Cancer phylogenetics using single-cell RNA-seq data

Jiří Moravec[1], Rob Lanfear[2], David Spector[3], Sarah Diermeier[1] and Alex Gavryushkin[1]

1) University of Otago, Te Whare Wānanga o Otāgo, New Zealand

2) Australian National University

3) Cold Spring Harbor Laboratory (CSH)

scRNA-seq is exciting!

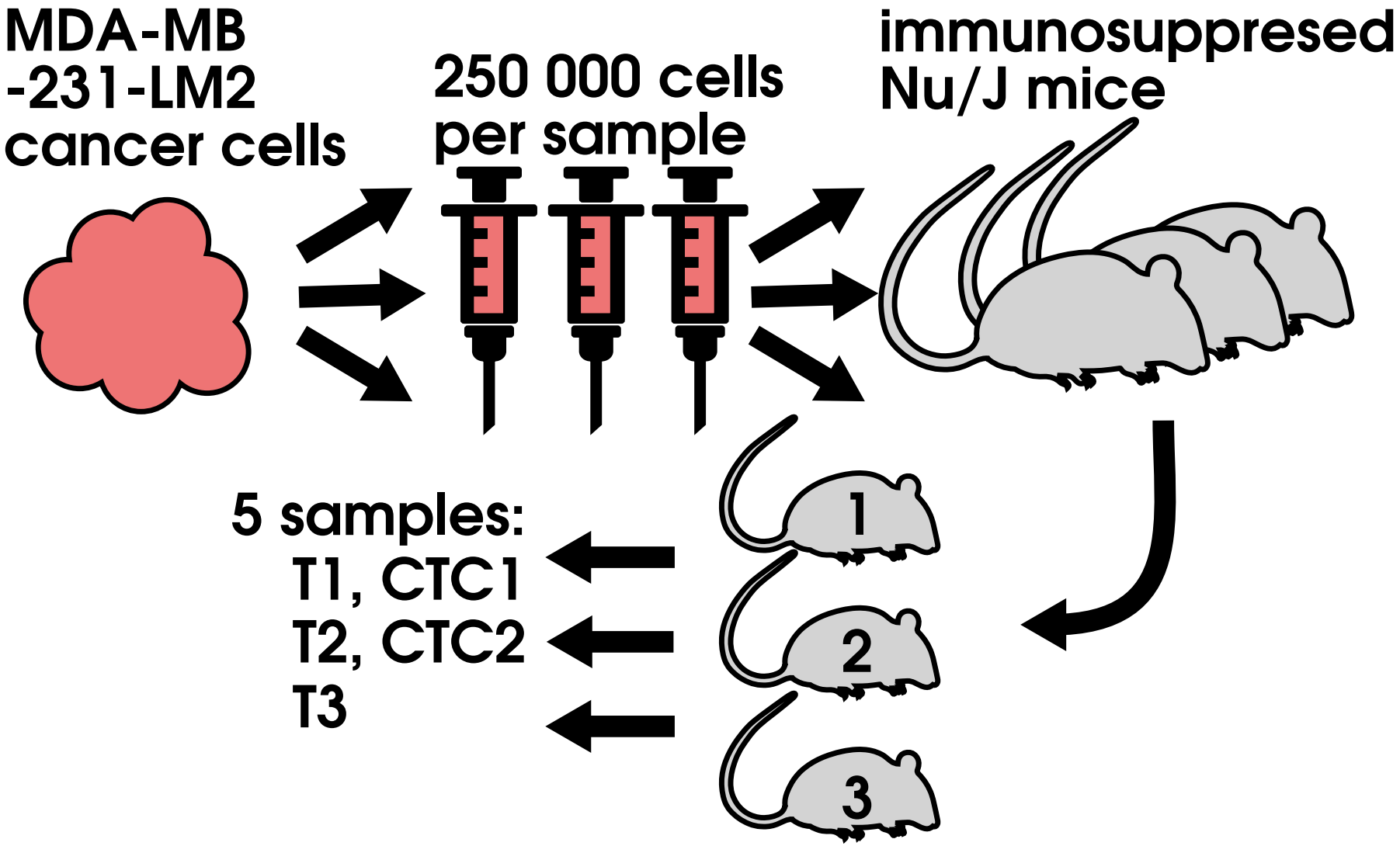
Like scDNA-seq:

- guranteed single origin of DNA
- we can detect SNVs

But better!

- expression levels are commonly used
- expression levels are influenced by epigenetics
- we can detect CNVs

Experiment design



Data:

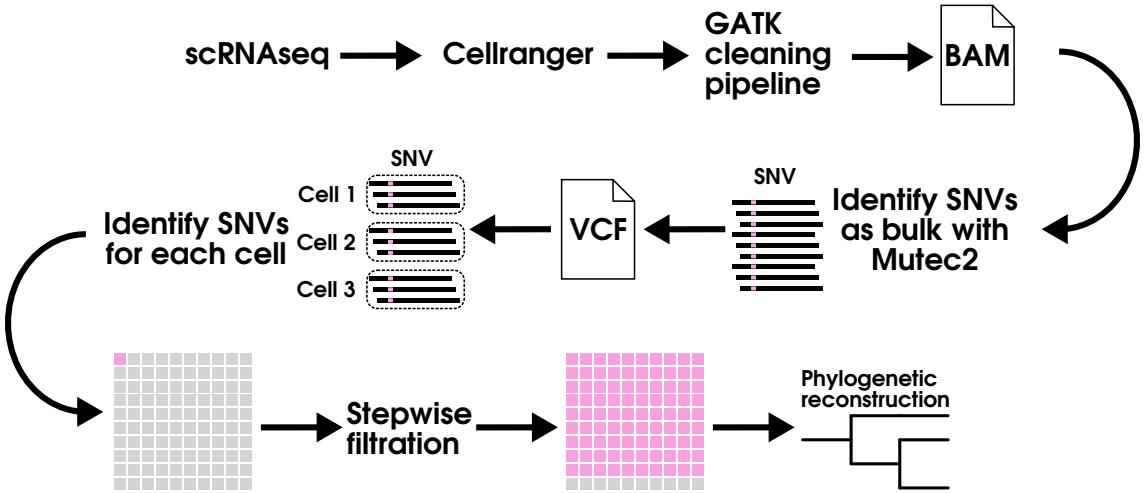
The expression levels for 43k genes varies greatly!

Sample	Cells	UMI	UMI/Cell	density
T1	713	428k	600	0.66 %
T2	2777	69k	25	0.052 %
T3	810	652k	805	1 %
CTC1	3108	129k	41	0.08 %
CTC2	8275	28k	3	0.005 %
total:	15683	1.3M	83	0.11 %

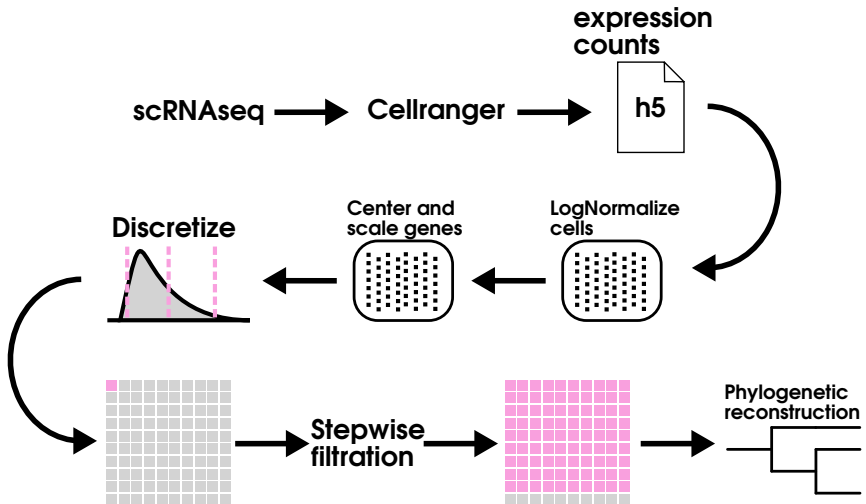
Missing data are integral part of scRNAseq problem!

Workflow

SNV:

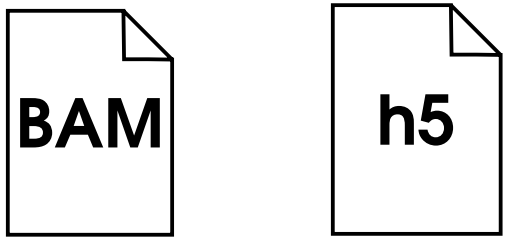


Expression:

