early Mouse Development

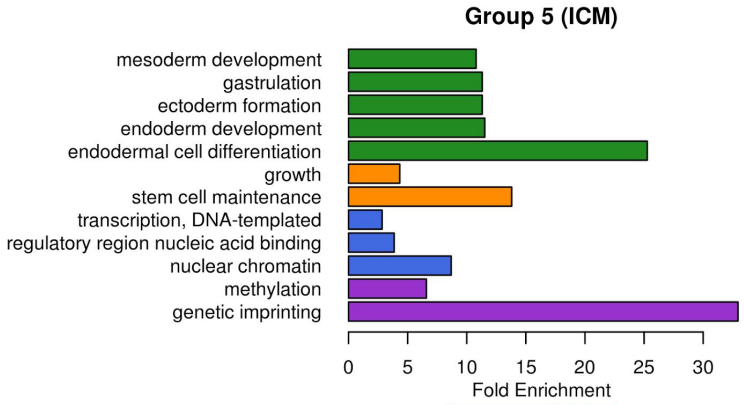


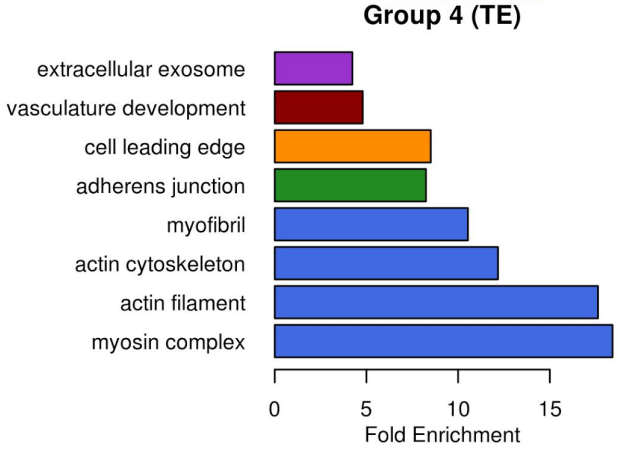




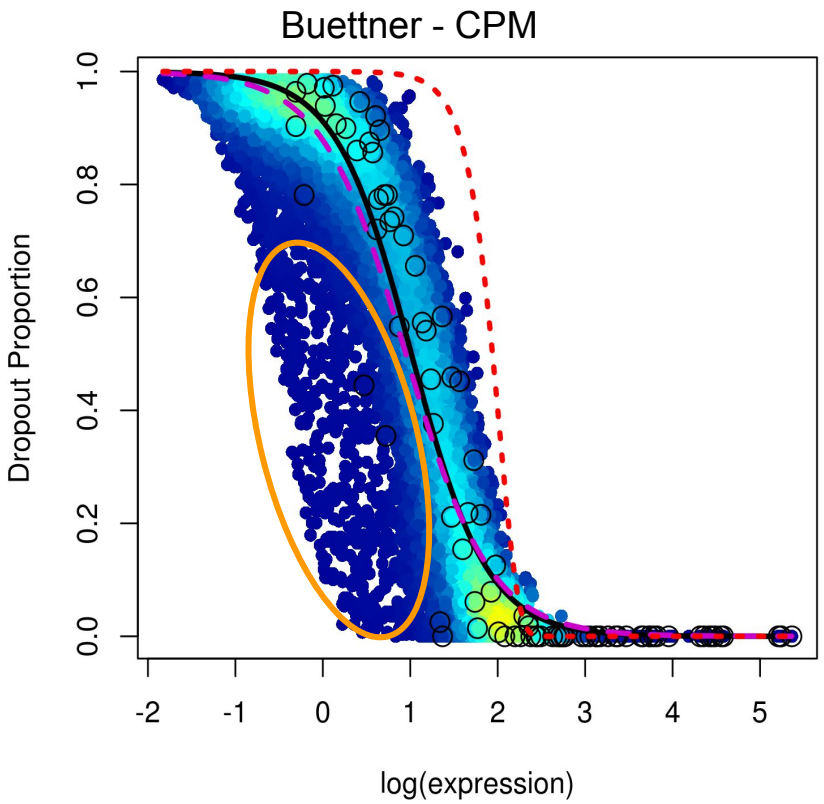
Identification of TE and ICM

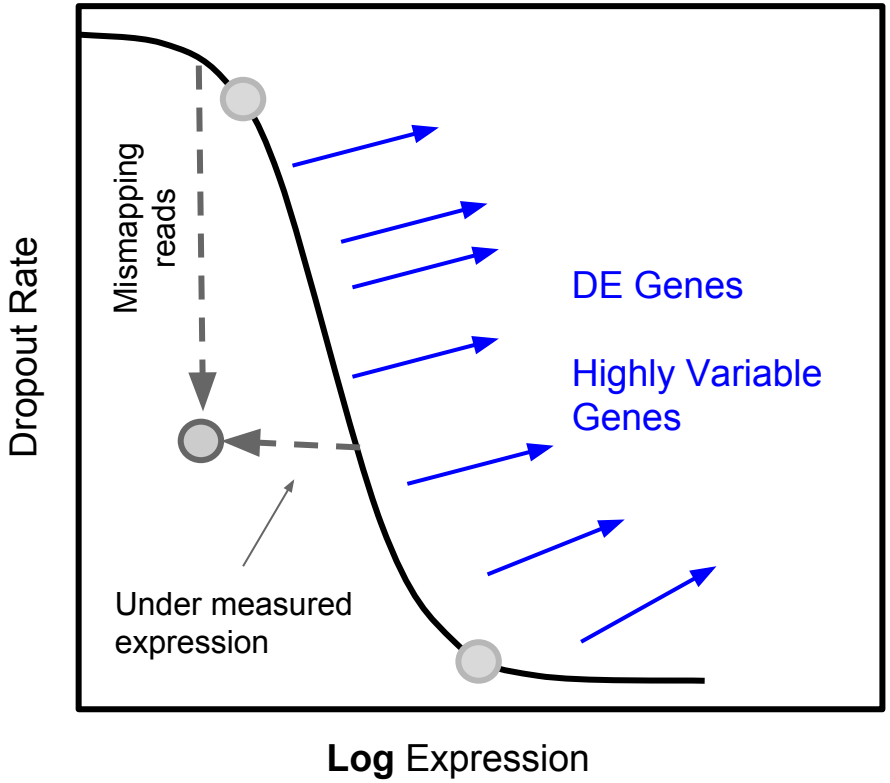






What are outliers to the left?

