Memorial Sloan Kettering
Cancer Center

survClust: An Outcome Weighted Learning Approach for Identifying Clinically Relevant Patient Subgroups from Large-scale Sequencing Data

joint work with – Adam B. Olshen,
Venkatraman E.Seshan,
and Ronglai Shen
ISMCO 2019

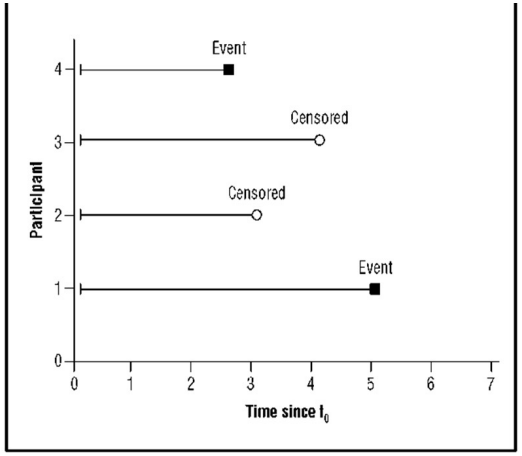
October 15, 2019

Arshi Arora
Research Biostatistician II
Memorial Sloan Kettering Cancer Center

Motivation

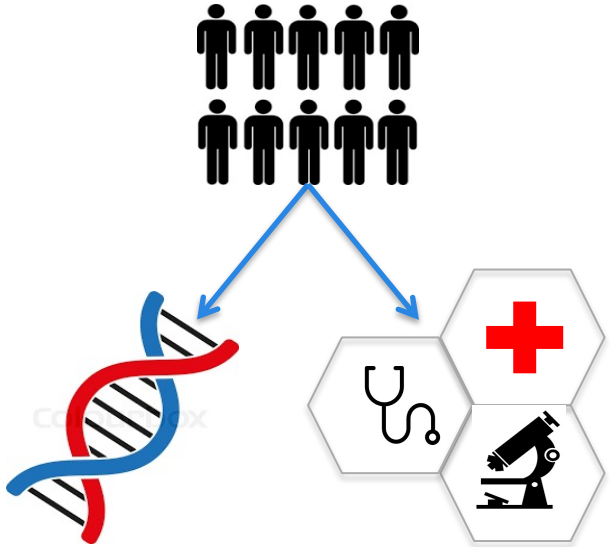
- TCGA generated multidimensional omics data across 10,000 tumors across 33 tumor types
- The main TCGA studies primarily focused on molecular subtype analysis using unsupervised clustering
- We aim to develop a supervised learning approach for patient outcome weighted stratification

+



References:
1. 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.

survClust



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

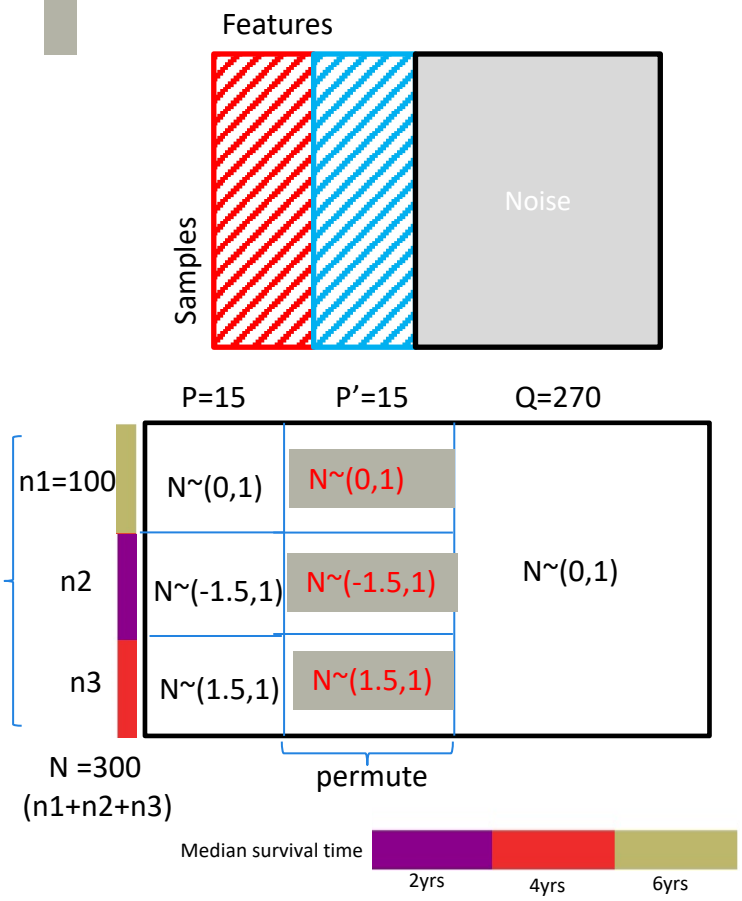
- Overall Survival
- Progression Free Survival
- Response (future work)

