Memorial Sloan Kettering
Cancer Center

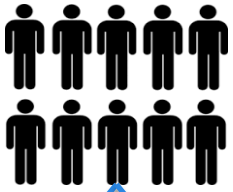
Multi-Omics Supervised Integrative Clustering *(MOSAIC)* on scNMT-seq mouse gastrulation dataset

Arshi Arora
BIRSBiointegration 2020
joint work with Dr. Ronglai Shen

June 17, 2020

Arshi Arora
Research Biostatistician II
Memorial Sloan Kettering Cancer Center

Motivation



- DNA Methylation
- mRNA expression
- miRNA expression
- Copy Number
- Somatic Mutation
- Protein
- Mutation signature
- Single Cell Sequencing
- …

- Overall Survival (time-event)
- Progression Free Survival (time-event)
- Response (categorical)

***M**ulti-**O**mics **S**upervised **I**ntegrated **C**lustering or (**MOSAIC**)*

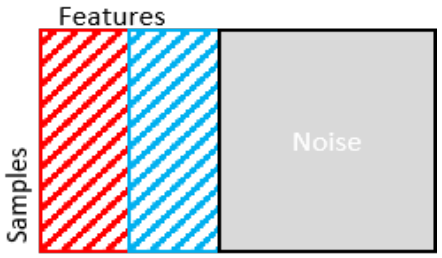
survClust*

*Arora A, Olshen AB, Seshan VE, and Shen R. Pan-cancer identification of clinically relevant genomic subtypes using outcome-weighted integrative clustering. Biorxiv

Memorial Sloan Kettering Cancer Center

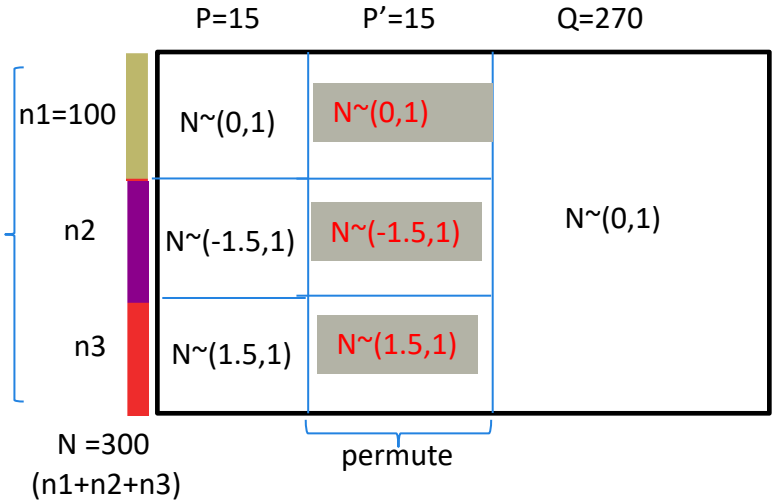
unsupervised vs supervised clustering via simulation

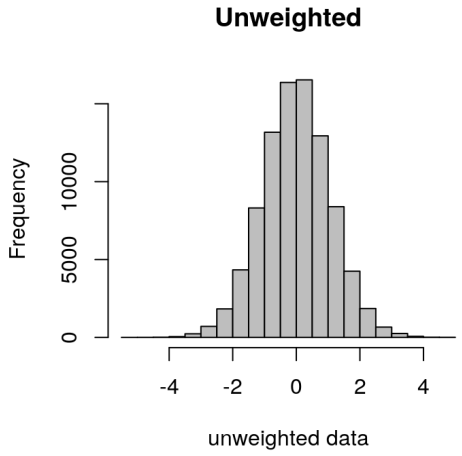
Typical data set

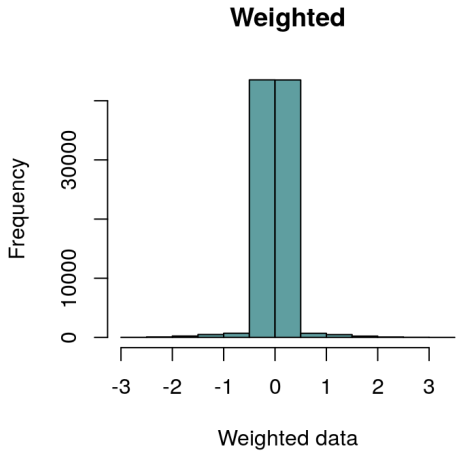


Kmeans clustering vs simulated truth			
	1	2	3
1	68	0	0
2	32	41	28
3	0	59	72

MOSAIC 3-class vs simulated truth			
	1	2	3
1	100	0	0
2	0	100	0
3	0	0	100





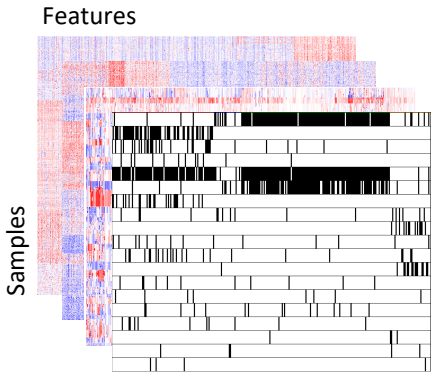


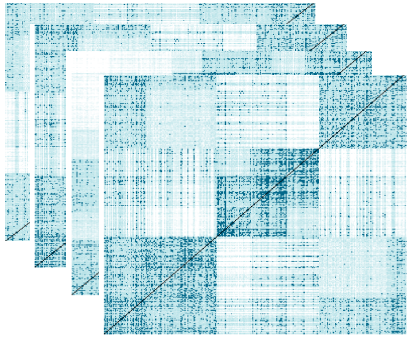
* Unsupervised clustering solution was arrived by running *k-means* algorithm