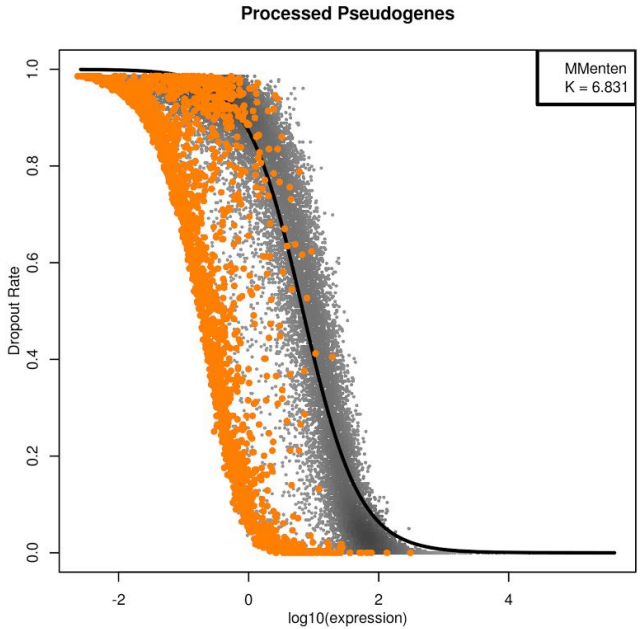




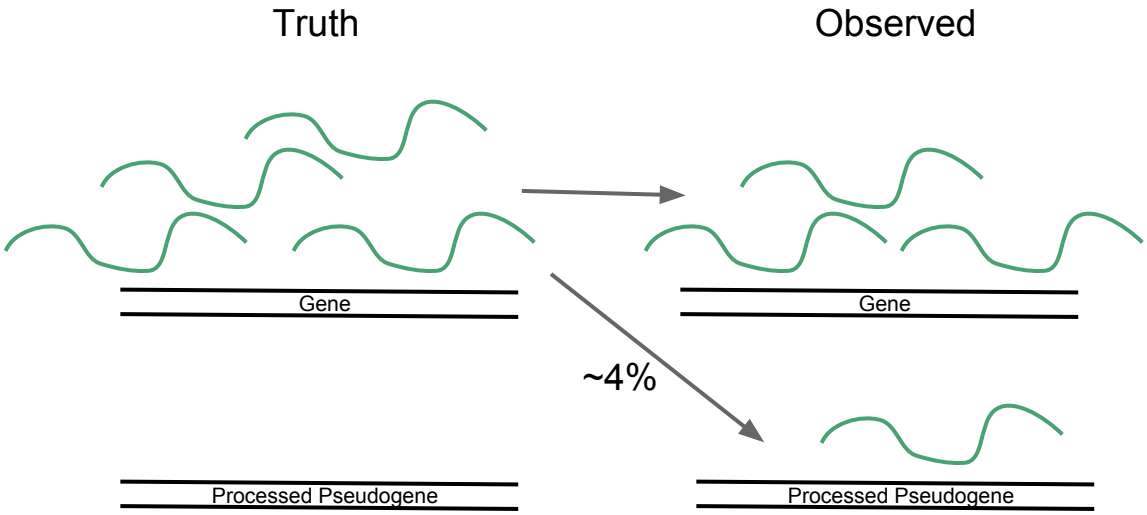
Processed Pseudogenes = True Negatives

- Identical sequence to original transcript
- Lacks introns
- Lacks promoters & regulatory sequences
  - Assumed to not be transcribed
- >3,000 identified in the mouse genome
  - only 150 have confirmed expression

Processed Pseudogenes - Mismapping Reads



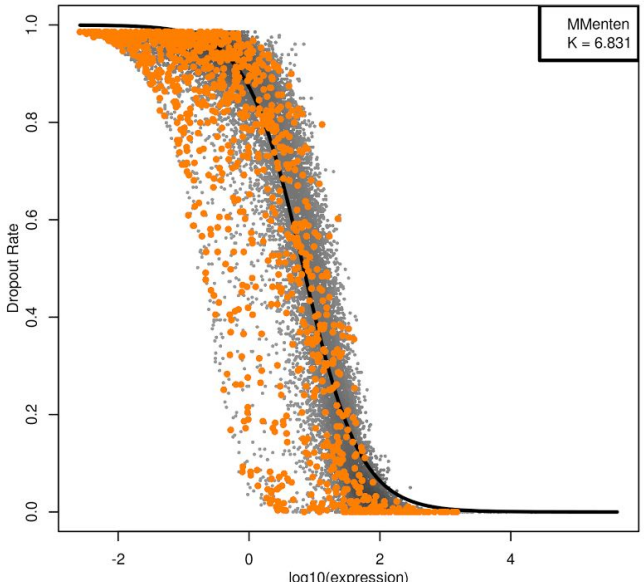
Processed Pseudogenes
Left shifted by 1.4 (p ~ 0)



1% sequencing error rate x 100bp reads:
4% of reads have 3+ sequencing errors

Under-Measured Expression

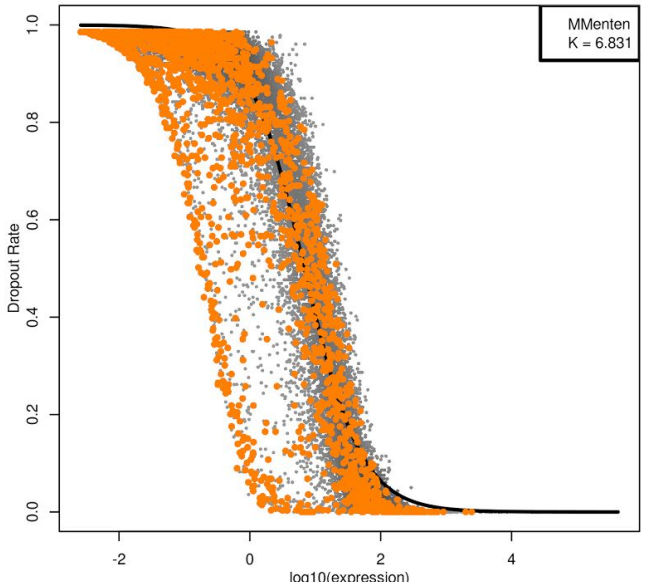
Paralogs



Duplication node: Mus musculus
Left shifted by 0.66 ($p < 10^{-40}$)

