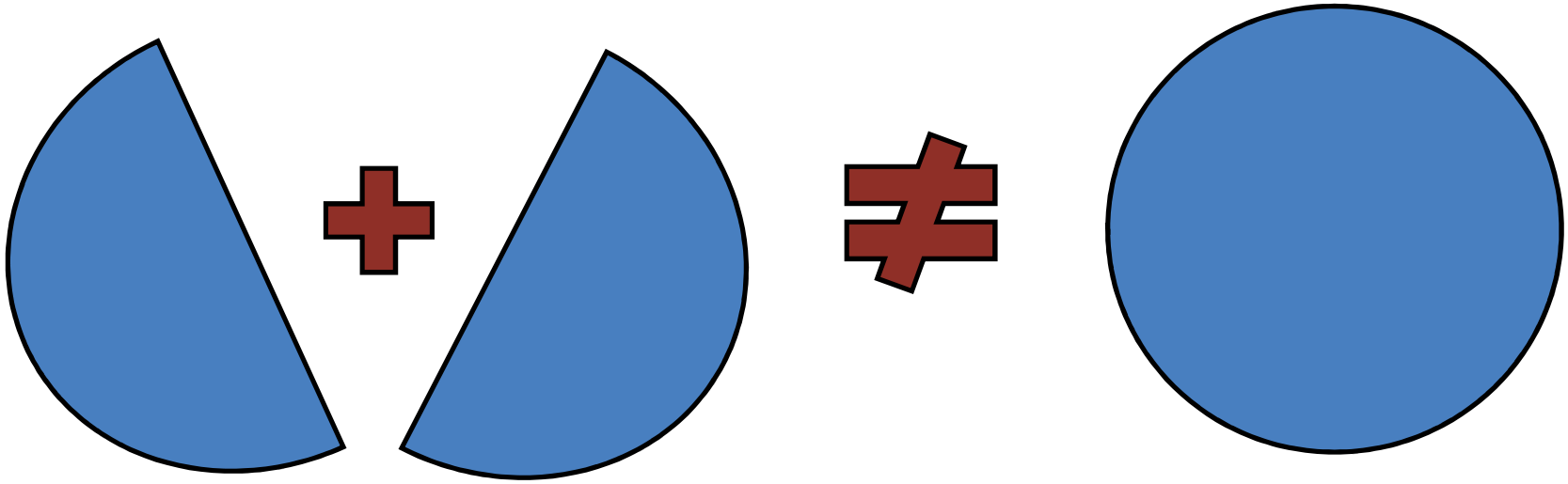
UNIVERSITY OF CAMBRIDGE

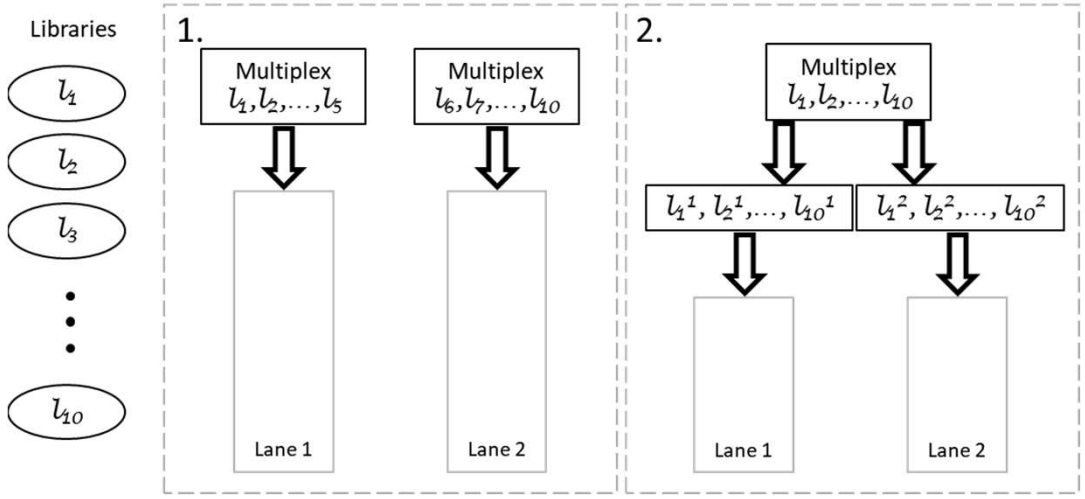
wellcome-MRC cambridge stem cell institute

The sum of two halves may be different from the whole. Effects of splitting sequencing samples across lanes.