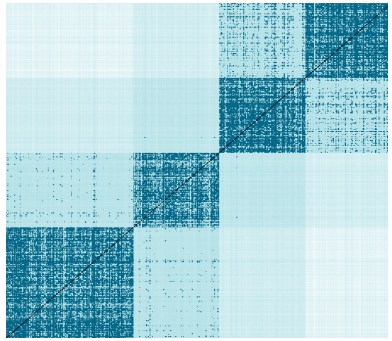
Memorial Sloan Kettering Cancer Center

MOSAIC Workflow

Raw Data	getDist	combineDist	MOSAIC
Various molecular platforms	Weighted Distance Matrix	Integrated Distance Matrix	Cluster membership with weighted outcome







MOSAIC 3-class vs simulated truth			
	1	2	3
1	100	0	0
2	0	100	0
3	0	0	10

$$I_w = \frac{\sum_{m=1}^{M} D_m}{M}$$

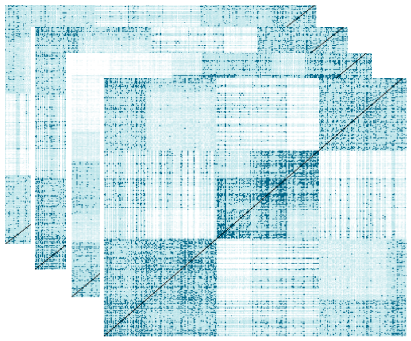
Where,

D_m = weighted distance matrix of mth data type

Step 2- getDist

getDist

Weighted Distance Matrix



Consider a data type $\boldsymbol{X_m}$ (where, m=1, .., M data types) of varying samples(N_m) and features (p_m)

$\boldsymbol{a_p}$ and $\boldsymbol{b_p}$ are a pair of samples measured for p features

The weighted distance[1] –

$$d_w(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^T \boldsymbol{W} (\boldsymbol{a} - \boldsymbol{b})}$$

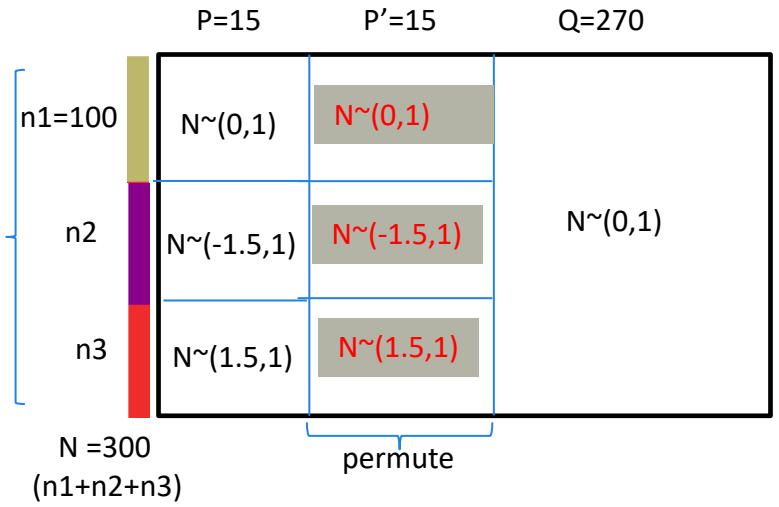
Where $\boldsymbol{W}$ is a $p \times p$ diagonal weight matrix with $\boldsymbol{W} = diag\{w_1, \ldots, w_p\}$.

$$\boldsymbol{X'} = \boldsymbol{X} * \boldsymbol{W}^{1/2}$$

$$d_w(\boldsymbol{a'}, \boldsymbol{b'}) = d_w(\boldsymbol{b'}, \boldsymbol{a'}) = \sqrt{\sum_{j=1}^{p} (a_j' - b_j')^2}$$

References:
1. Xing, Eric P., et al. "Distance metric learning with application to clustering with side-information." *Advances in neural information processing systems*. 2003.

Step 2- getDist – calculation of weights



$$w_{jc} = \log\left[\frac{l\left(x_{ijc}\middle|\mu_{jc},\sigma_{jc}^2\right)}{l\left(x_{ijc}\middle|\mu_j,\sigma_j^2\right)}\right]$$

$$w_j = \max\left(w_{j1}, w_{j2}, \ldots w_{jk}\right)$$

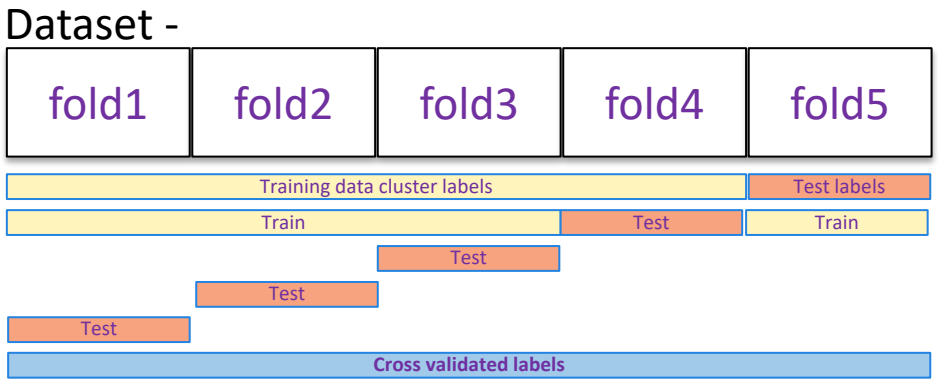
Where x_{ijc}, is the expression value of m[th] datatype for i[th] sample and j[th] feature

μ_{jc} = mean of a feature j only considering samples belonging to cluster c, where $c = 1,2,3 \ldots k$, σ_{jc}^2 = standard deviation of a feature j only considering samples belonging to cluster c

μ_j= population mean, all samples across all clusters,, σ_j^2= population standard deviation, considering all samples

Overfitting is avoided by cross-validation

- We did 5-fold cross validation for 50 rounds of cross validation to arrive at a consolidated solution for a particular k cluster