multimapping reads =
under counting

Short Genes

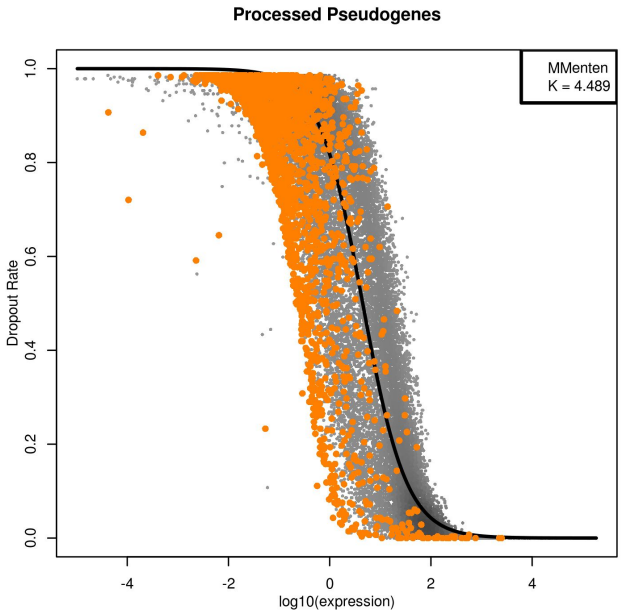


CDS < 300 n.t.
Left shifted by 0.21 ($p < 10^{-45}$)

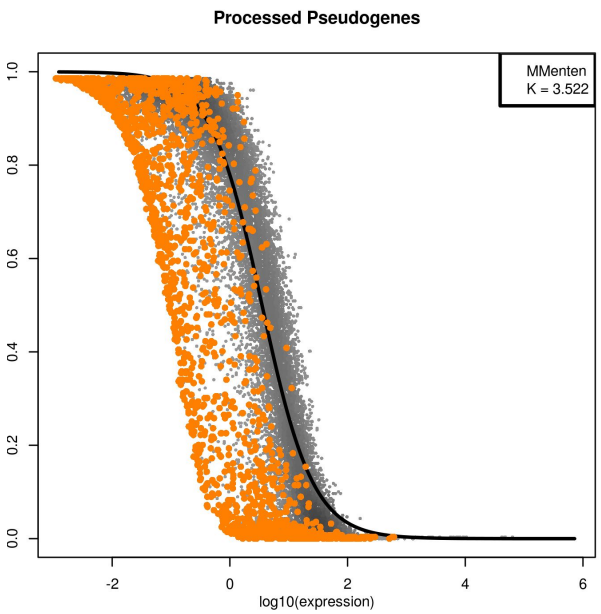
fewer unique fragments =
fewer unique reads