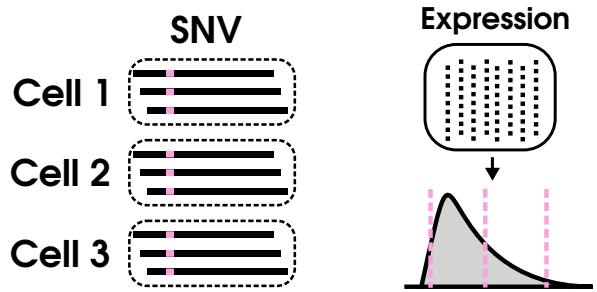


scRNAseq

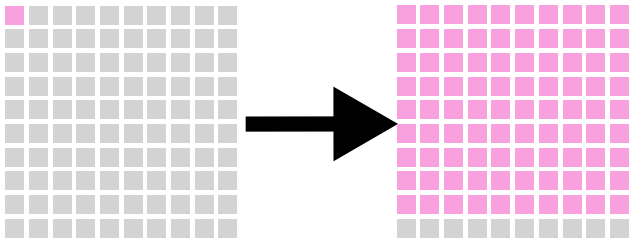
Mapping + demultiplexing



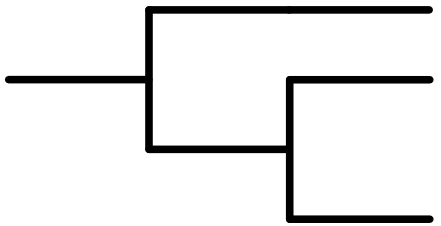
Transformation



Filtering



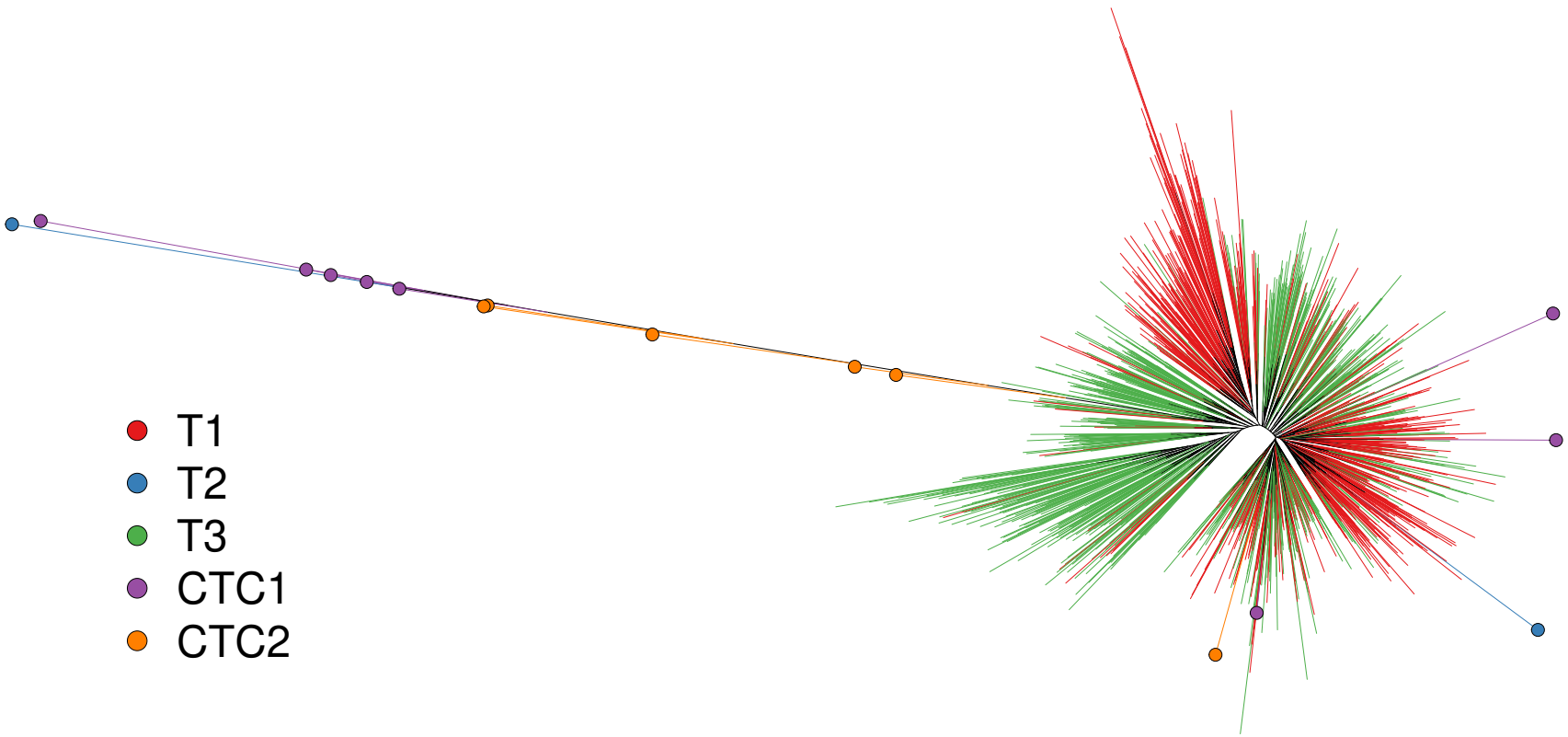
Phylogenetic reconstruction



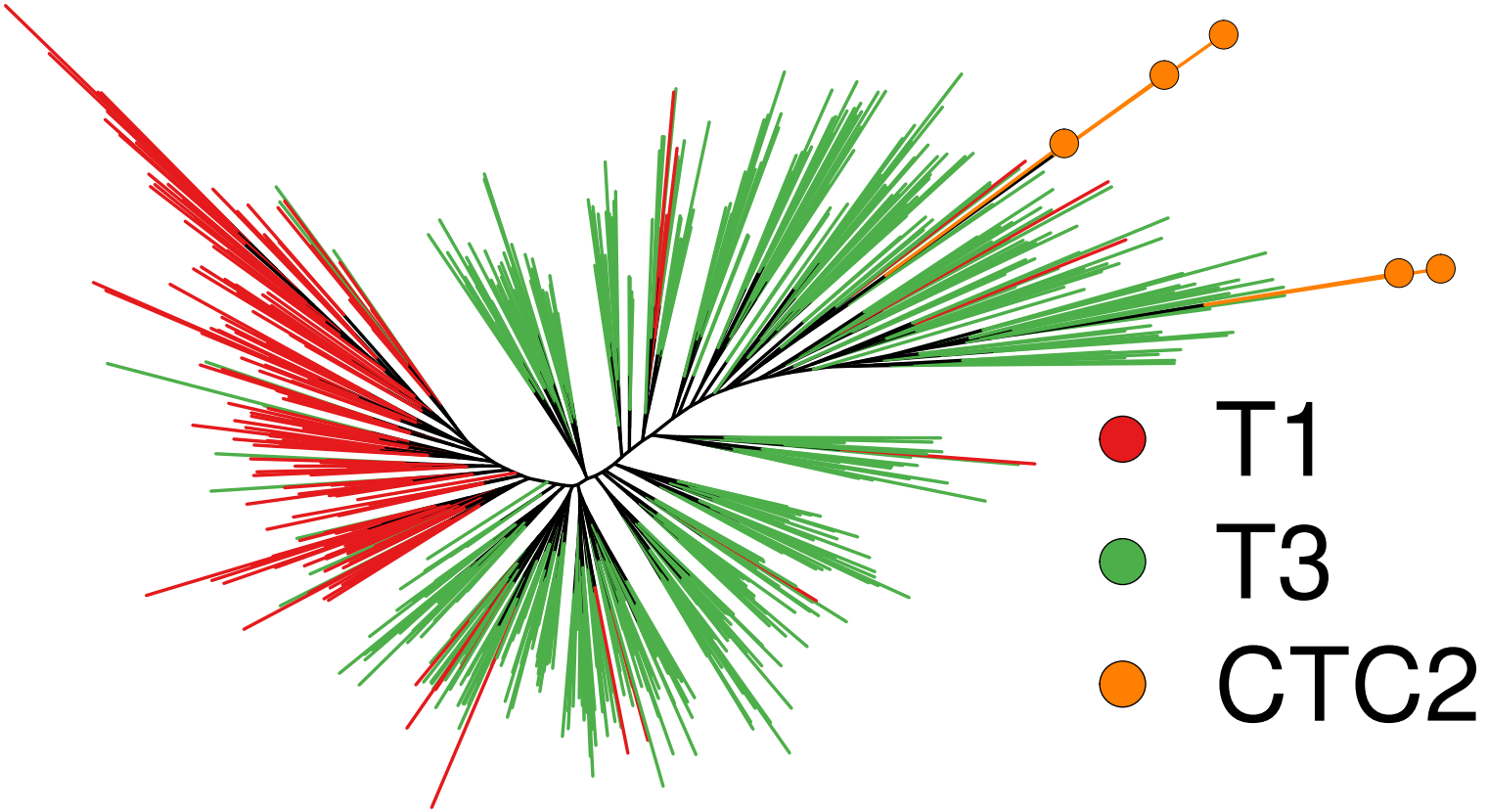
Trees

- Expression and SNV
- Maximum Likelihood method using IQtree
- Progressive filtering to remove missing data
- filtered datasets: 20%, 50% and 90% data density
- running time for IQtree on the 20% expr: 6 hours