Concludes one round of cross-validation

- Perform 50 such rounds – with random 5 splits of the data
- Collect 50 cross validated survClust predicted class labels for each k = 2 to 7

scNMT seq Mouse gastrulation – Input data

	#cells	features	missing	features missing >50% samples	final features	final missing
acc_DHS	826	290	0.19	0	290	0.19
acc_p300	826	138	0.34	0	138	0.34
acc_cgi	826	4459	0.33	0	4459	0.33
acc_CTCF	826	898	0.37	0	898	0.37
acc_promoter	826	16518	0.28	0	5000	0.30
acc_genebody	826	17139	0.14	0	5000	0.24
met_DHS	826	66	0.24	3	63	0.22
met_p300	826	101	0.45	24	77	0.43
met_cgi	826	5536	0.42	511	5000	0.41
met_CTCF	826	175	0.48	51	124	0.46
met_promoter	826	12092	0.40	595	5000	0.42
met_genebody	826	15837	0.22	140	5000	0.24
rna	826	18345	0.00	0	5000	0.00

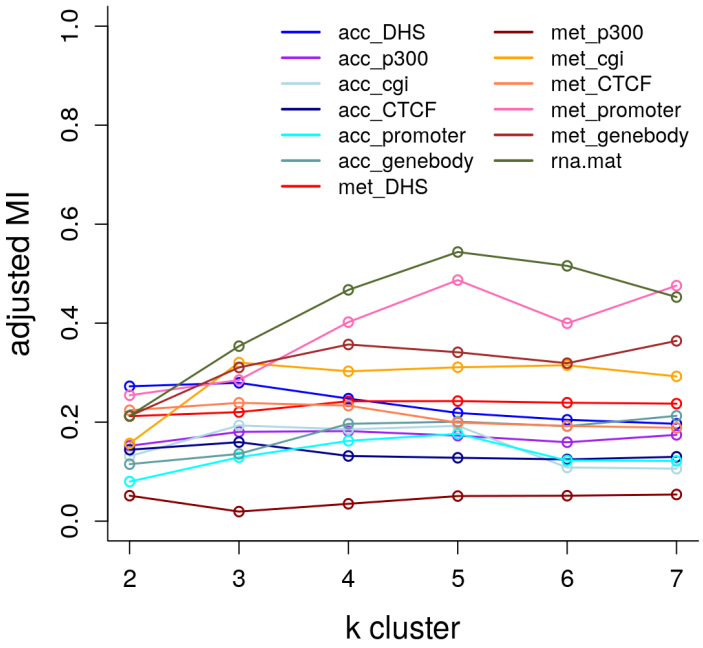
Results – MOSAIC with Stage

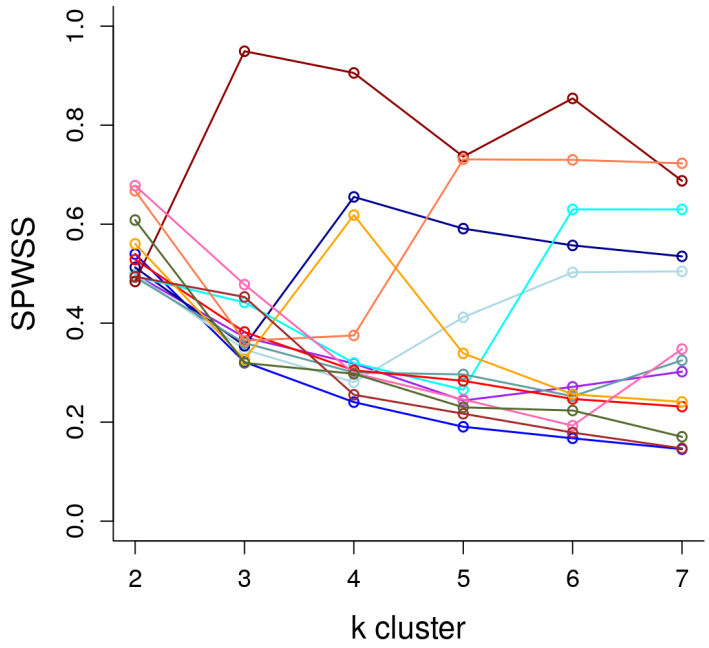
MOSAIC was run on 13 data types wrt stage. For 5 folds and 50 rounds of CV.

stage			
E4.5	**E5.5**	**E6.5**	**E7.5**
104(12.59%)	108(13.08%)	271(32.81%)	343(41.53%)