Tophat2 maps more reads to processed pseudogenes

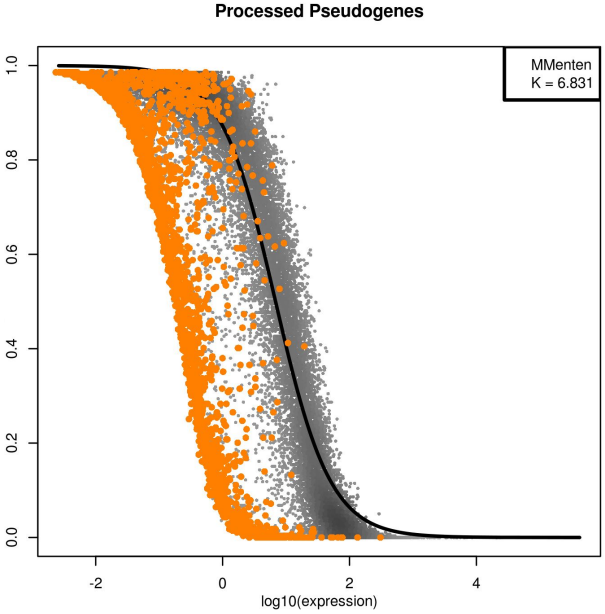
Kallisto



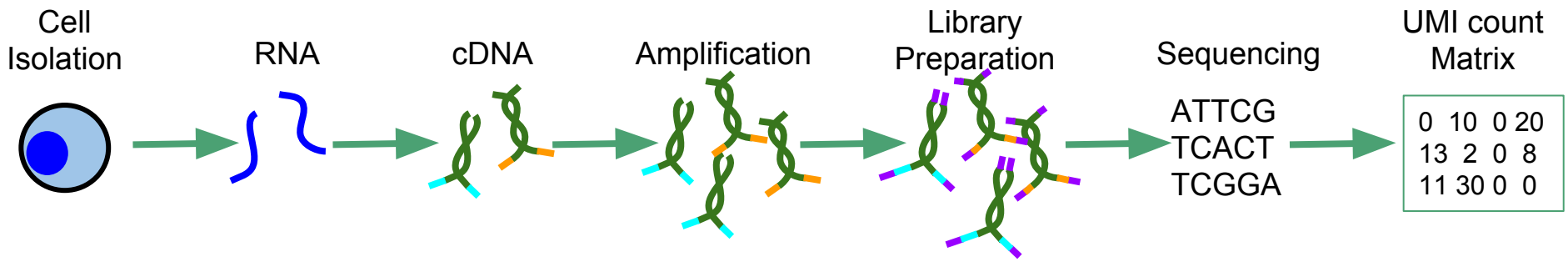
Tophat2



STAR



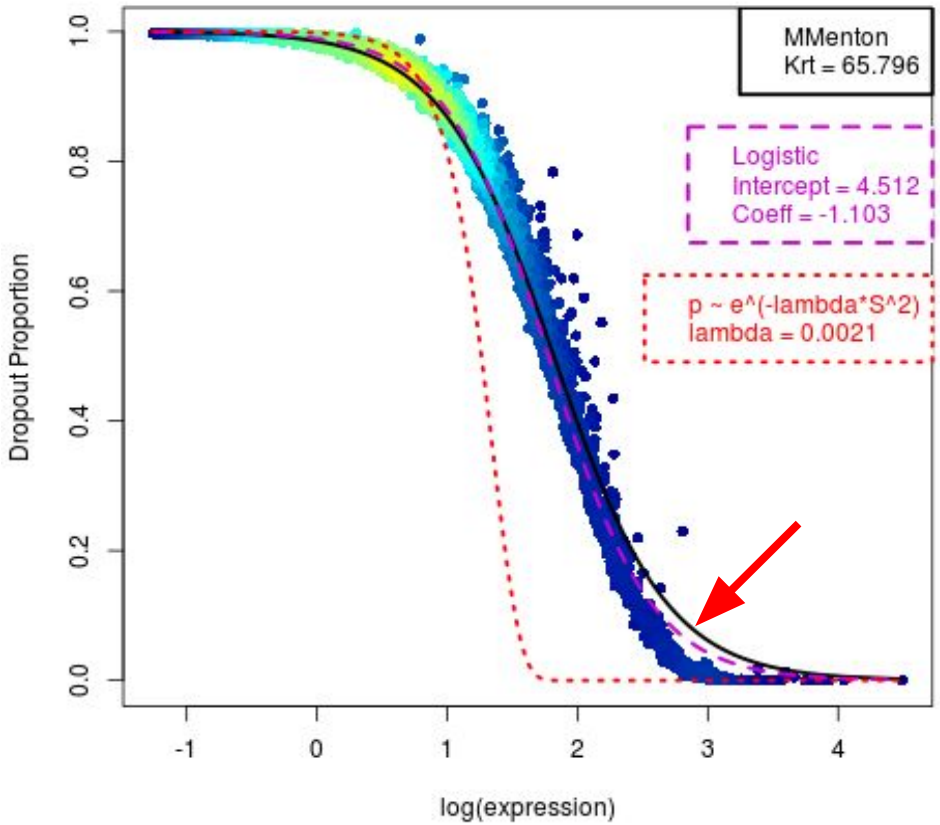
Unique Molecular Identifiers (UMIs)



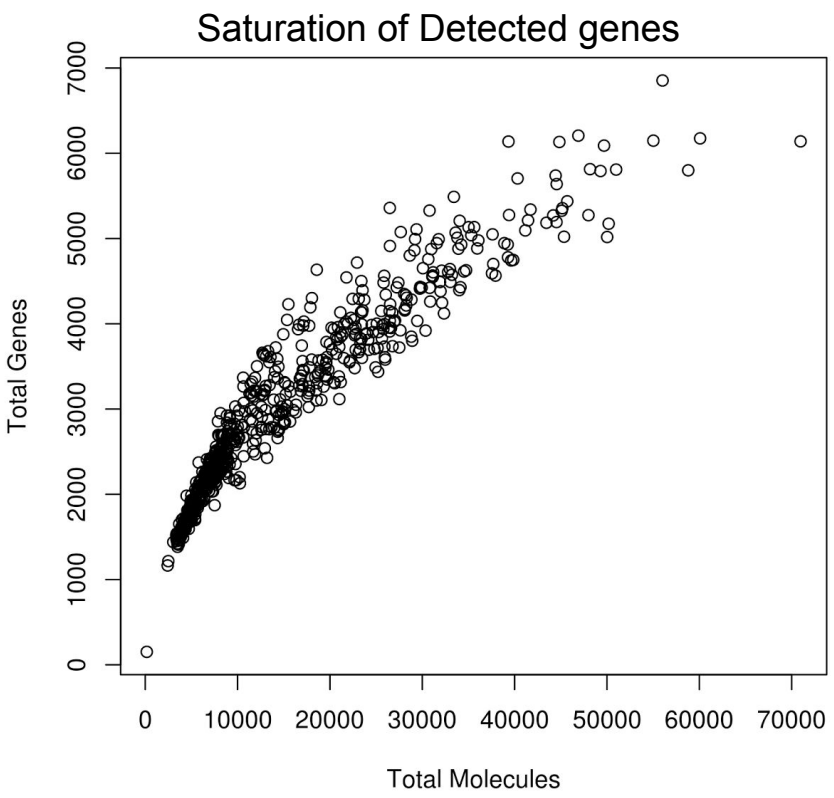
Enables:

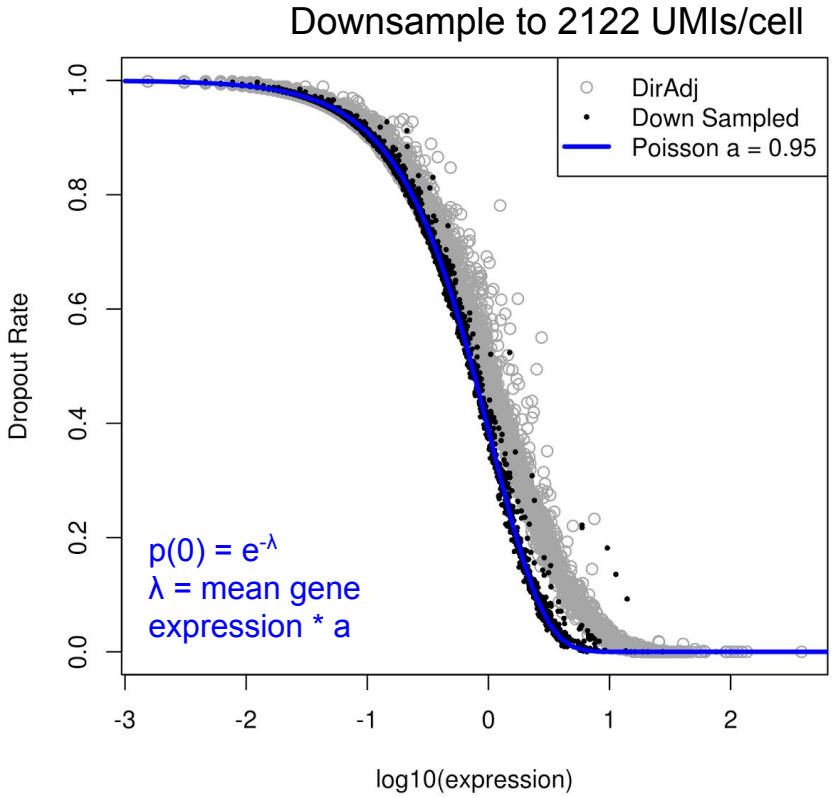
- Correction for PCR duplicates (amplification noise)