Eleanor C. Williams*, Ruben Chazarra-Gil*, Arash Shahsavari*, Irina Mohorianu@
Core Bioinformatics Group, Wellcome-MRC Cambridge Stem Cell Institute
Contact: ecw63@cam.ac.uk

Aims



- Simulating across-lane sample splitting to illustrate variability introduced through sequencing design
- For bulk experiments, differences are observed in differential expression and enrichment analysis
- At the single cell level, we see changes in the identification of cell subpopulations
- We identify changes in biological interpretation when splitting across lanes is performed

Statistical Background

Sample 1
'Ground Truth'

AATC	10
TACT	7
GGAT	5
ATGC	5
GCGT	2
CCGT	1

Sequencing Depth: 30

Subsample at 50%

Subsample 1

AATC	5
TACT	3
GGAT	2
ATGC	5
GCGT	0
CCGT	0

Subsample 2

AATC	7
TACT	5
GGAT	1
ATGC	1
GCGT	0
CCGT	1

Subsample 3

AATC	4
TACT	6
GGAT	1
ATGC	3
GCGT	1
CCGT	0

Add 2
subsamples

SS1 + SS2

AATC	12
TACT	8
GGAT	3
ATGC	6
GCGT	0
CCGT	1

SS1 + SS3

AATC	9
TACT	9
GGAT	3
ATGC	9
GCGT	1
CCGT	0

SS2 + SS3

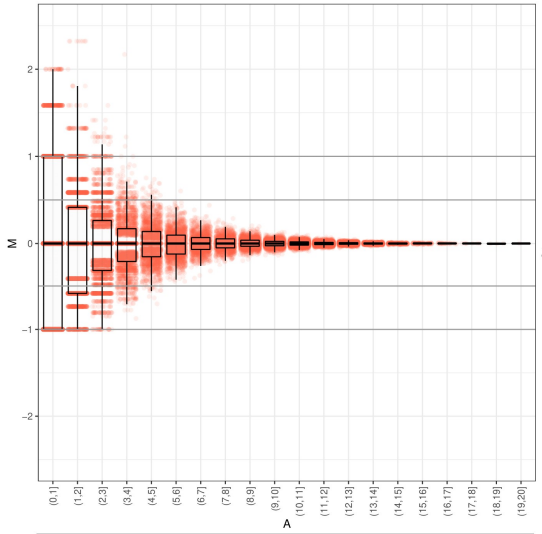
AATC	11
TACT	11
GGAT	2
ATGC	4
GCGT	1
CCGT	1

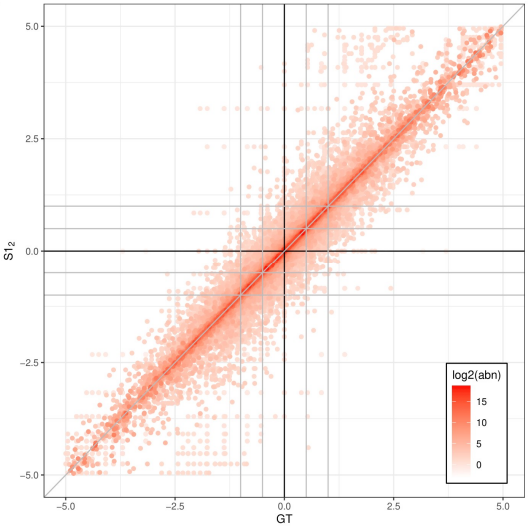
mRNAseq Case Study

Yang et al, Cell Systems, May 22;8(5):427-445.e10. (2019)
Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency

GT v $S1_2$

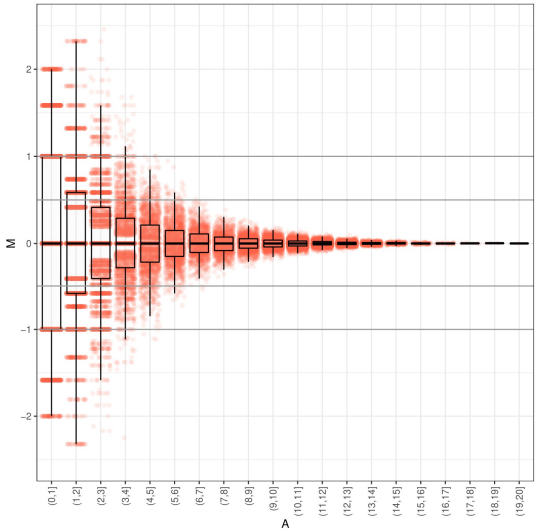
Ground truth vs simulated k=2 sample

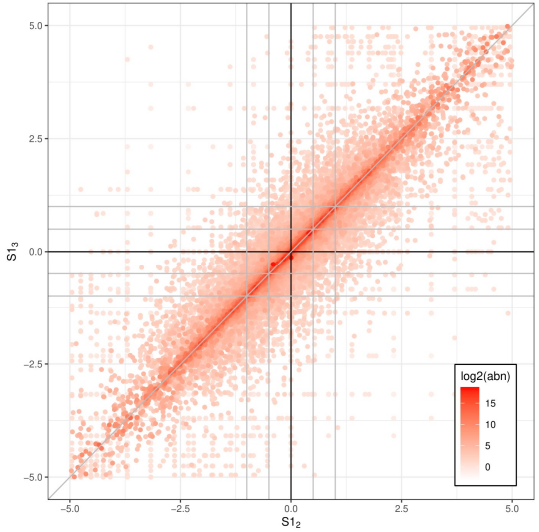


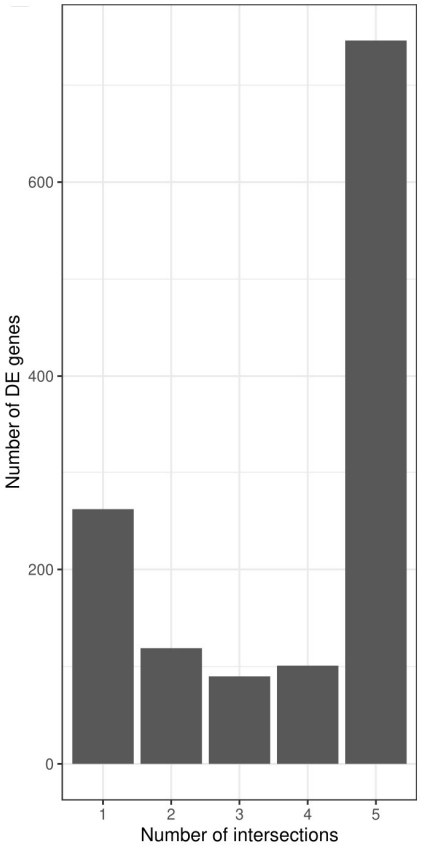


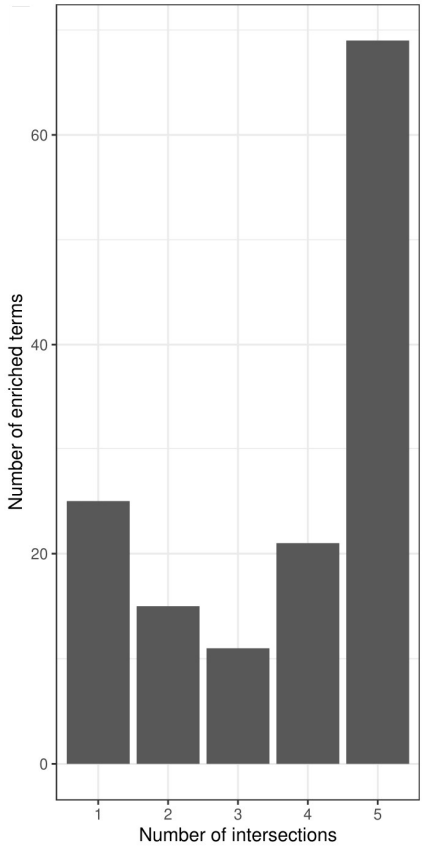
$S1_2$ v $S1_3$