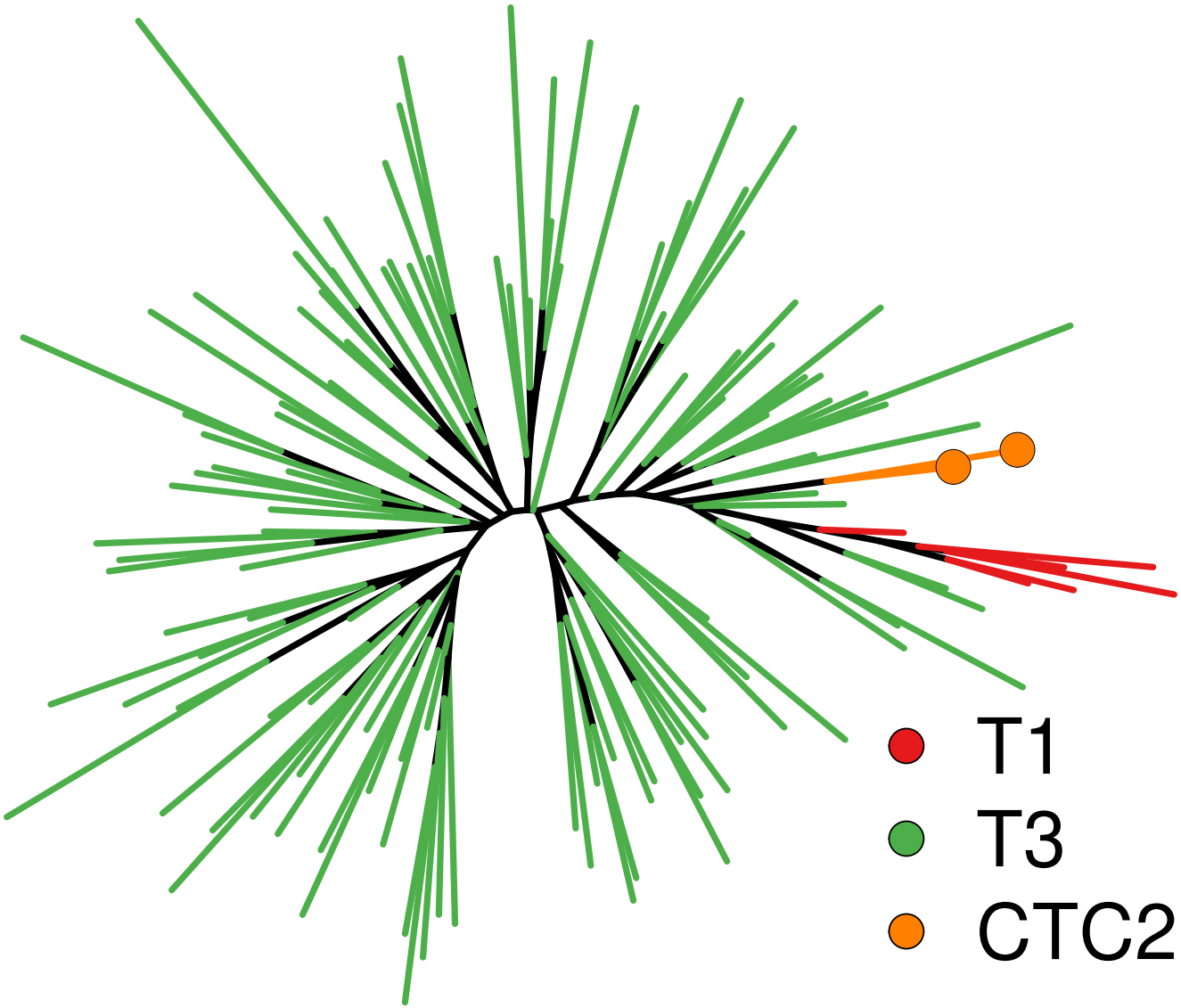
Expression 20% density



Expression 50% density

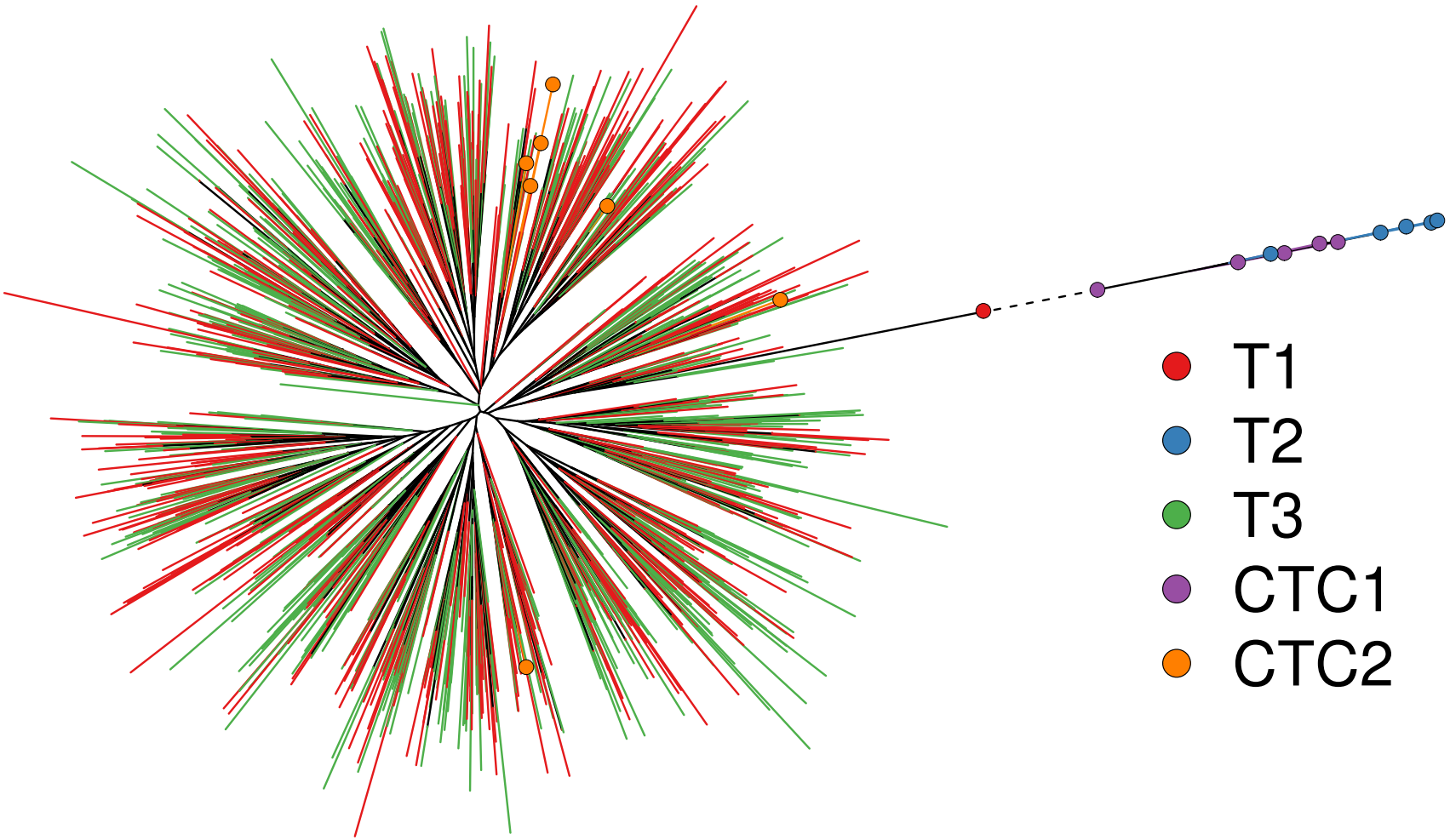


Expression 90% density

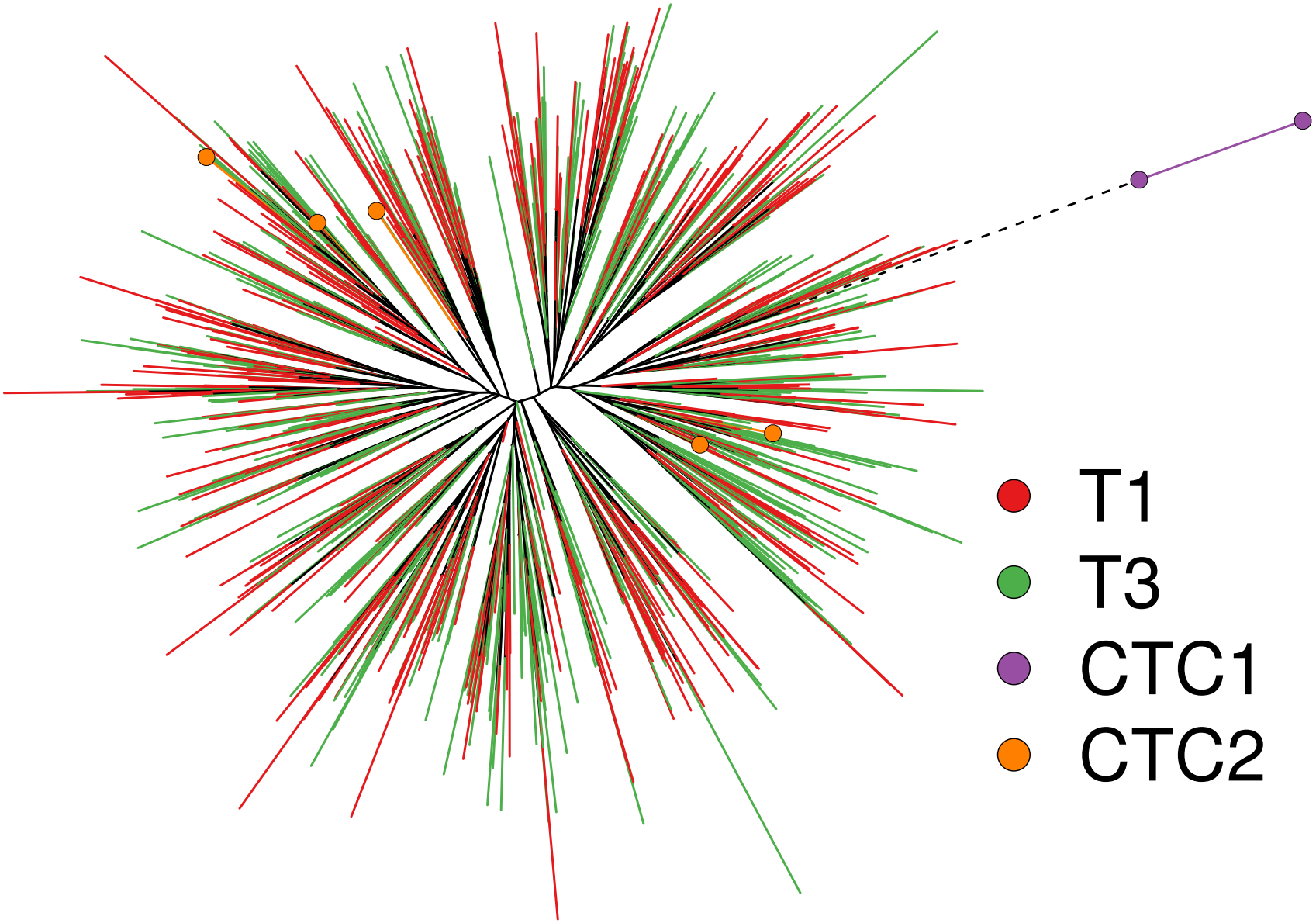


SNV 20% density

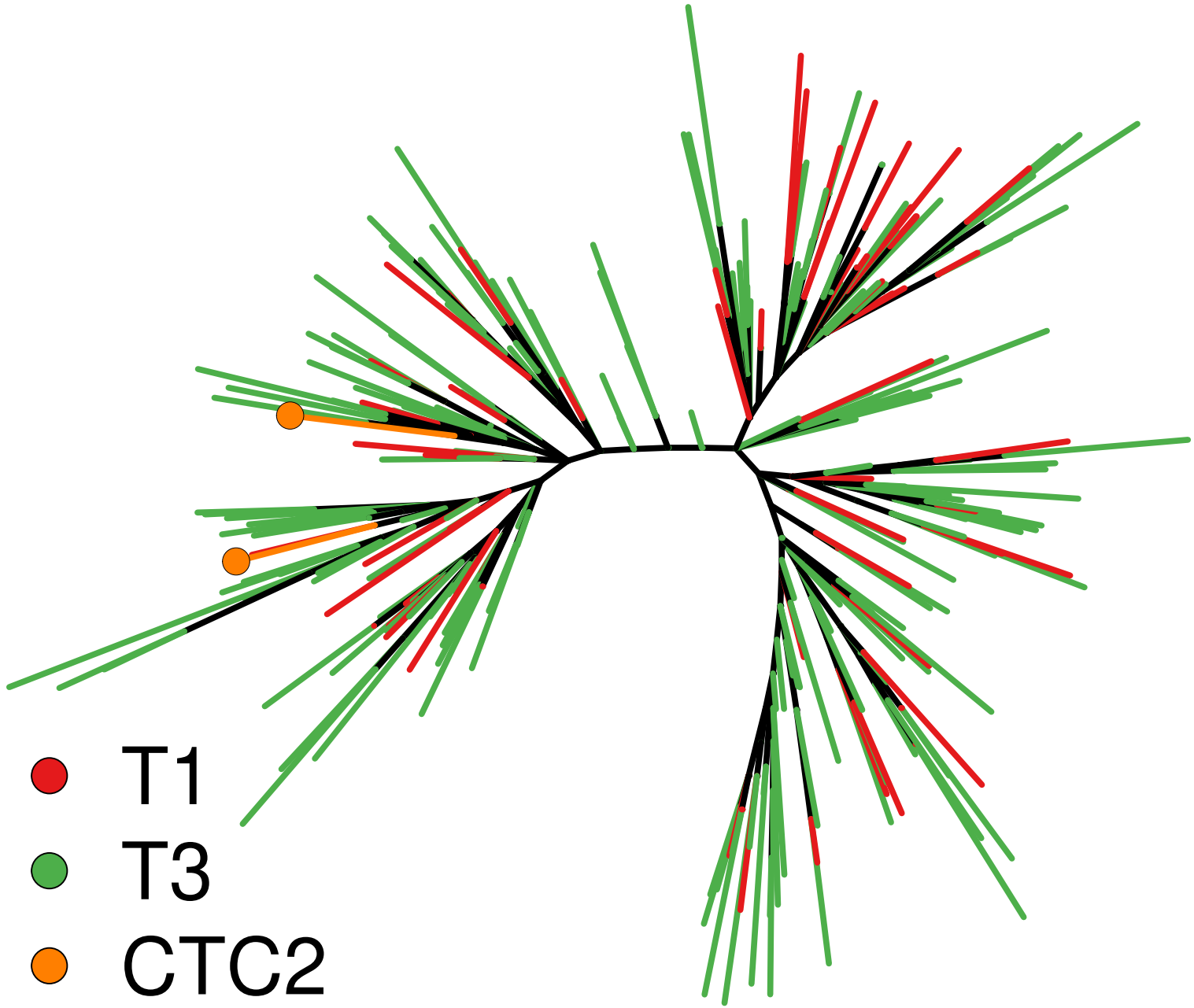
SNV 20% density