None of the proposed models fit corrected UMIs



Cell-specific detection rates obscure true relationship



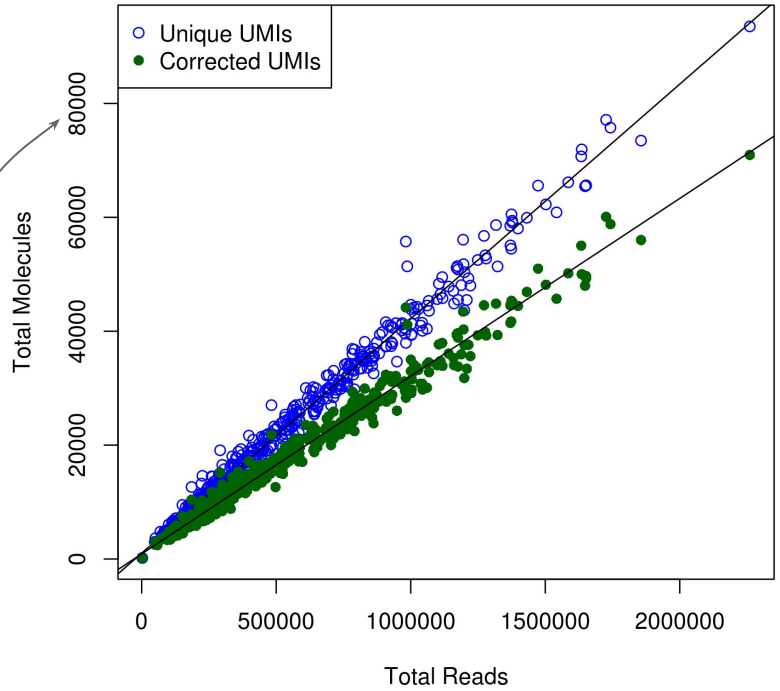


The PoissonUMIs Model

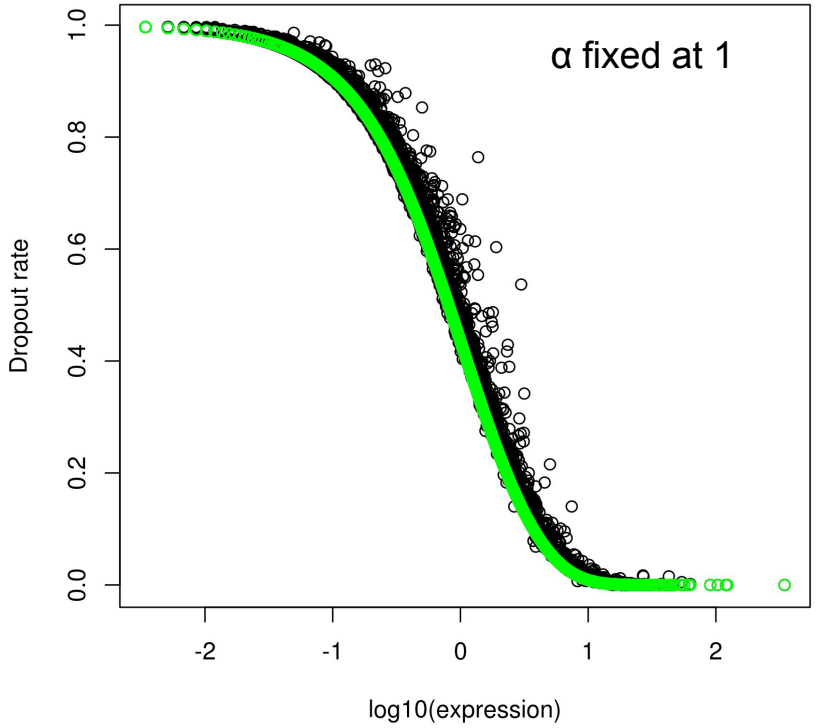
$M_{ij} \sim \text{Poisson}(\lambda)$
$\lambda = m_i * m_j * total * \alpha$

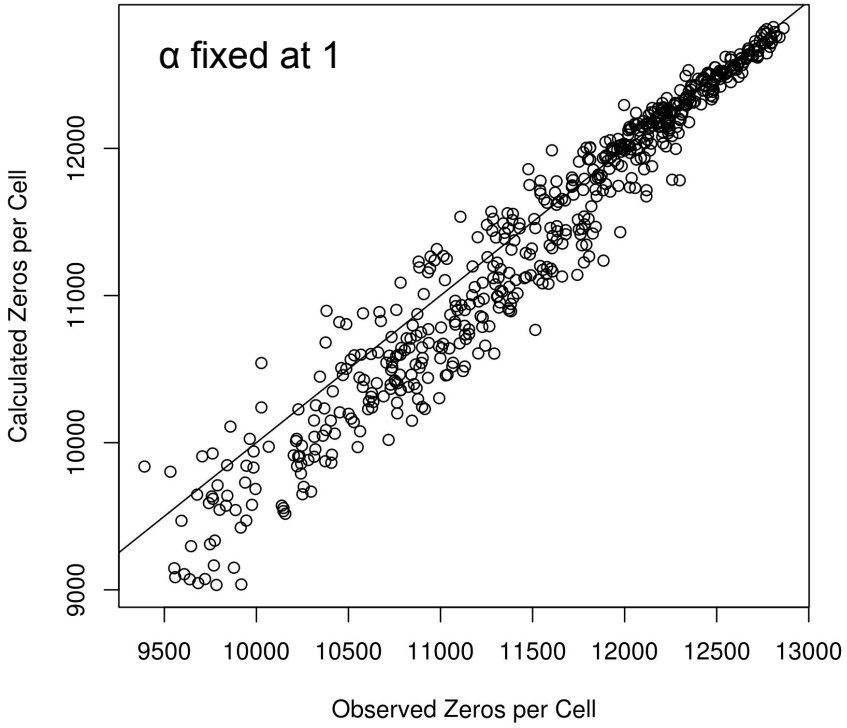
M_{ij} = Molecules of gene j in cell i
m_i = proportion of molecules in cell i
m_j = proportion of molecules for gene j
total = total detected molecules
α = scaling factor