Simulated k=2 sample vs simulated k=3 sample



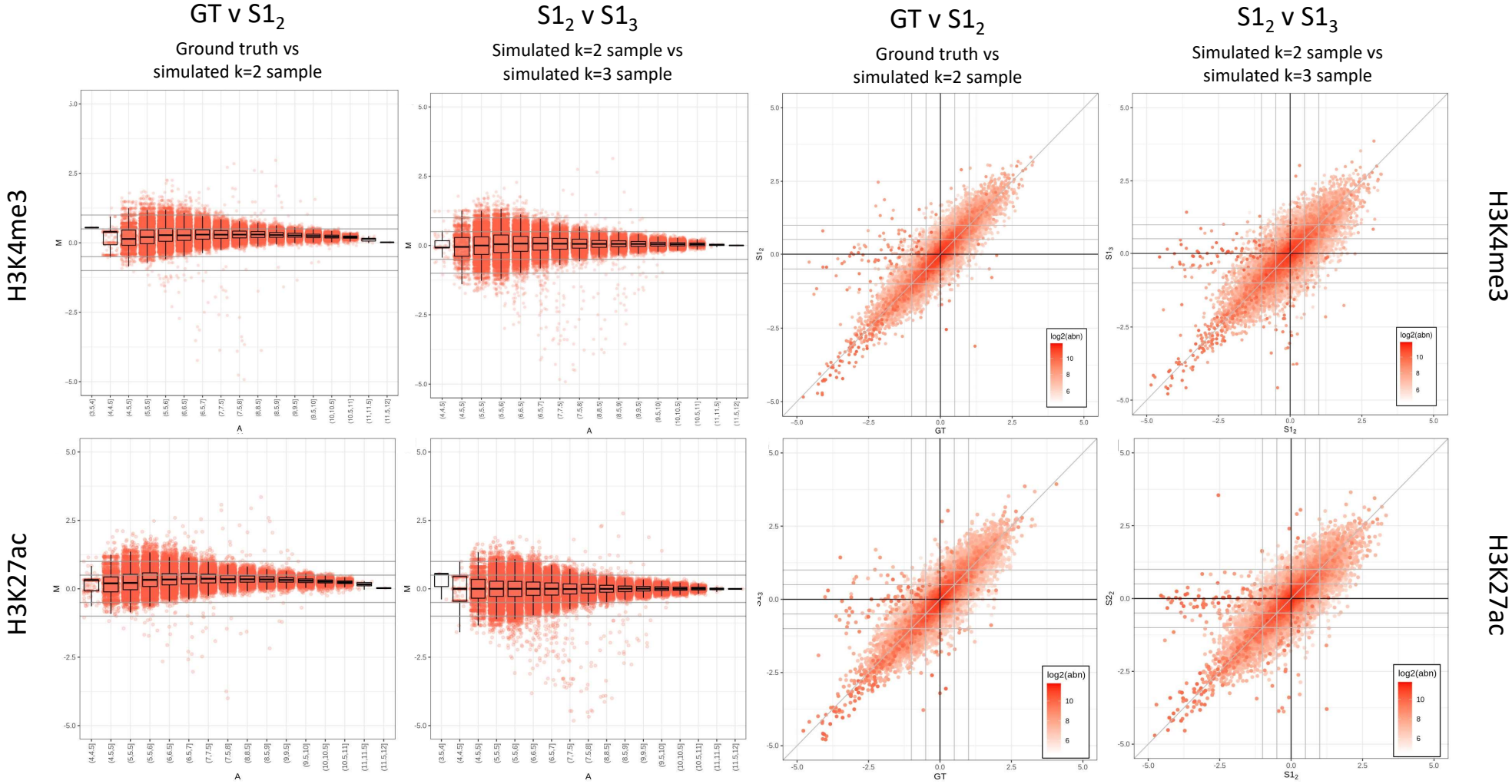






ChIPseq Case Study

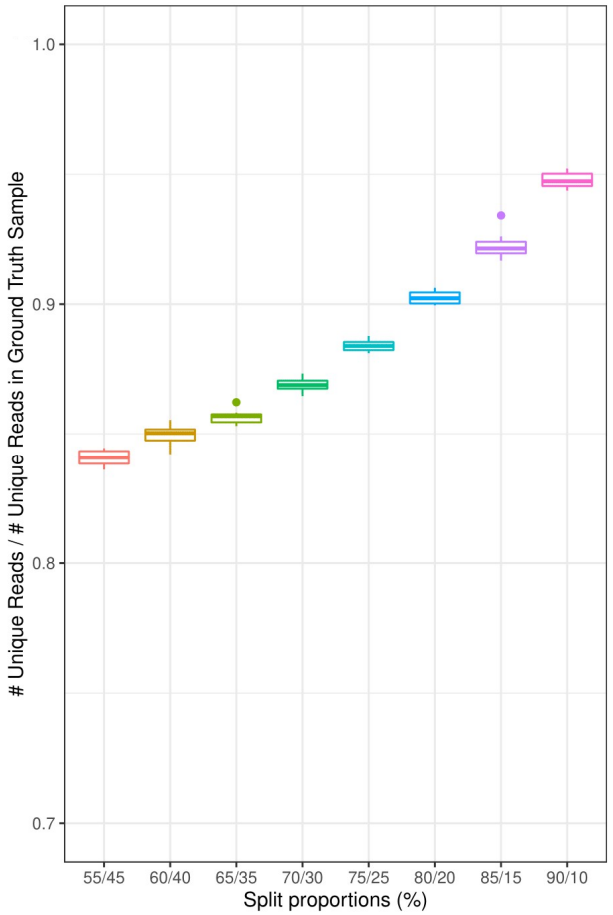
Yang et al, Cell Systems, May 22;8(5):427-445.e10. (2019)
Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency



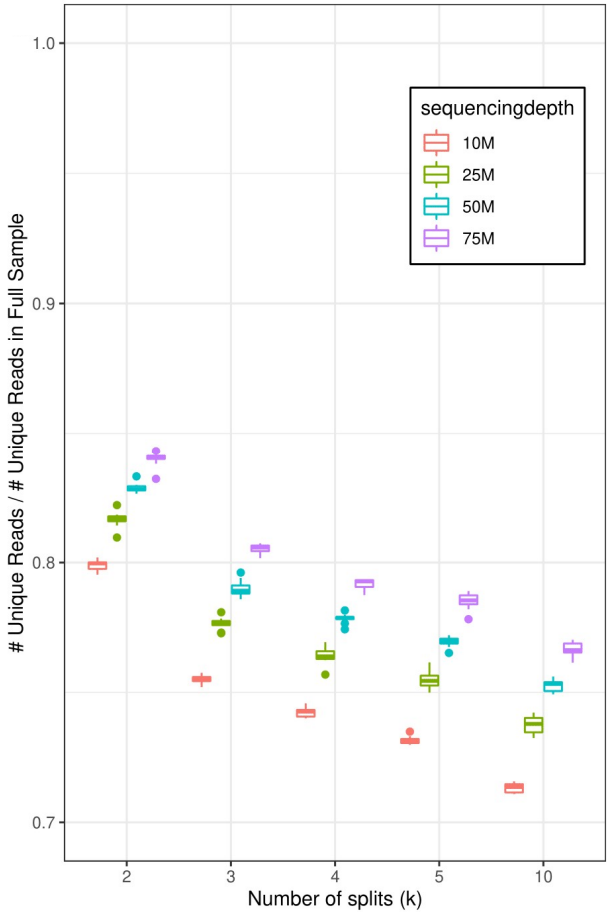
mRNAseq Case Study

Yang et al, Cell Systems, May 22;8(5):427-445.e10. (2019)
Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency

Varying split proportion



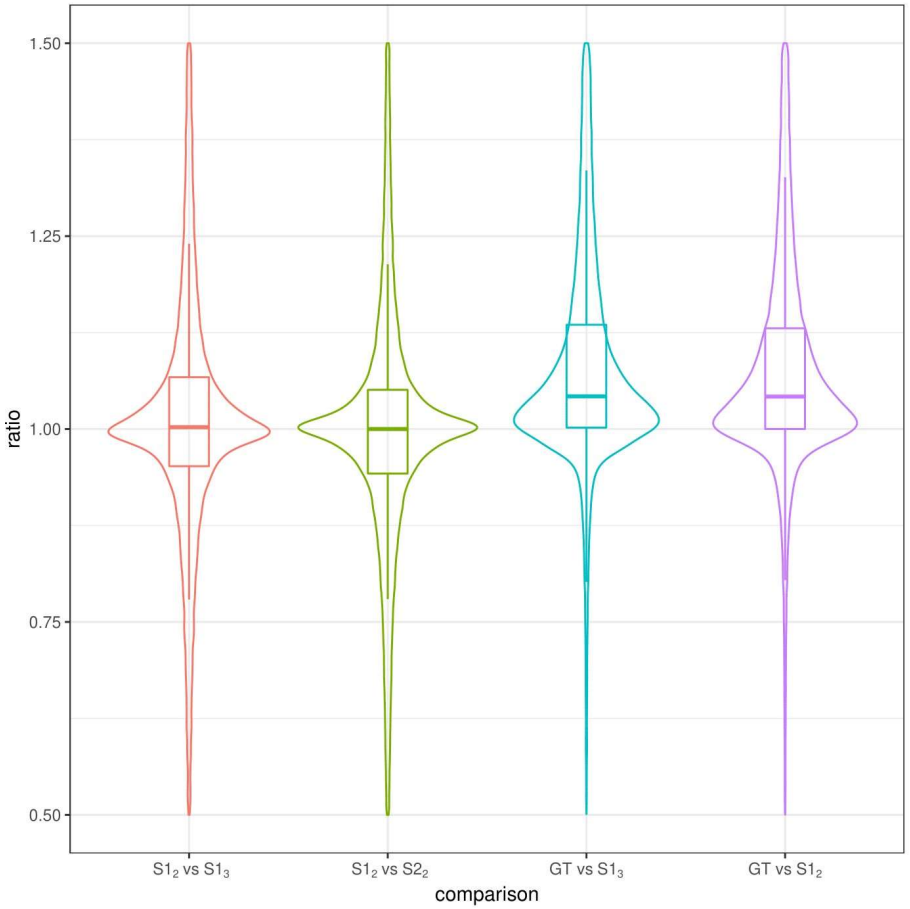
Varying number of splits and sequencing depth