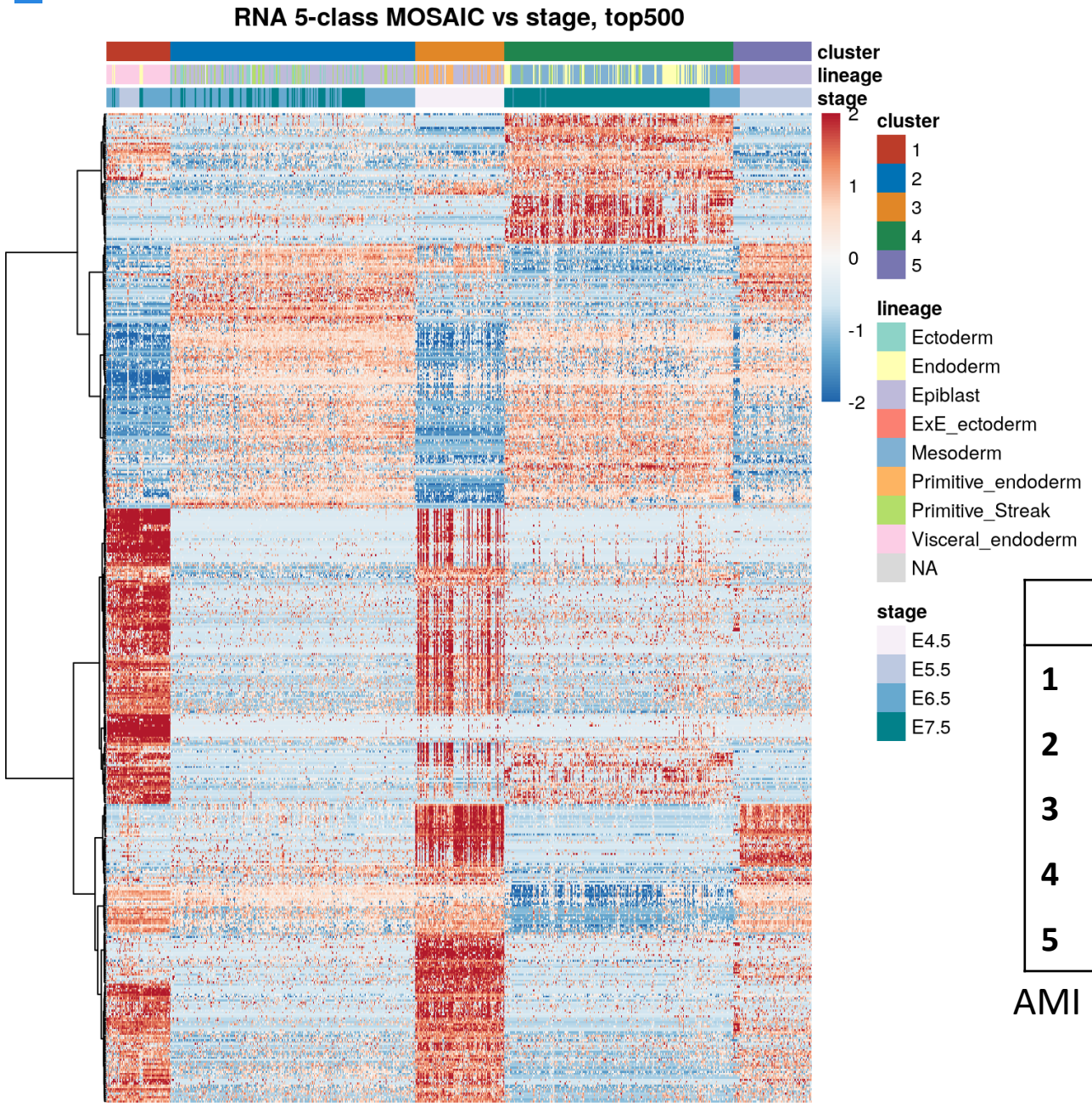
A *k* was picked as follows –

- Highest **adjusted Mutual Information (MI)**
- Lowest **Standardized Pooled Within Sum of Squares** (SPWSS)