Account for different counting methods

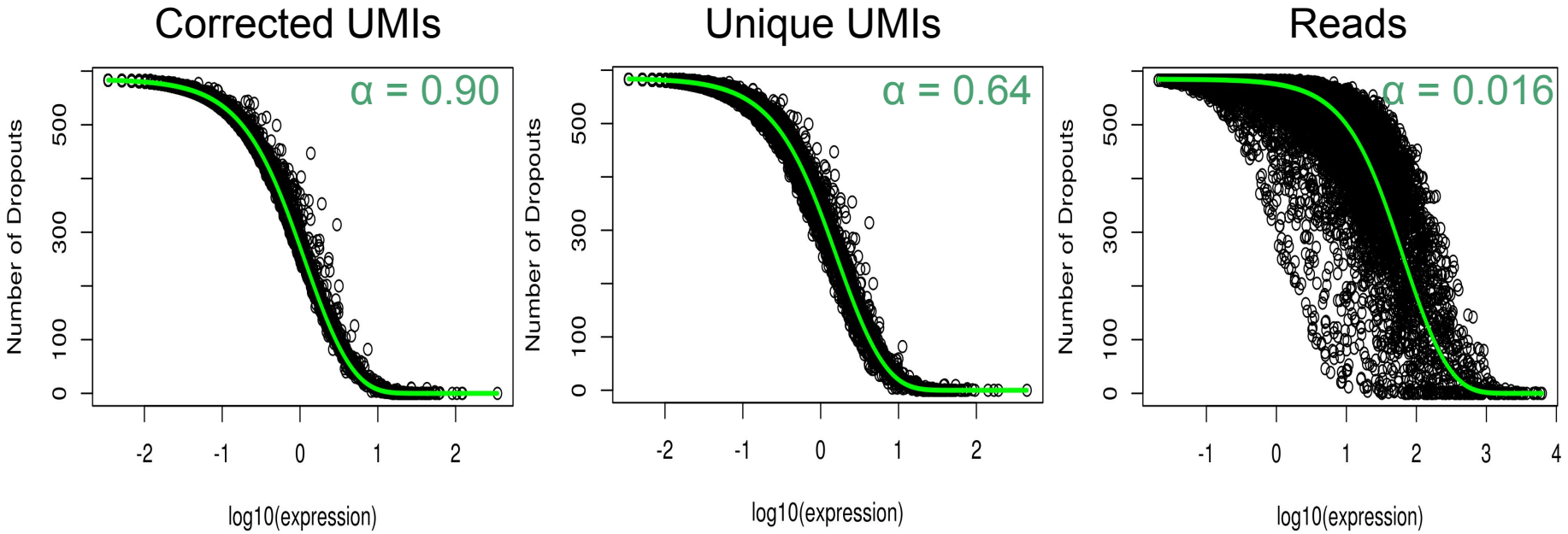


Poisson model accounting for differences in read depth



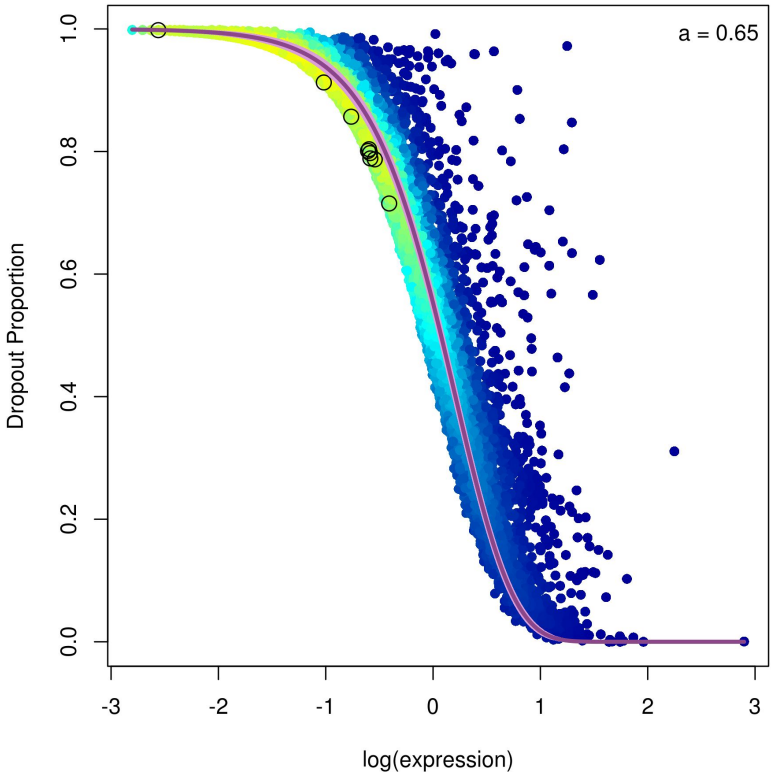


Fitted alpha reflects quantification method

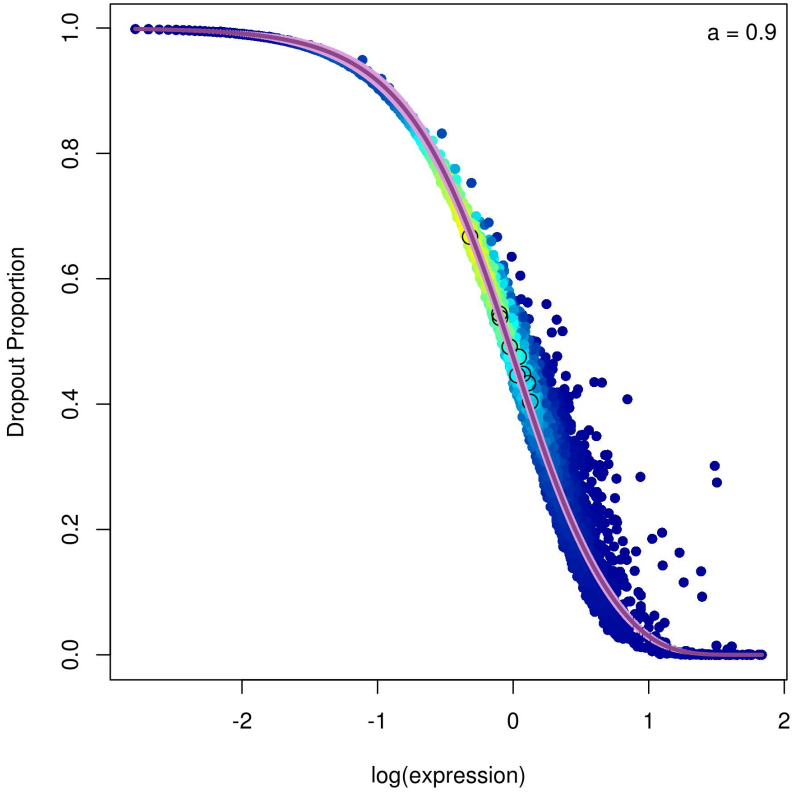


Fitting the model to other UMI datasets

Linnarsson α = 0.65

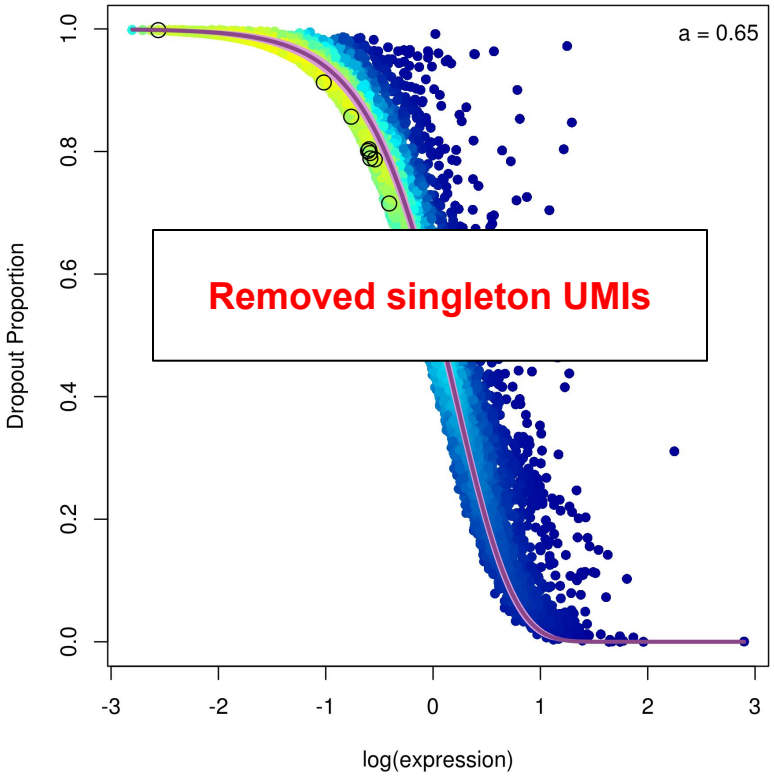