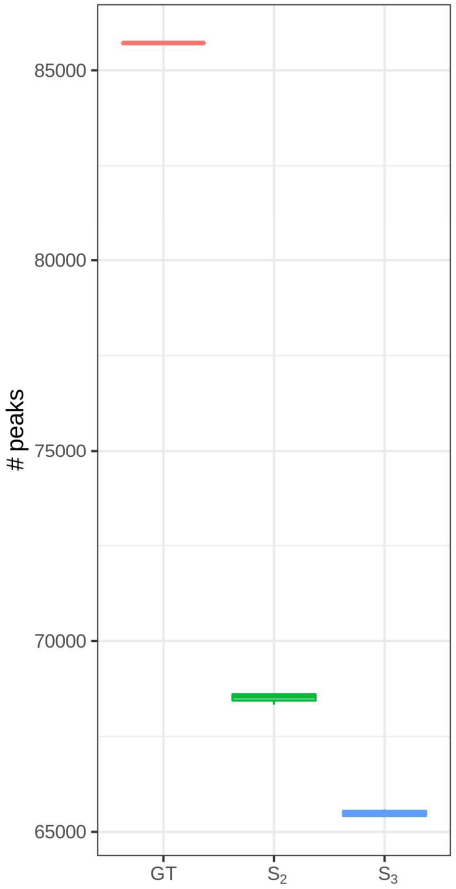
ChIPseq Case Study

Yang et al, Cell Systems, May 22;8(5):427-445.e10. (2019)
Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency

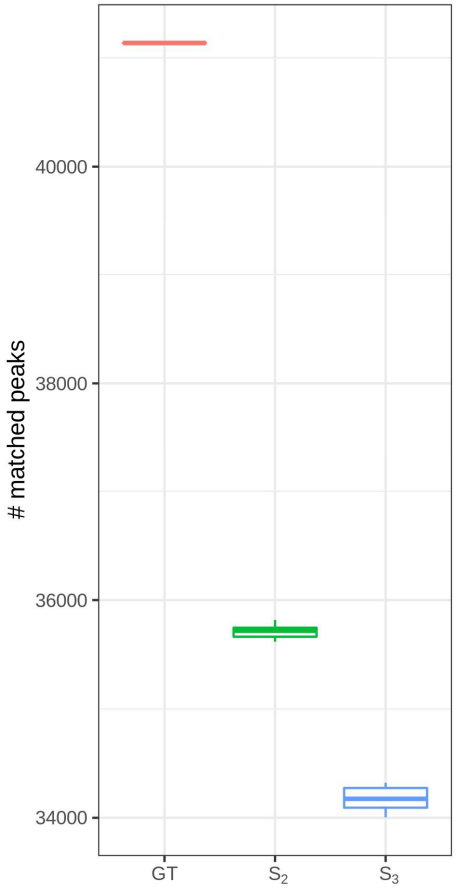
Ratio of peak lengths



Number of peaks



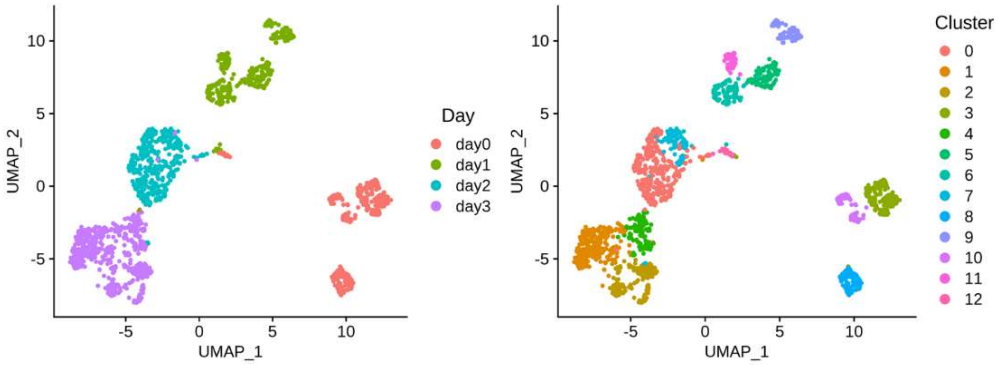
Number of peaks matched between time points