MOSAIC on RNA data type with Stage



	E4.5	E5.5	E6.5	E7.5
1	0	24	45	6
2	0	0	187	100
3	104	0	0	0
4	0	0	31	237
5	0	84	8	0

AMI = 0.55

Memorial Sloan Kettering Cancer Center

RNA MOSAIC solution vs kmeans

	E4.5	E5.5	E6.5	E7.5
1	0	24	45	6
2	0	0	187	100
3	104	0	0	0
4	0	0	31	237
5	0	84	8	0

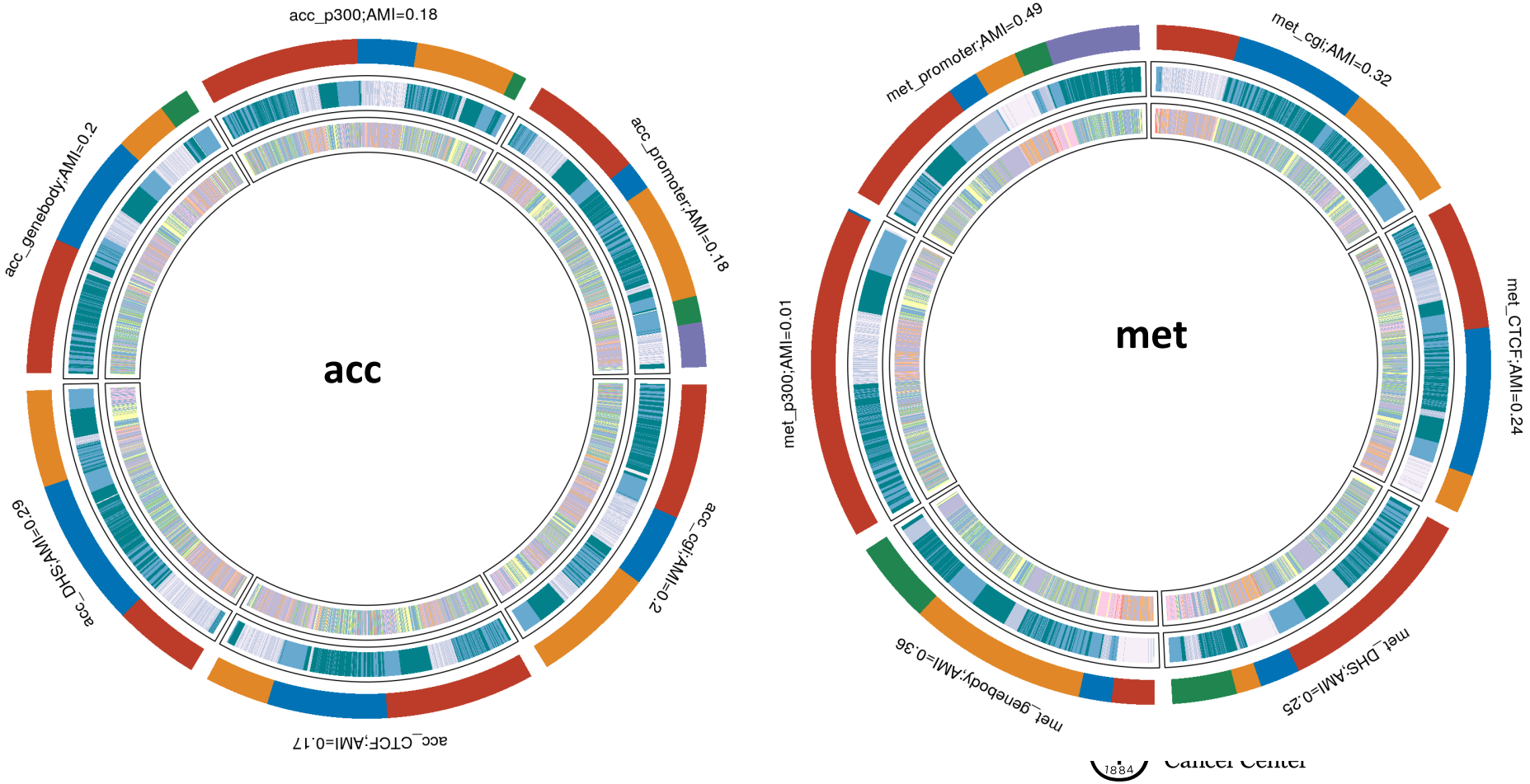
AMI = 0.55, AMI for lineage 0.56

	E4.5	E5.5	E6.5	E7.5
1	0	0	30	228
2	3	7	74	20
3	58	0	0	0
4	0	77	125	89
5	43	24	42	6

AMI = 0.34, add AMI for lineage =0.51

	Ectoderm	Endoderm	Epiblast	ExE_ectoderm	Mesoderm	Primitive_endoderm	Primitive_Streak	Visceral_endoderm	NA
E4.5	0	0	60	0	0	43	0	0	1
E5.5	0	0	84	0	0	0	0	24	0
E6.5	0	0	146	8	28	0	43	45	1
E7.5	43	81	44	0	141	0	33	0	1

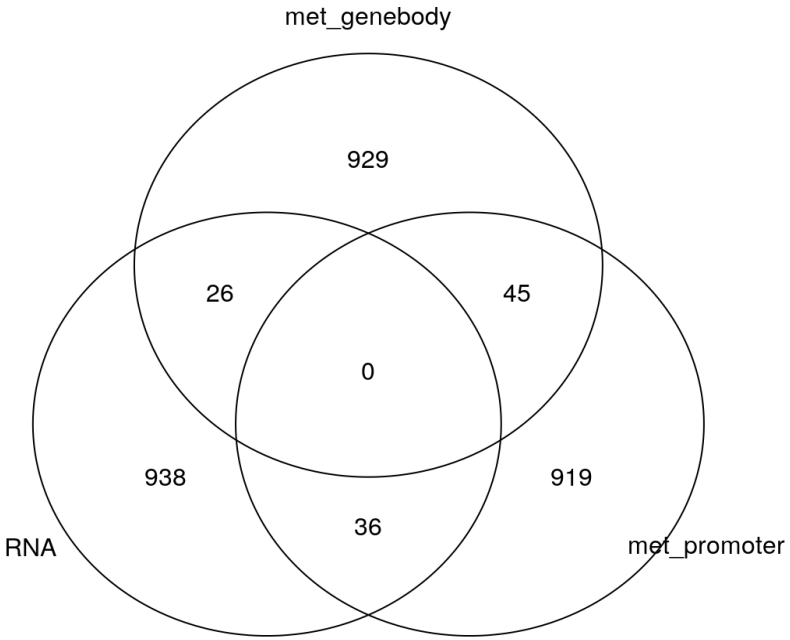
MOSAIC solutions for other data types



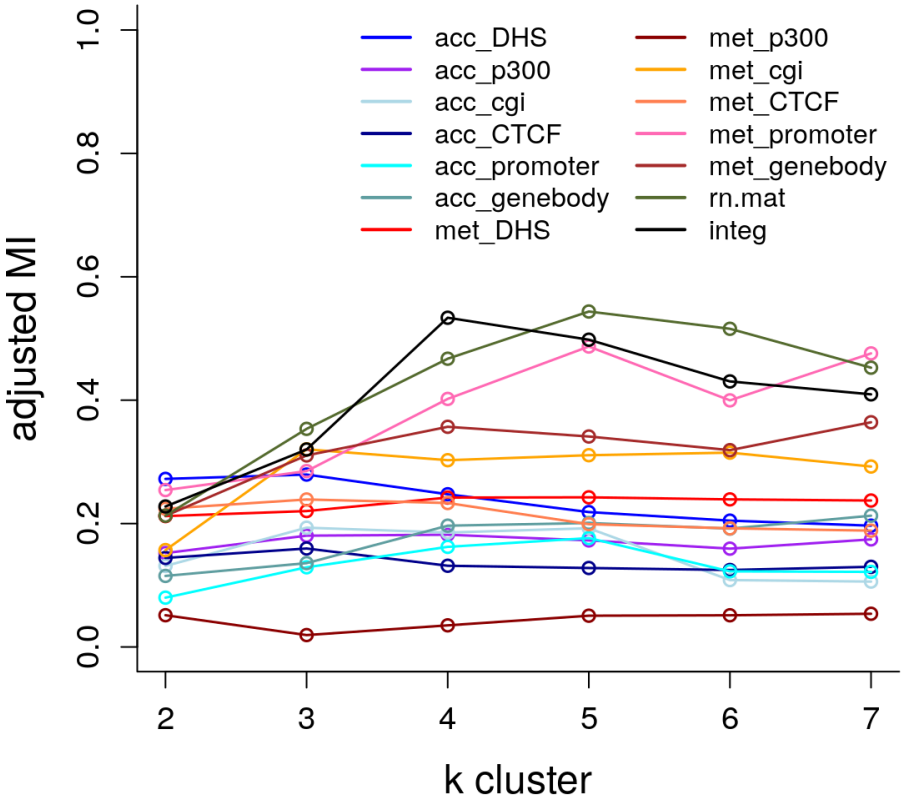
Integrating 5 data types and stage as outcome

Data type	AMI	Features
RNA	0.56	5000
met_promoter	0.49	5000
met_genebody	0.36	5000
met_cgi	0.32	5000
acc_DHS	0.29	290