survClust

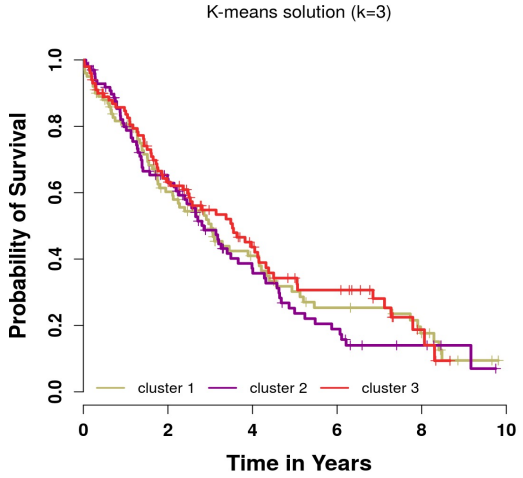
Memorial Sloan Kettering Cancer Center

unsupervised vs supervised clustering via simulation

Typical data set

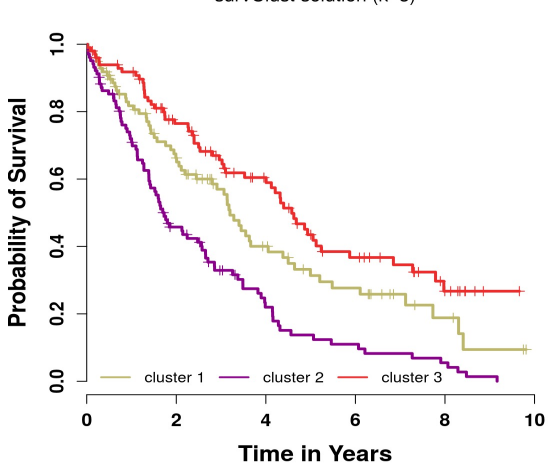


Unsupervised clustering solution*



	Cluster 1	Cluster 2	cluster 3
Truth Cluster 1	29	34	38
Truth Cluster 2	71	0	0
Truth Cluster 3	0	66	62

survClust solution



	Cluster 1	Cluster 2	cluster 3
Truth Cluster 1	0	0	100
Truth Cluster 2	0	100	0
Truth Cluster 3	100	0	0

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

survClust Workflow

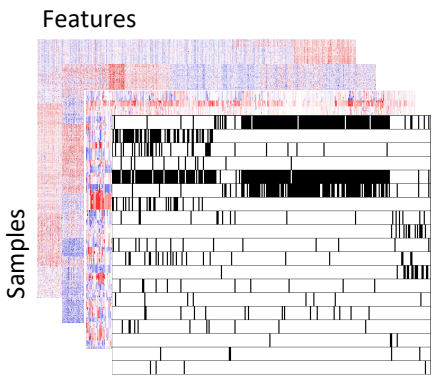
Raw Data → **getDist** → **combineDist** → **survClust**