Kirschner α = 0.90

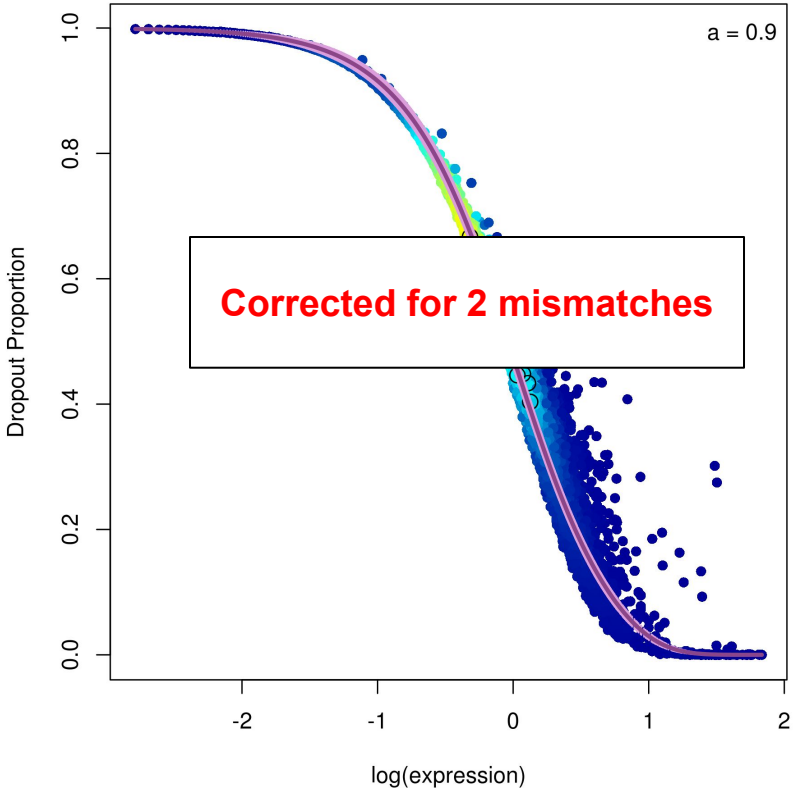


Fitting the model to other UMI datasets

Linnarsson α = 0.65

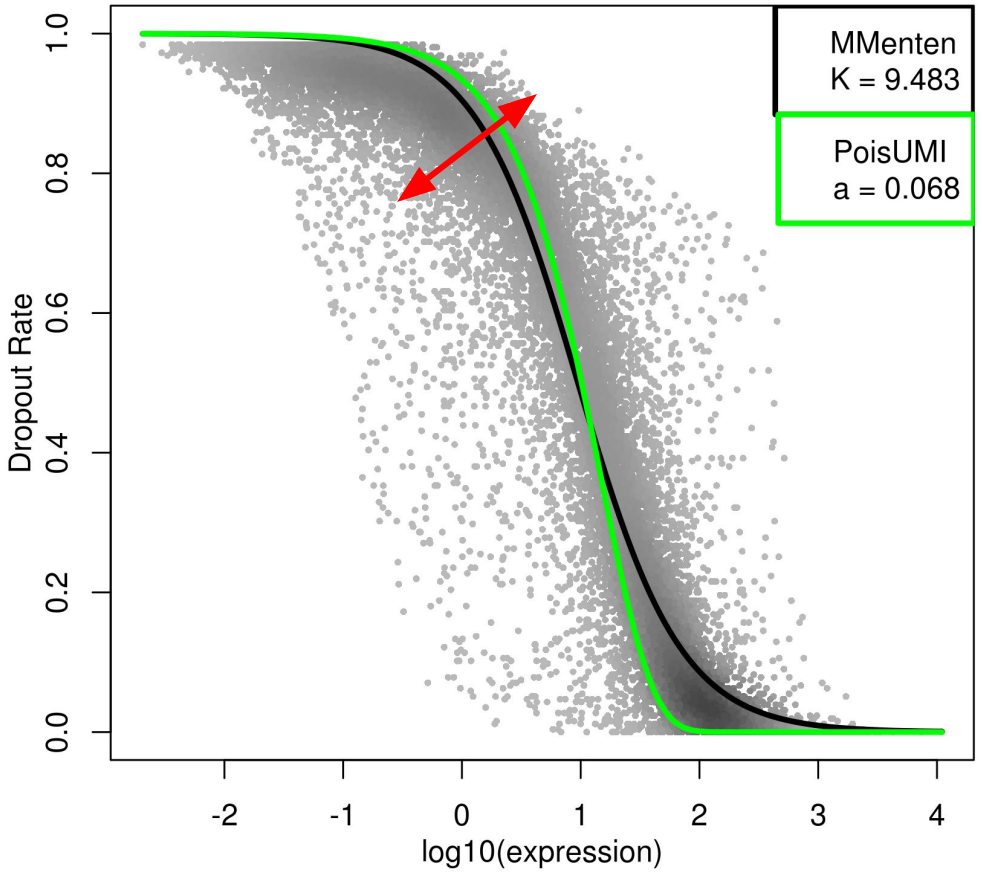


Kirschner α = 0.90



Summary

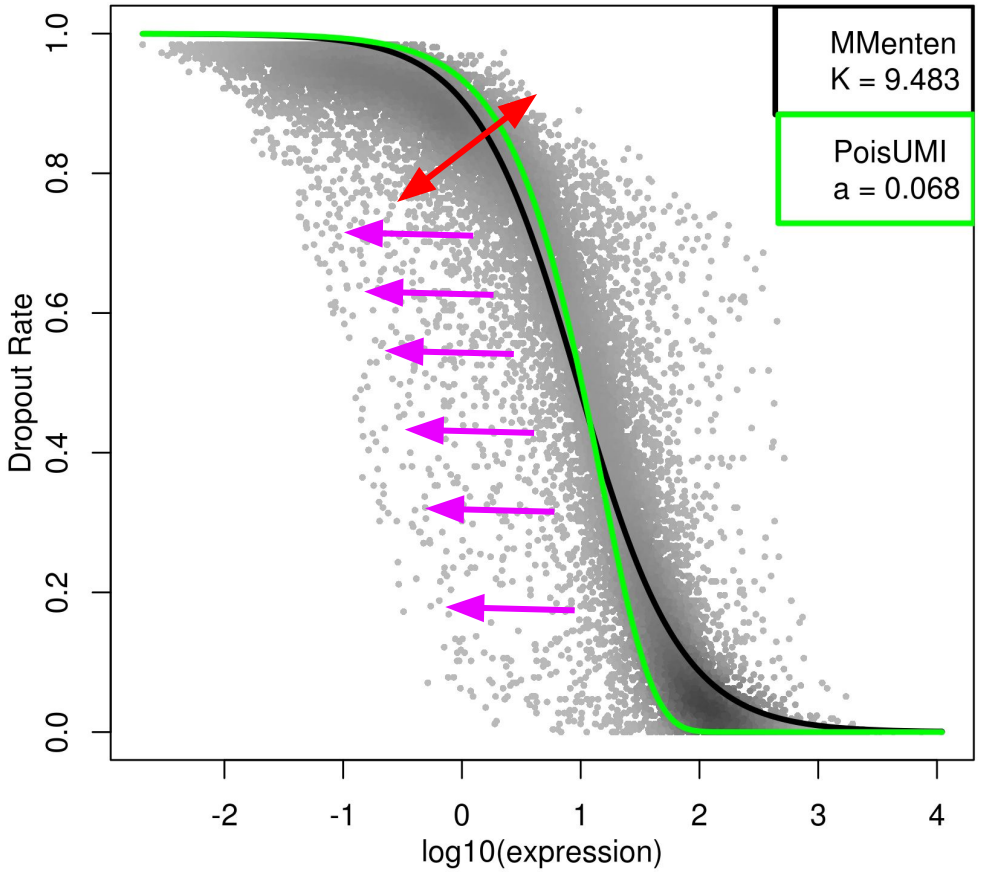
Amplification noise



Summary

Amplification noise

Mismapping / Miscounting