Overlap between top 1000 genes



Integrating 5 data types and stage as outcome – AMI tracks close to rna



Integrated solution

AMI = 0.53, stage				
	E4.5	**E5.5**	**E6.5**	**E7.5**
1	0	1	211	337
2	0	83	7	0
3	1	22	52	6
4	103	2	1	0

AMI = 0.62, RNA k5 solution					
rnak5	**1**	**2**	**3**	**4**	**5**
Integ 1	0	280	0	268	1
2	0	7	0	0	83
3	72	0	1	0	8
4	3	0	103	0	0

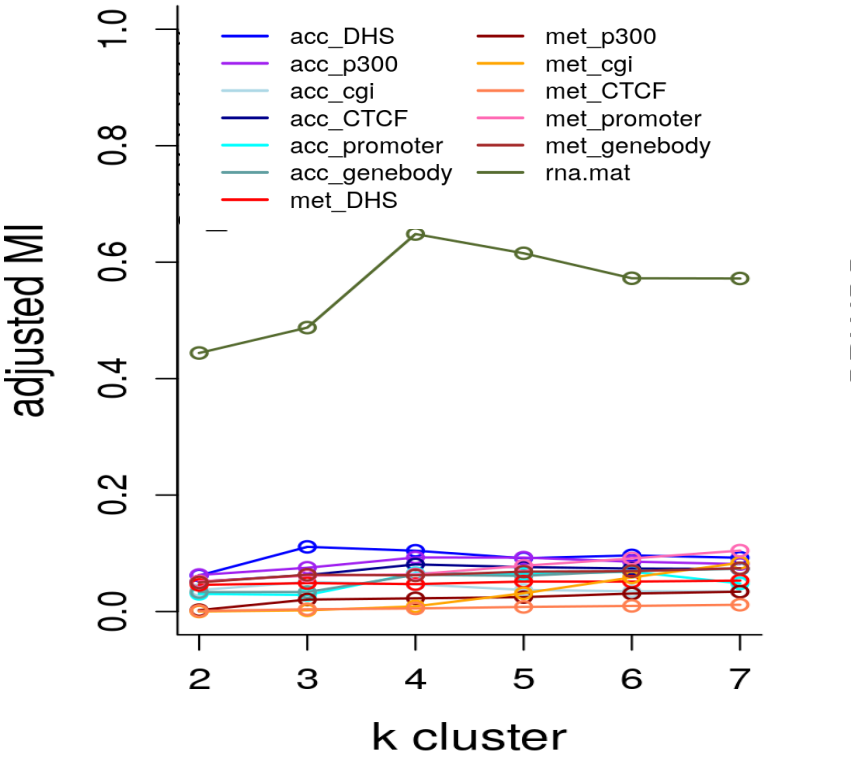
AMI = 0.33, lineage								
	Ectoderm	**Endoderm**	**Epiblast**	**ExE_ectoderm**	**Mesoderm**	**Primitive_endoderm**	**Primitive_Streak**	**Visceral_endoderm**
1	43	75	185	0	169	0	75	0
2	0	0	89	0	0	0	1	0
3	0	6	0	8	0	0	0	66
4	0	0	60	0	0	43	0	3

Results – MOSAIC with Lineage

MOSAIC was run on 13 data types wrt stage. For 5 folds and 50 rounds of CV.

Ectoderm	Endoderm	Epiblast	ExE_ectoderm	Mesoderm	Primitive_endoderm	Primitive_Streak	Visceral_endoderm	<NA>
43(5.21%)	81(9.81%)	334(40.44%)	8(0.97%)	169(20.46%)	43(5.21%)	76(9.2%)	69(8.35%)	3(0.36%)

Ectoderm	Endoderm	Epiblast	Mesoderm	Primitive_Streak
43(6.12%)	81(11.52%)	334(47.51%)	169(24.04%)	76(10.81%)



Memorial Sloan Kettering Cancer Center

RNA MOSAIC with lineage vs kmeans

	Ectoderm	Endoderm	Epiblast	Mesoderm	Primitive_Streak
1	0	2	0	168	12
2	0	0	142	0	0
3	43	0	192	1	61
4	0	79	0	0	3

AMI = 0.65, AMI with stage 0.48

	E4.5	E5.5	E6.5	E7.5
1	0	0	30	228
2	3	7	74	20
3	58	0	0	0
4	0	77	125	89
5	43	24	42	6

AMI for stage =0.34, add AMI for lineage =0.51

Conclusion

- MOSAIC finds supervised clusters, with an outcome of interest in mind. Where kmeans might give mixed results. Supervised clustering is much more efficient and helps in sorting out different signals
- MOSAIC can run with missing data. However interpretations should be made carefully.
- MOSAIC reduces computation space from sample x feature to sample x sample
- Efficient in dealing with noisy features

Future Work:

- Imputation of missing data – area where a lot of research has been done.
- In scNMT mouse data, stages have a temporal relationship, perhaps model ordinal relationship.
- Joint modeling of stage and lineage
- Integrated solution can be further improved

References

- Shen, R. et al. Integrative subtype discovery in glioblastoma using iCluster. 7, e35236 (2012).
- Olshen, A.B., Venkatraman, E., Lucito, R. & Wigler, M.J.B. Circular binary segmentation for the analysis of array-based DNA copy number data. 5, 557-572 (2004).
- Xing, E.P., Jordan, M.I., Russell, S.J. & Ng, A.Y. in Advances in neural information processing systems 521-528 (2003).
- Torgerson, W.S. Theory and methods of scaling. (1958).
- Hartigan, J.A. & Wong, M.A.J.J.o.t.R.S.S.S.C. Algorithm AS 136: A k-means clustering algorithm. 28, 100-108 (1979).
- Mardia, K.V.J.C.i.S.-T. & Methods Some properties of clasical multi-dimesional scaling. 7, 1233-1241 (1978).
- Legendre, P. & Gallagher, E.D.J.O. Ecologically meaningful transformations for ordination of species data. 129, 271-280 (2001).
- Tibshirani, R., Walther, G. & Hastie, T.J.J.o.t.R.S.S.S.B. Estimating the number of clusters in a data set via the gap statistic. 63, 411-423 (2001).
- Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. 500, 415 (2013).
- Robertson, A.G. et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. 171, 540-556. e525 (2017).
- Hoadley, Katherine A., et al. "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer." *Cell* 173.2 (2018): 291-304.