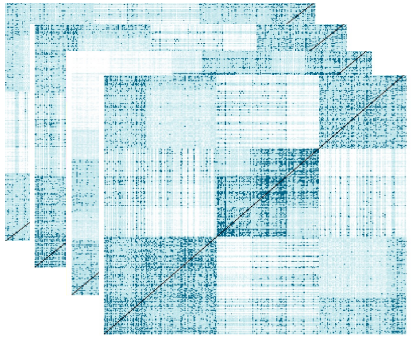
Various molecular platforms

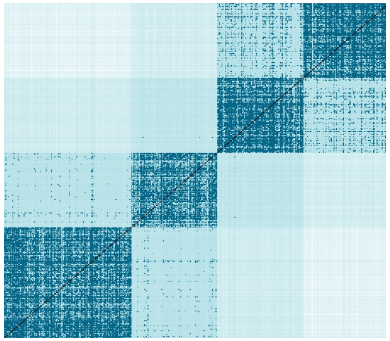
Weighted Distance Matrix

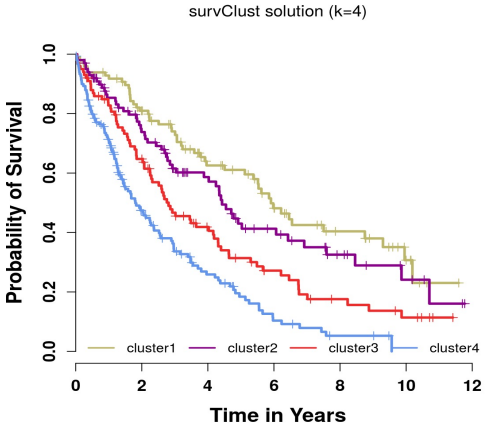
Integrated Distance Matrix

Cluster membership with survival difference





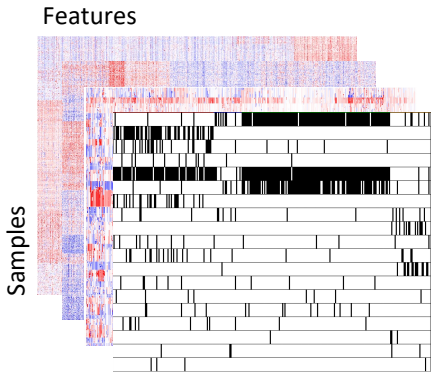




Step 1 – prepare input data

Raw Data

Various molecular platforms

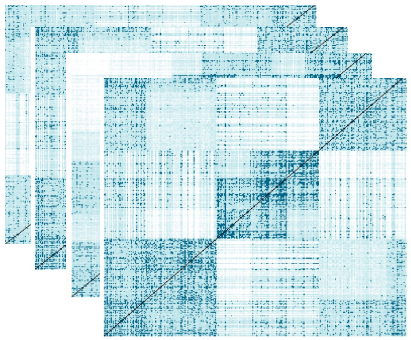


- Continuous data should be standardized across features (columns)
- This ensure that weights are interpretable.

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] –

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

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

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

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

- The scaling factor or weights w_p are obtained by fitting a univariate cox proportional model for each p –

$$h(t|\boldsymbol{x_j}) = h_o \times \exp(\boldsymbol{x_j^T} * \beta)$$

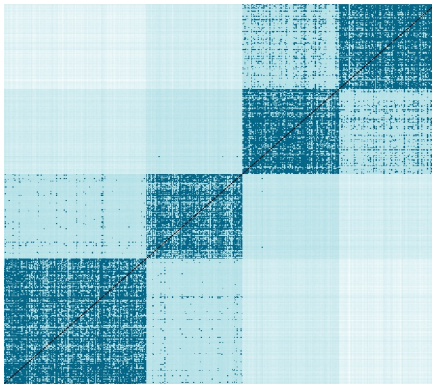
where j is the j[th] feature from $1 \ldots p$ features. t represents the survival time, $h(t)$ is the hazard function determined by p covariate, coefficient β determines the impact of covariate also known as w_p, and h_o is defined as baseline hazard.