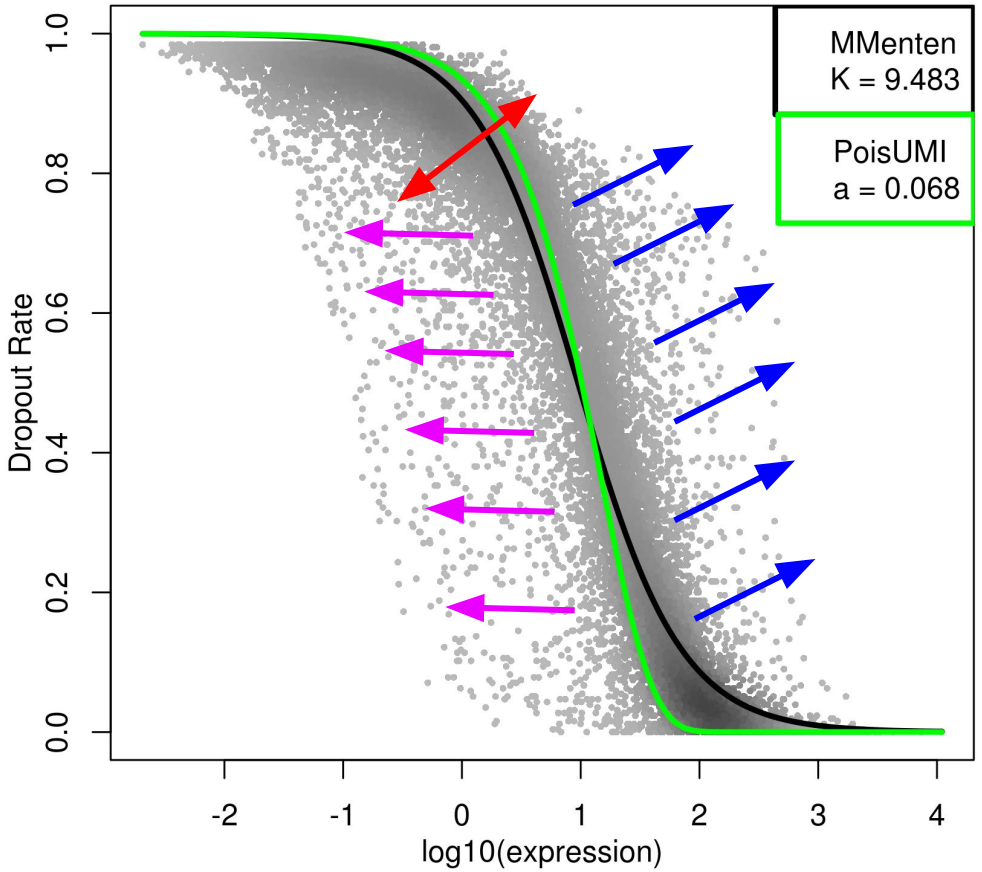


Summary

Amplification noise

Mismapping / Miscounting

Differential Expression



Acknowledgements

Wellcome Trust Sanger Institute
Martin Hemberg
Vladimir Kiselev

Availability

M3Drop : https://github.com/tallulandrews/M3Drop
PoissonUMIs: https://github.com/tallulandrews/PoissonUMIs

EMBL Rome
Christophe Lancrin
Isabelle Bergiers