Thanks!
Questions?

Acknowledgements

Ronglai Shen, PhD
Associate Attending Biostatistician

Adam B. Olshen, PhD
Venkatraman E. Seshan, PhD

Memorial Sloan Kettering Cancer Center

EXTRA Slides

Step 2- getDist – calculation of weights

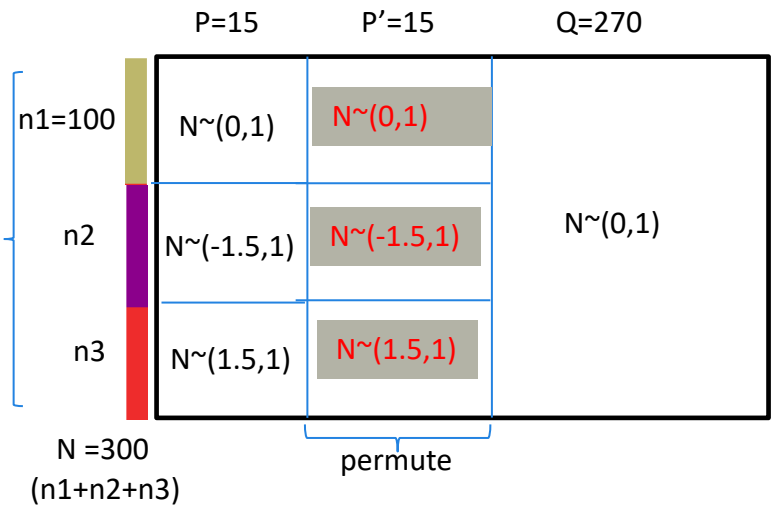
$$w_{jc} = \log\left[\frac{l\left(x_{ijc}\middle|\mu_{jc},\sigma_{jc}^2\right)}{l\left(x_{ijc}\middle|\mu_j,\sigma_j^2\right)}\right]$$

$$l\left(x_{ijc}\middle|\mu_{jc},\sigma_{jc}^2\right) = \sum_{i=1}^{n_c} log\left[\frac{1}{\sqrt{2\pi\sigma_{jc}^2}} exp\frac{-1}{2}\left\{\frac{\left(x_{ijc}-\mu_{jc}\right)^2}{\sigma_{jc}^2}\right\}\right]$$

$$l(x_{ijc}|\mu_{jc},\sigma_{jc}^2) = \sum_{i=1}^{n_c} log\left[\frac{1}{\sqrt{2\pi\sigma_j^2}} exp\frac{-1}{2}\left\{\frac{\left(x_{ijc}-\mu_j\right)^2}{\sigma_j^2}\right\}\right]$$

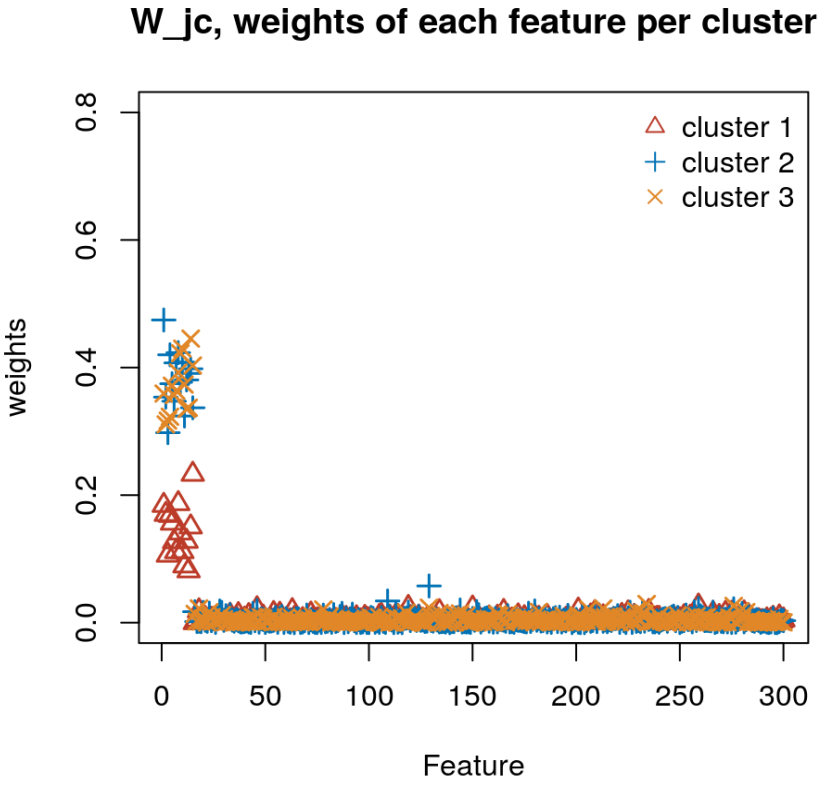
$$w_j = \max\left(w_{j1}, w_{j2}, \ldots w_{jk}\right)$$