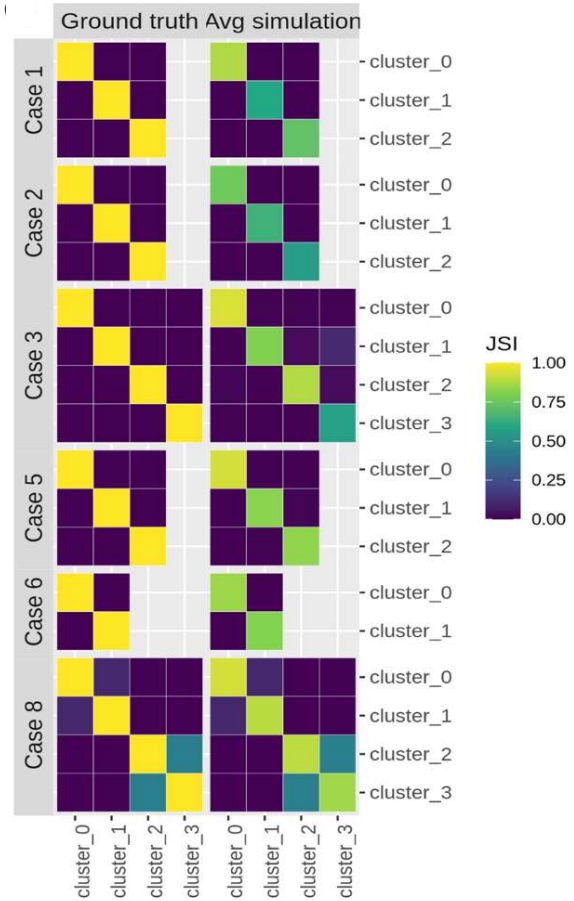


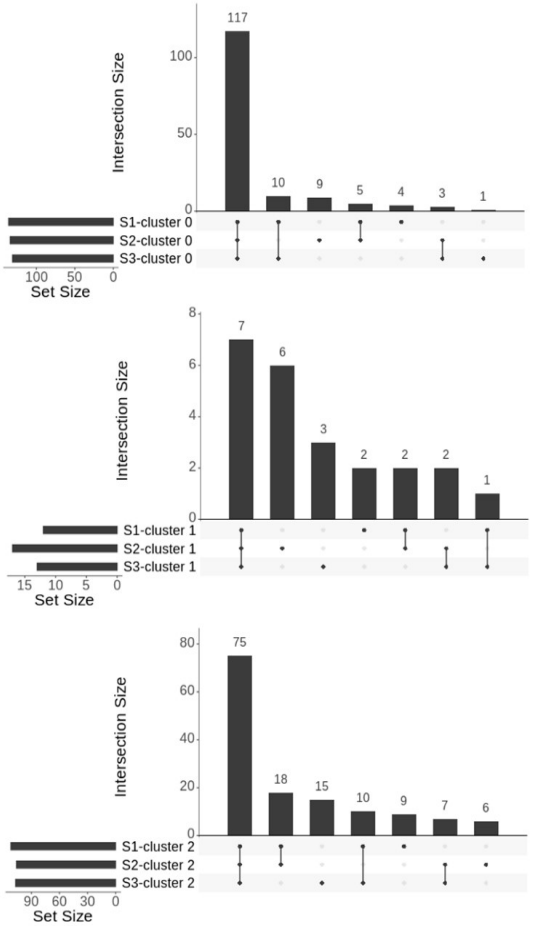
Smart-Seq2 Case Study

Cuomo et al; Nature Communications volume 11, Article number: 810 (2020)
Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression

	Study case	N cells	Donor	N cells (donor)	Time point	N cells (time point)	Cluster	N cells (cluster)
1	Case 1	105	hayt	105	day2	105	0	105
2	Case 2	106	pahc	106	day3	106	1	66
							4	40
3	Case 3	168	melw	94	day0	168	3	168
			qunz	74				
4	Case 5	168	hayt	168	day1	61	9	61
					day3	107	1	107
5	Case 6	95	melw	47	day1	95	5	45
			vils	48			6	50
6	Case 8	217	melw	95	day0	95	3	95
			naah	122	day3	122	2	122