SNV 50% density



SNV 90% density



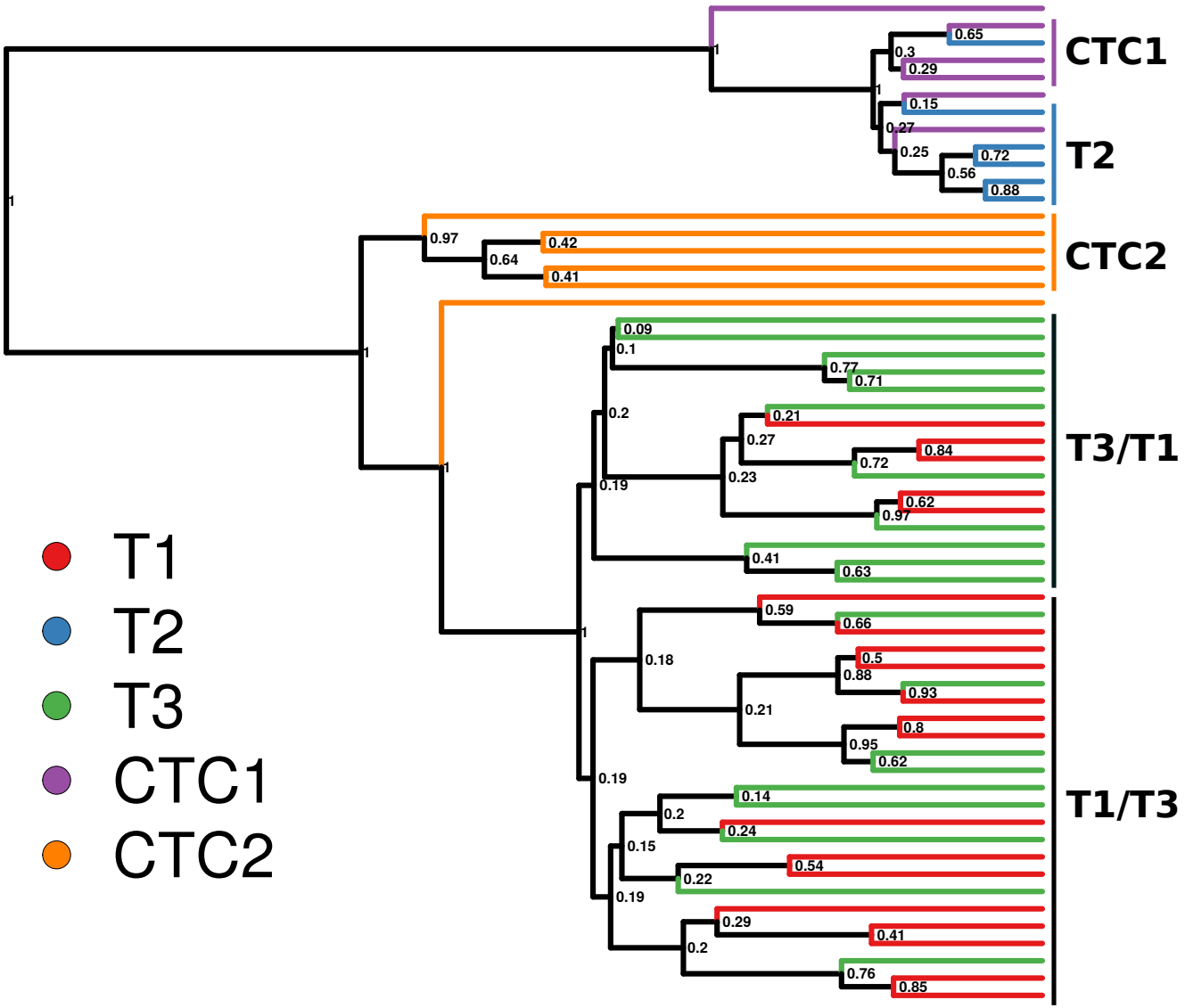
Conclusion

- Expression works well!
- SNV seems to fail.
- Statistical support for these trees is deceivingly high
- Bayesian analysis required, but datasets are too big or missing samples