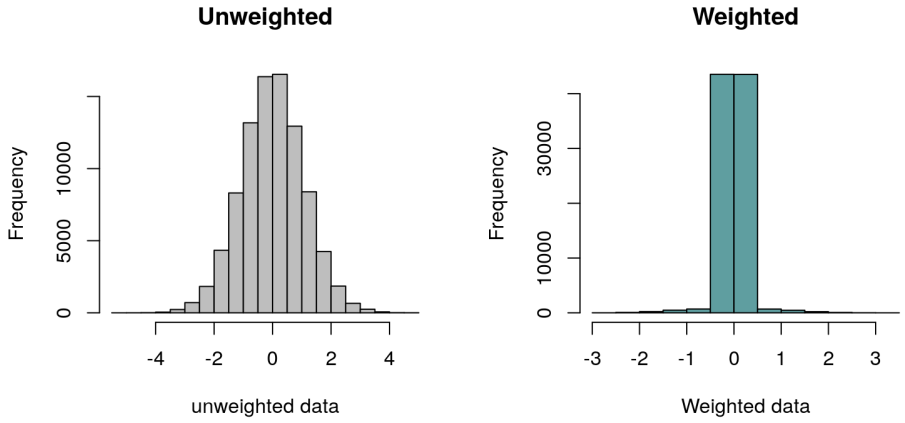
More on weights



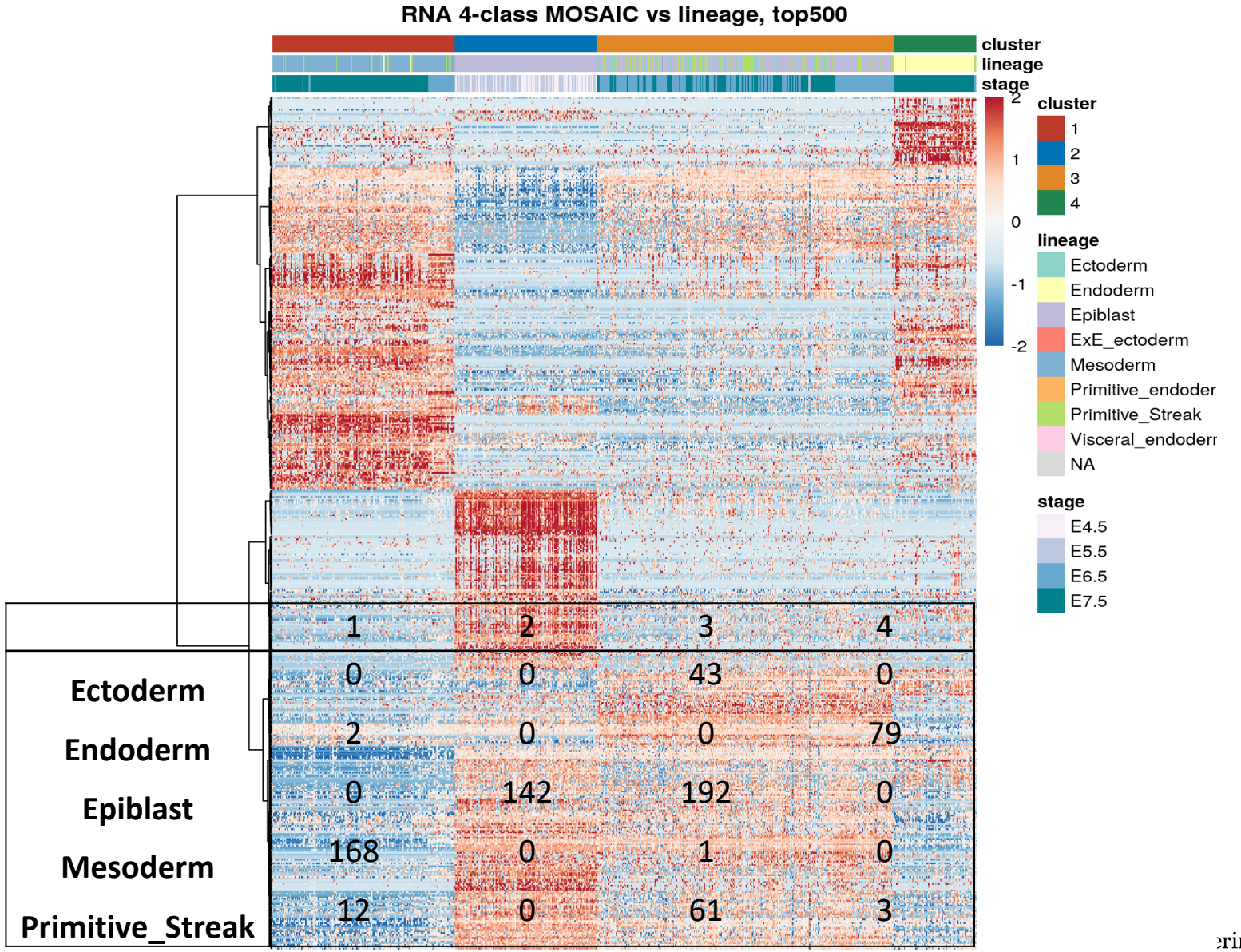
$$w_{jc} = \log\left[\frac{l\left(x_{ijc}\middle|\mu_{jc},\sigma_{jc}^2\right)}{l\left(x_{ijc}\middle|\mu_j,\sigma_j^2\right)}\right]$$

$$w_j = \max\left(w_{j1}, w_{j2}, w_{j3}\right)$$





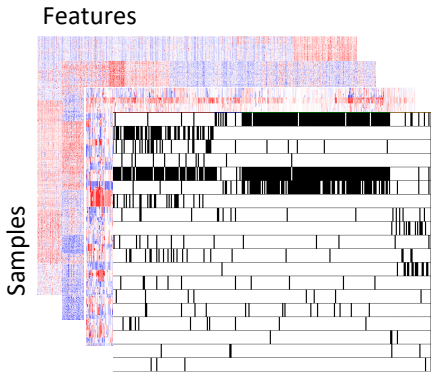
MOSAIC on RNA data type with Lineage



Step 1 – prepare input data

Raw Data

Various molecular platforms



- Continuous data should be standardized across features (columns)
- This ensure that weights are interpretable.

For proportion data – folded square root transformation

$$x_{ij} = \sqrt{x_{ij}} - \sqrt{(1 - x_{ij})}$$

Where i, ith sample , j, jth feature for a particular data type m, where m = met-DHS, … etc