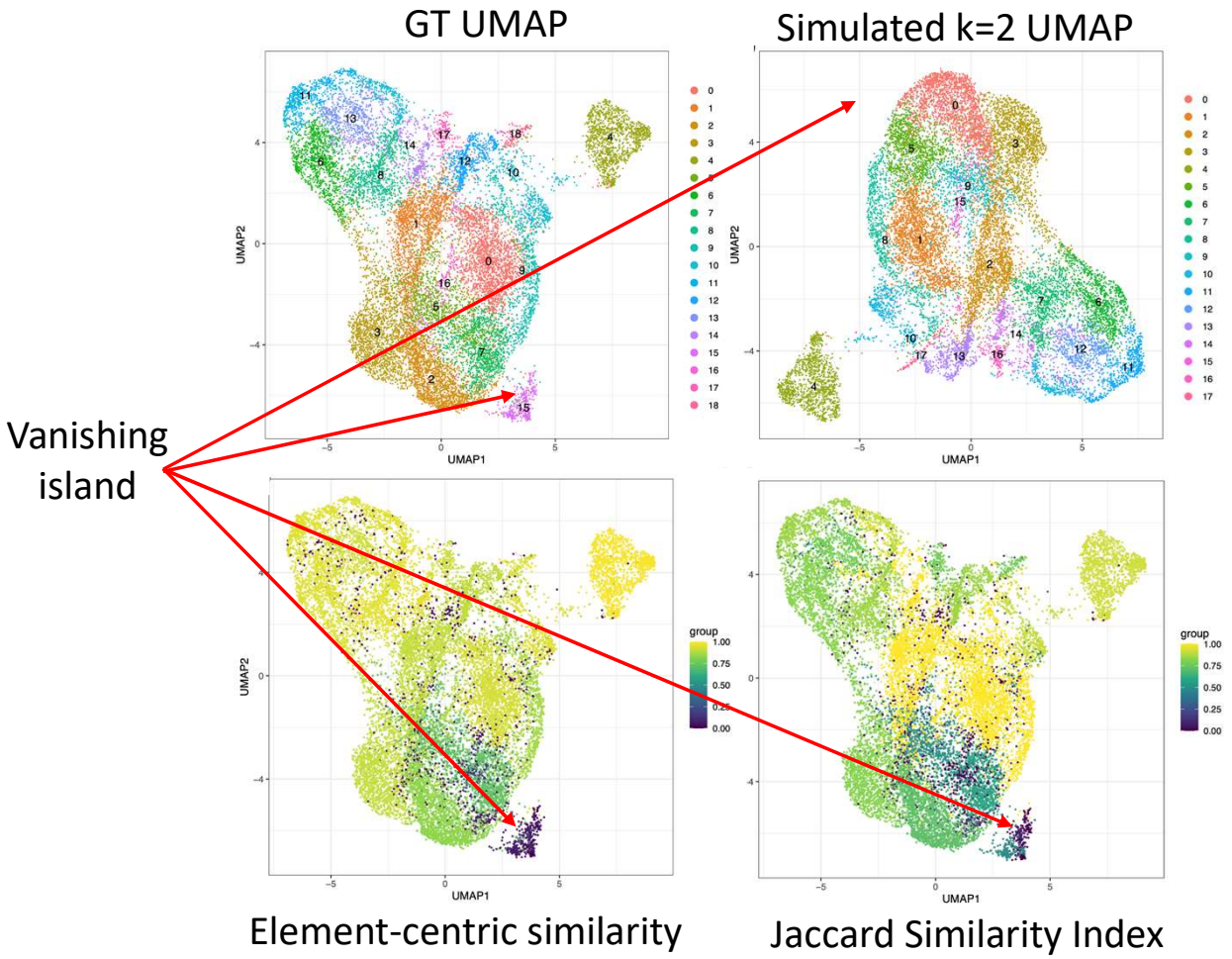




10x Case Study

Mende et al; bioRxiv (2020)

Quantitative and molecular differences distinguish adult human medullary and extramedullary haematopoietic stem and progenitor cell landscapes



Conclusions

- Splitting across lanes typically reduced read diversity, leading to over-representation for higher abundance fragments and under-representation or complete loss of lower abundance fragment
- While these changes may be difficult to see pre-alignment, we see knock-on effects in downstream analysis

- In bulk mRNAseq experiments this could lead to false positives in differential expression
- In bulk ChIPseq experiments, changes in peak calling and properties can be observed
- In single cell experiments, the topography of the UMAP can change, along with inferred cell type identities

- Ideally, don’t split across lanes but if it can’t be avoided then consistency is key

Acknowledgements

Ruben Chazarra-Gil

Arash Shahsavari

Irina Mohorianu

Core Bioinformatics Group

Ilias Moutospoulos

Alex Calderwood

Yu Zhang

Andi Munteanu

Chris Ellis

Antonella Santoro

Ingo Ringshausen

Ludovic Vallier

Rory Stark

Katarzyna Kania

Preprint on bioRxiv

UNIVERSITY OF CAMBRIDGE

wellcome-MRC cambridge stem cell institute