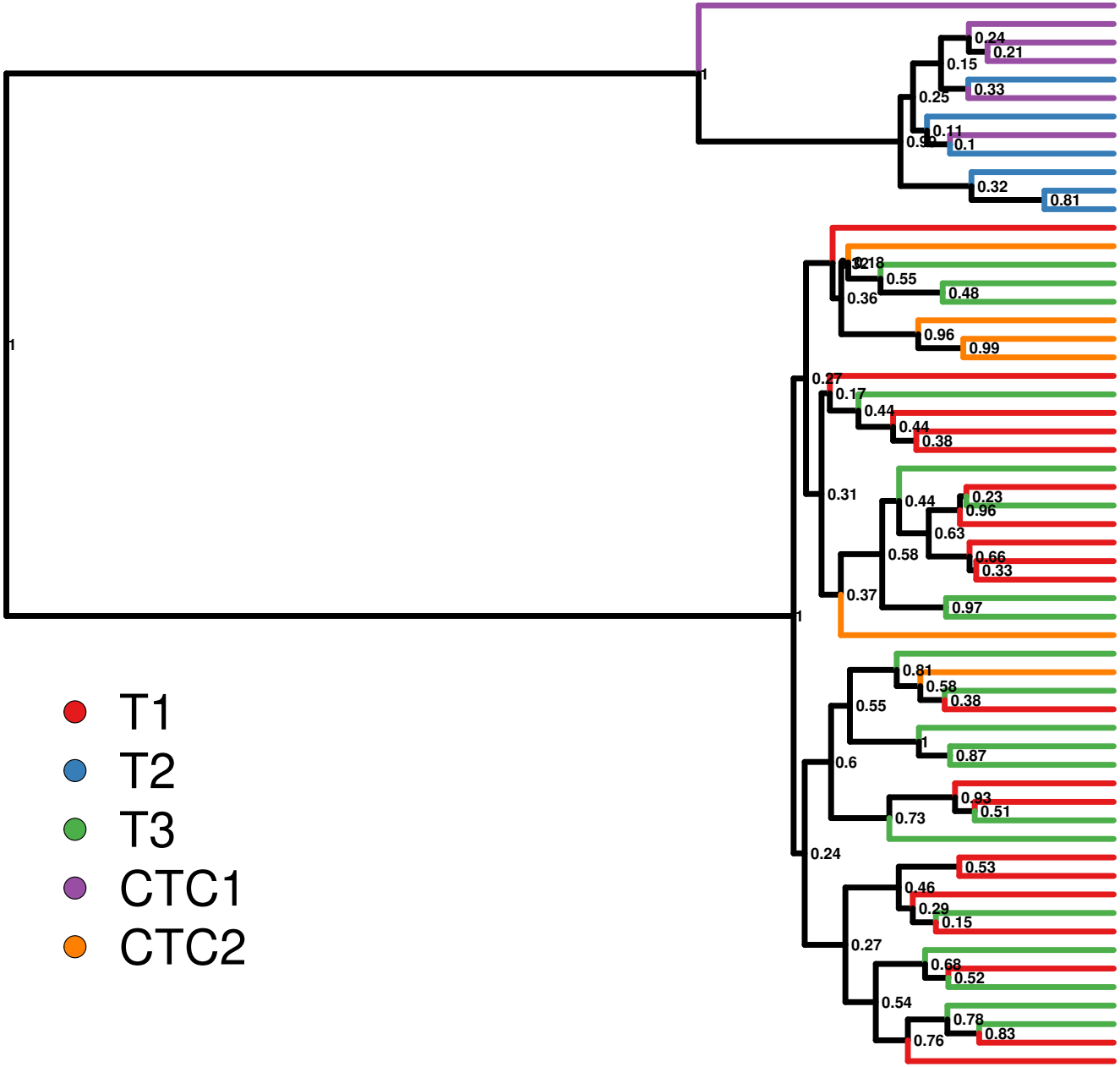
New filtering approach

- Data reduced to 58 sequences:
  - 20 sequences for T1 and T3
  - 6 sequences for T2, CTC1 and CTC2
- dataset filtered to:
  - 10% density (full dataset)
  - 50%
  - 90%
- Bayesian analysis using BEAST

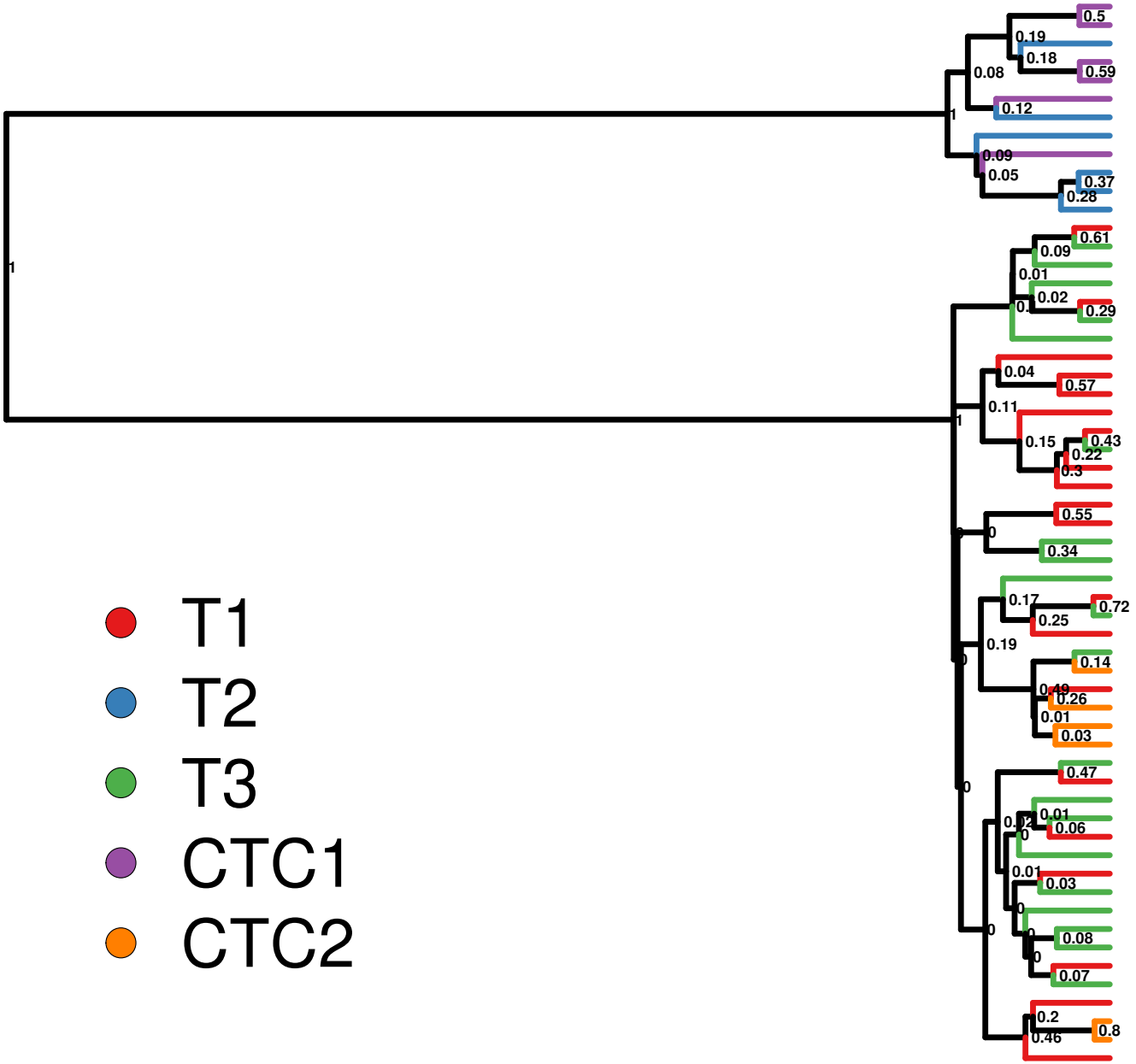
SNV 10% data density tree from Bayesian analysis



SNV 50% data density tree from Bayesian analysis



SNV 90% data density tree from Bayesian analysis



Conclusion

- scRNA-seq does contain phylogenetic signal!
- decent phylogeny can be constructed from both expression and SNVs
- a lot of space for improvement: scRNA-seq callers and error models

Acknowledgment

RUTHERFORD DISCOVERY FELLOWSHIPS

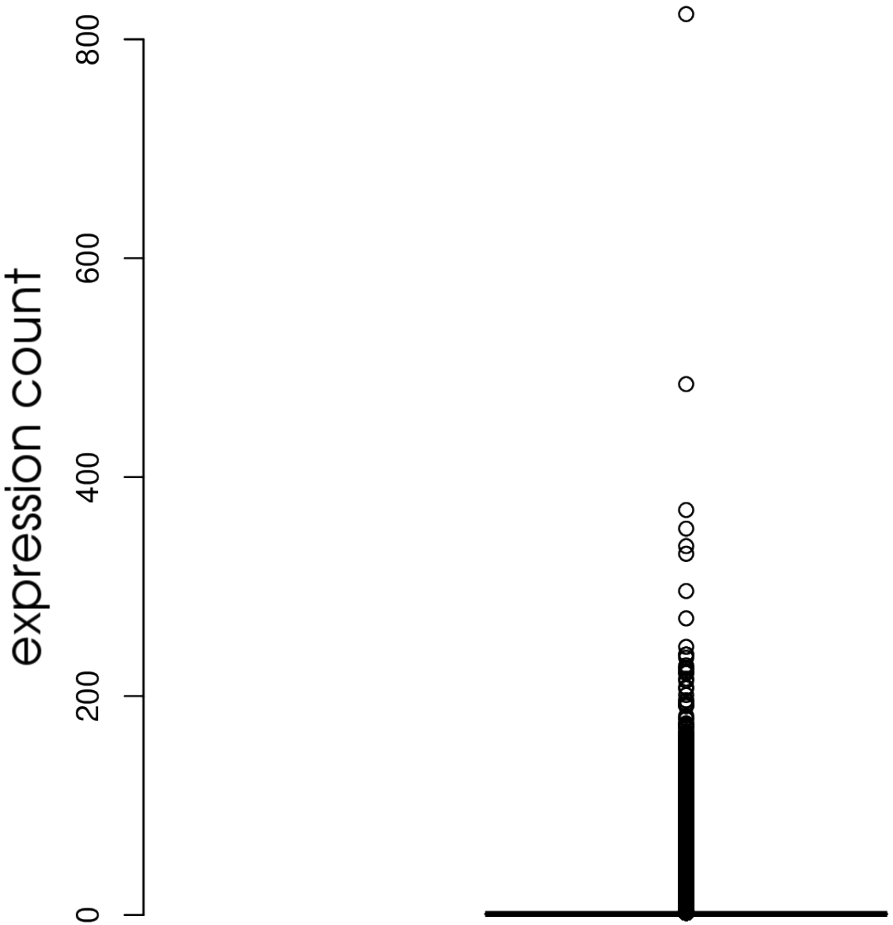
ROYAL SOCIETY TE APĀRANGI

MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

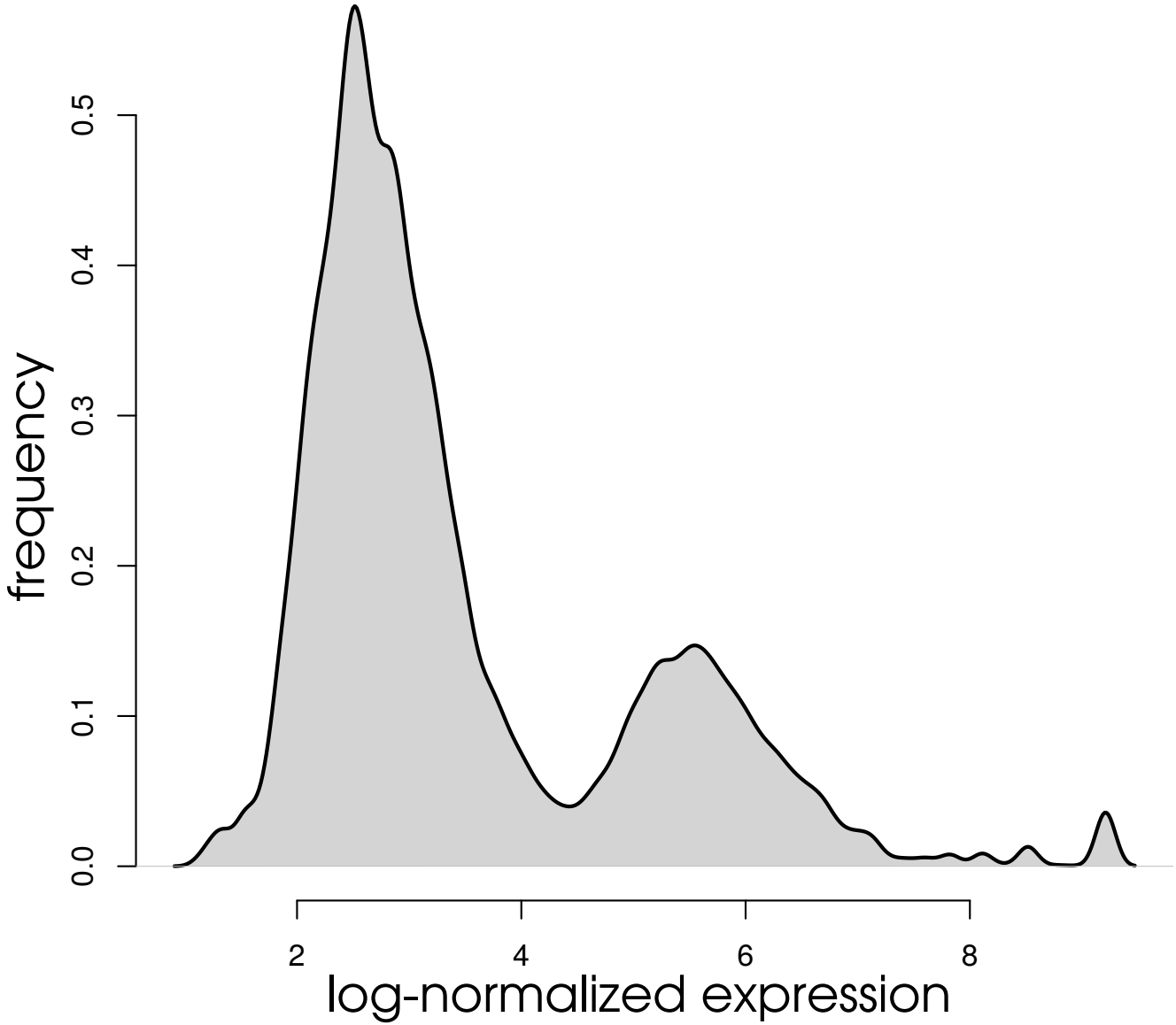
NIH
National Institutes of Health

Supplementary materials

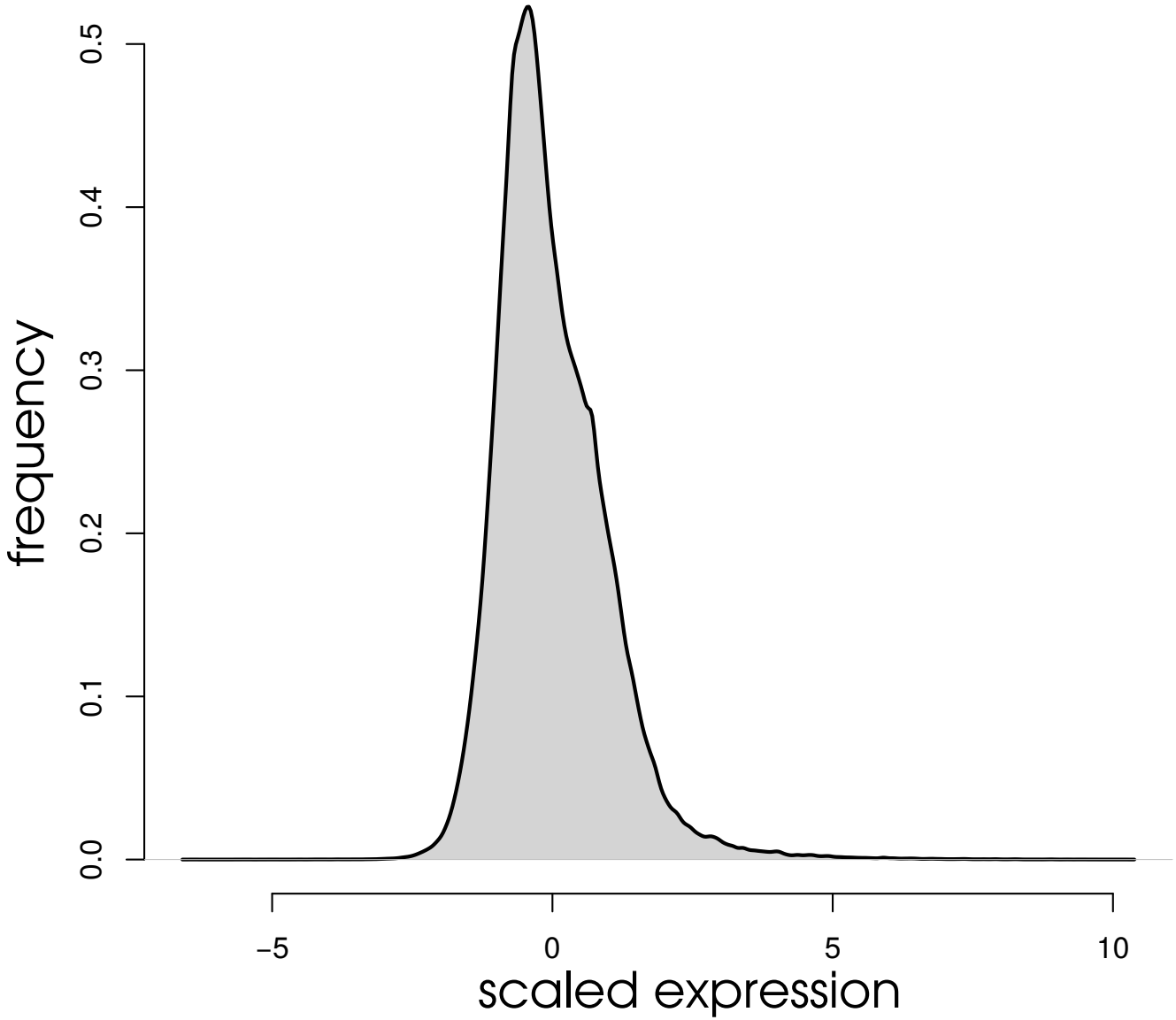
Boxplot of expression counts



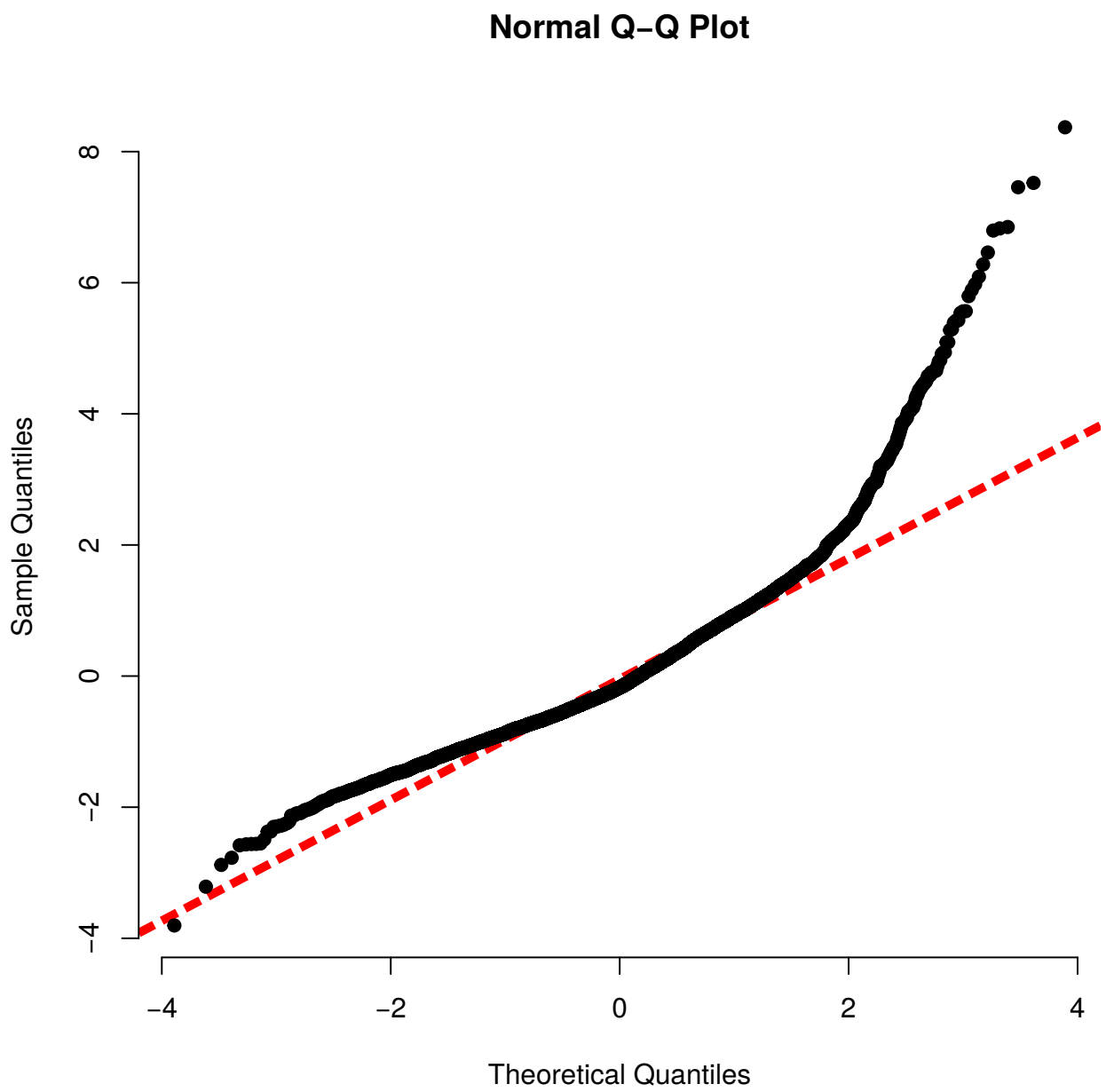
Log-normalization



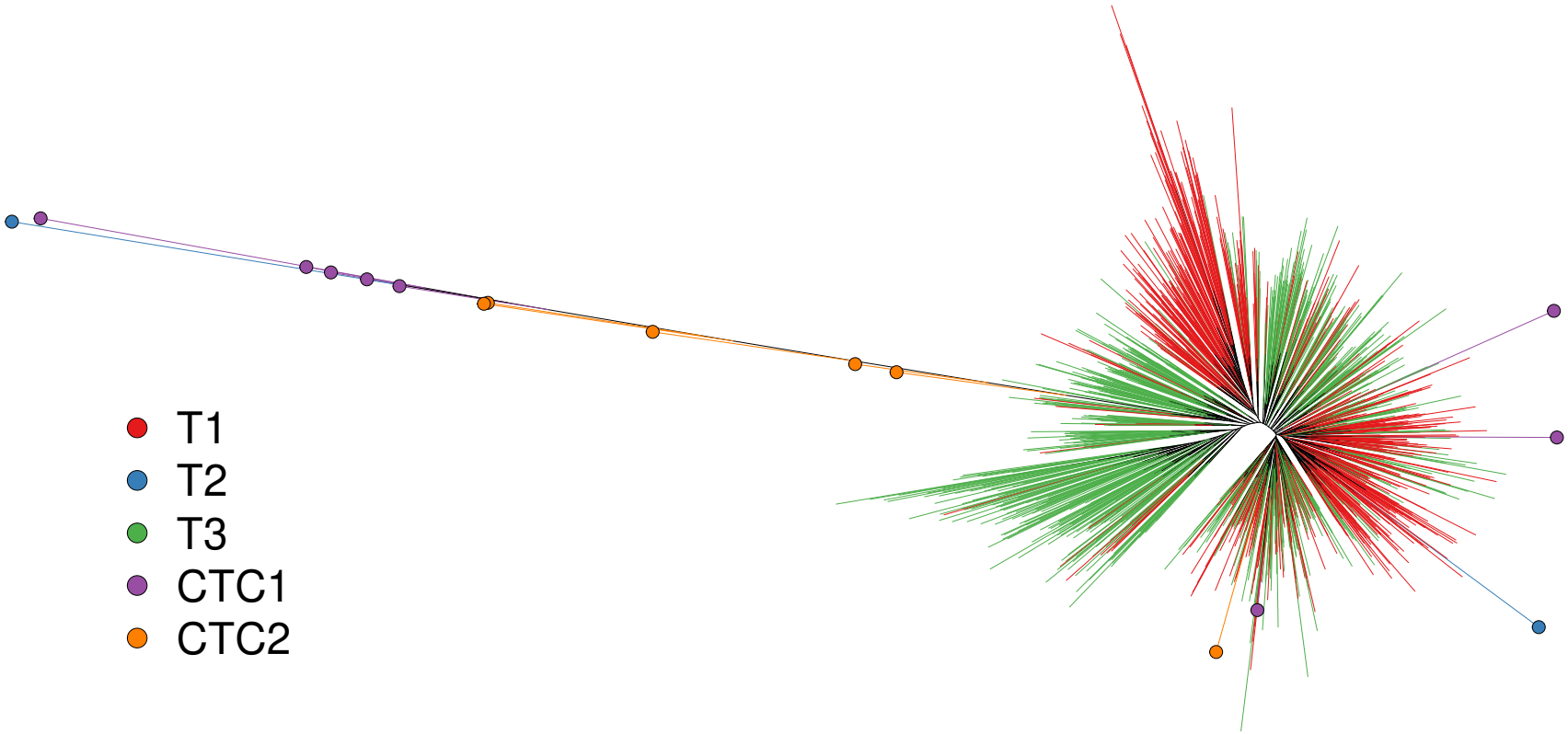
Scaling: $\mu = 0$, $\sigma = 1$



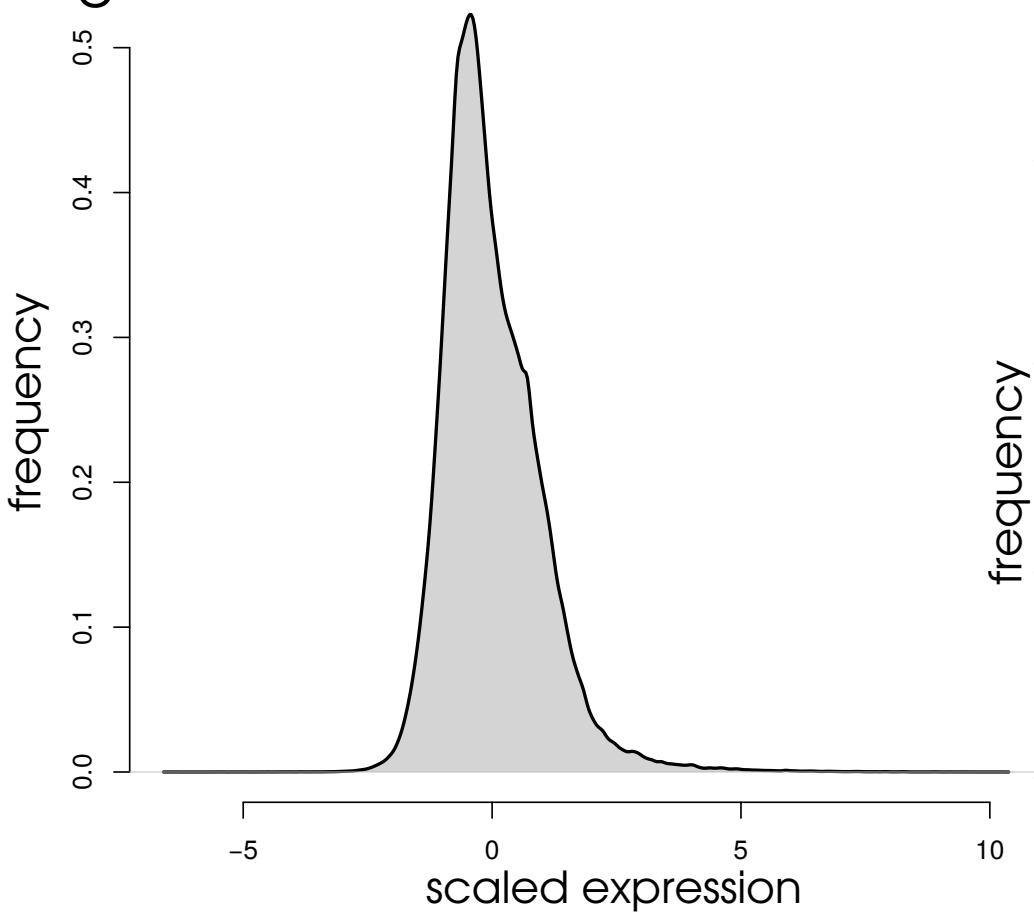
Normality test



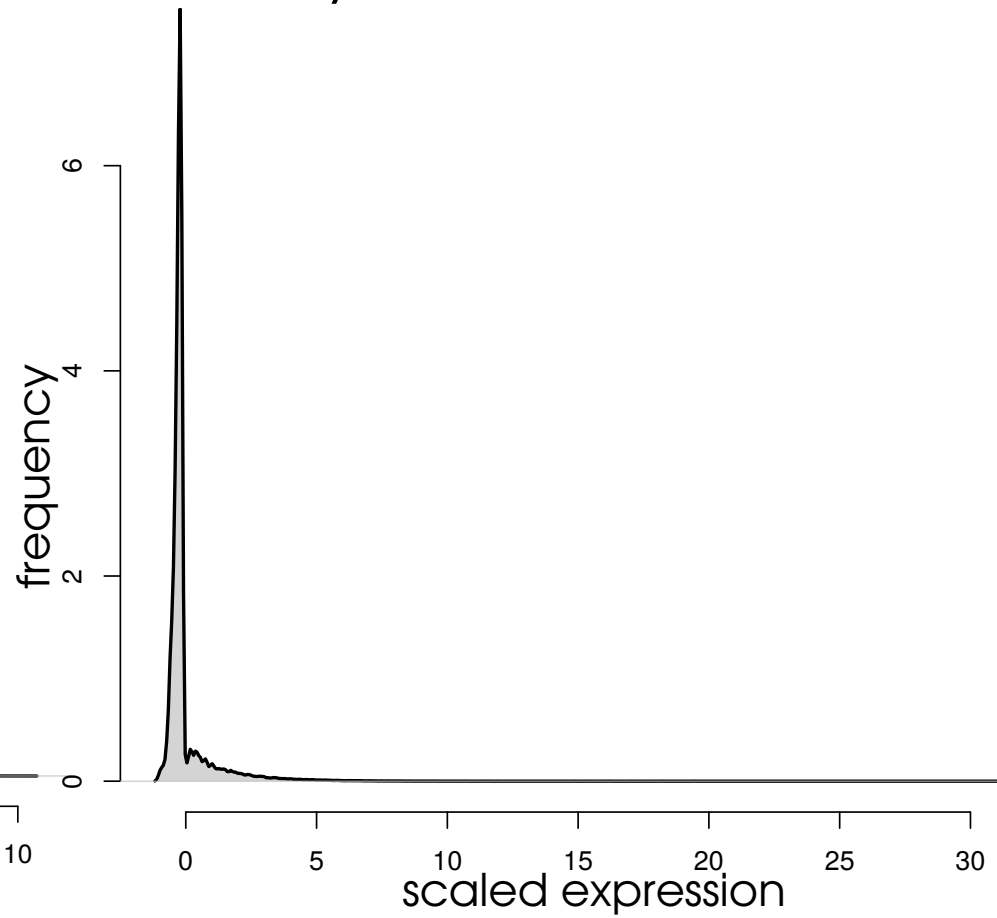

T1
T2
T3
CTC1
CTC2


T1
T2
T3
CTC1
CTC2

Lognormalized and scaled:



Scaled only:

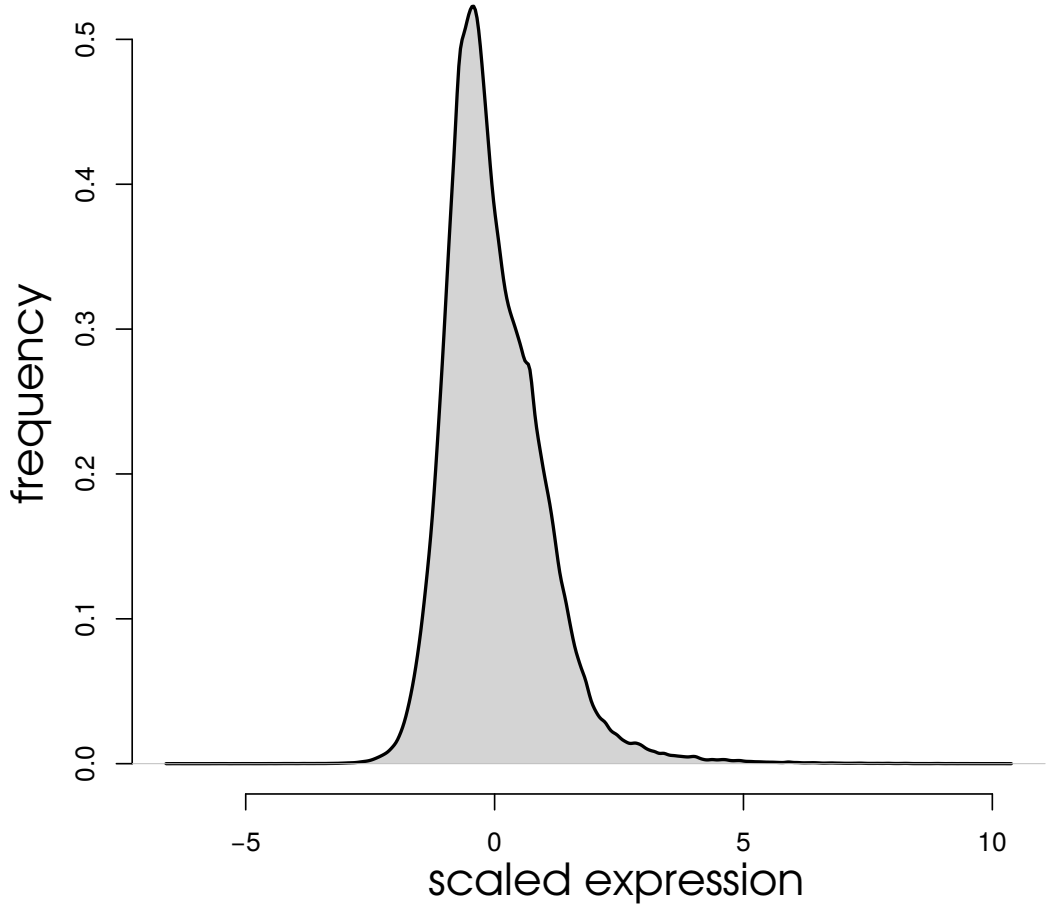


Discretization

Problem: Divide expression into groups.

Possible solutions:

- Centering around:
  - 0
  - mean
  - modus
- Intervals:
  - symetric
  - quantiles
  - HDI

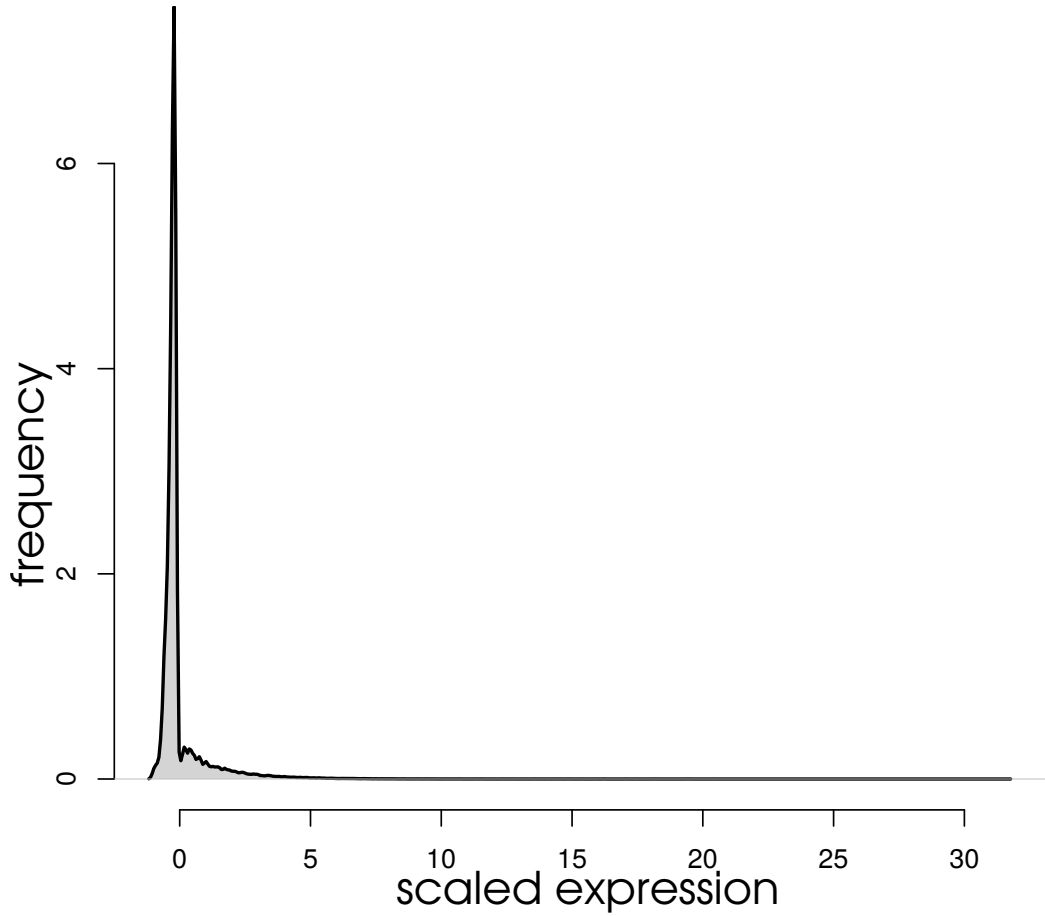


Discretization

Problem: Divide expression into groups.

Possible solutions:

- Centering around:
  - 0
  - mean
  - modus
- Intervals:
  - symetric
  - quantiles
  - HDI



Filtering data

- Calculate sums of columns and rows (colsum, rowsum)
- Find smallest colsum/rowsum
- Substract the least represented columns/rows from the rowsum/colsum
- Remove the columns/rows from the matrix

$$\begin{array}{cccc} 3 & 2 & 2 & 2 \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right) & \begin{array}{c} 3 \\ 2 \\ 2 \\ 1 \\ 1 \end{array} \end{array}$$

Filtering data

- Calculate sums of columns and rows (colsum, rowsum)
- Find smallest colsum/rowsum
- Substract the least represented columns/rows from the rowsum/colsum
- Remove the columns/rows from the matrix

$$\begin{array}{cccc} 2 & 2 & 2 & 1 \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right) & \begin{array}{c} 3 \\ 2 \\ 2 \\ 1 \\ 1 \end{array} \end{array}$$

Filtering data

- Calculate sums of columns and rows (colsum, rowsum)
- Find smallest colsum/rowsum
- Substract the least represented columns/rows from the rowsum/colsum
- Remove the columns/rows from the matrix

$$\begin{array}{cccc} 2 & 2 & 2 & 1 \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{array}\right) & & & \begin{array}{c} 3 \\ 2 \\ 2 \end{array} \end{array}$$

Bootstrap and posterior scores:

		Matrix density		
		0.2	0.5	0.9
Bootstrap	**Expression**	74.63	77.36	75.52
	SNV	74.26	70.99	61.92
Posterior	**SNV**	53	54	24

Statistical significance:

- Bootstrap $>$ 70
- Posterior $>$ 95