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

Integrate and perform survClust

combineDist

Integrated Weighted Distance Matrix



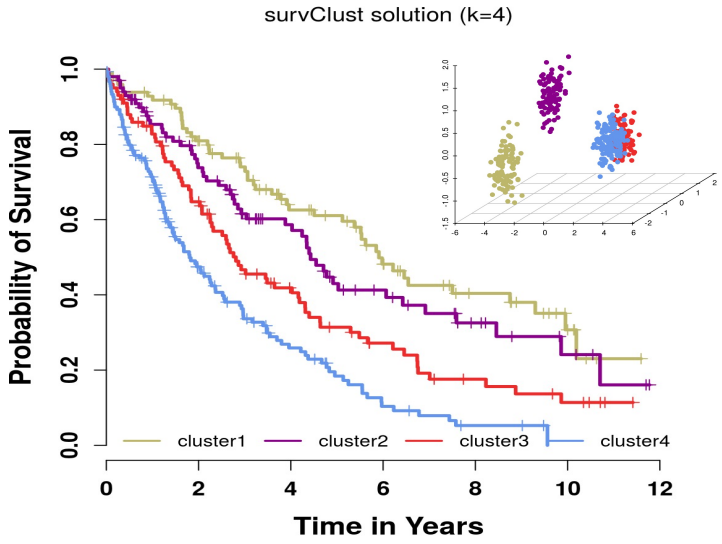
$$I_w = \frac{\sum_{m=1}^{M} D_m}{M}$$

Where,

D_m = weighted distance matrix of mth data type

survClust

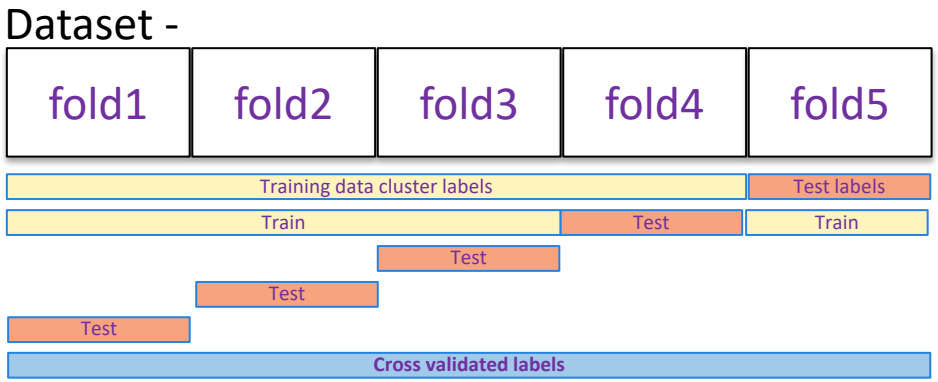
Cluster membership with survival difference



survClust then projects the integrated and weighted distance matrix in a lower dimensional space via multidimensional scaling and clustering sample points into subgroups via the K-means algorithm.

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

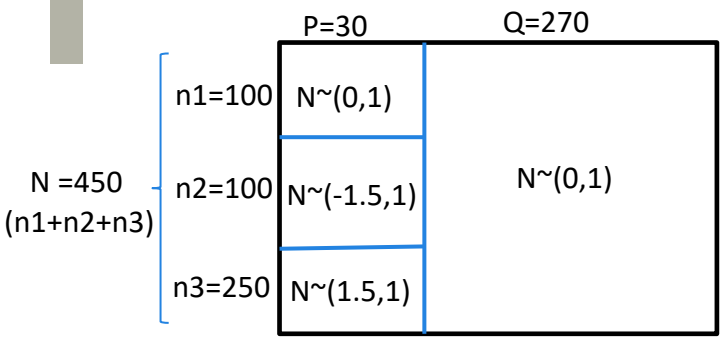
- Perform 50 such rounds – with random 5 splits of the data
- Collect 50 cross validated survClust predicted class labels for each k = 2 to 7

An example of cross validation and how to pick k

simulate two data types –

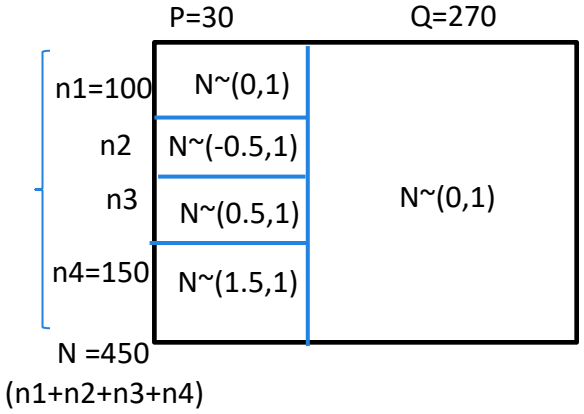
data type 1

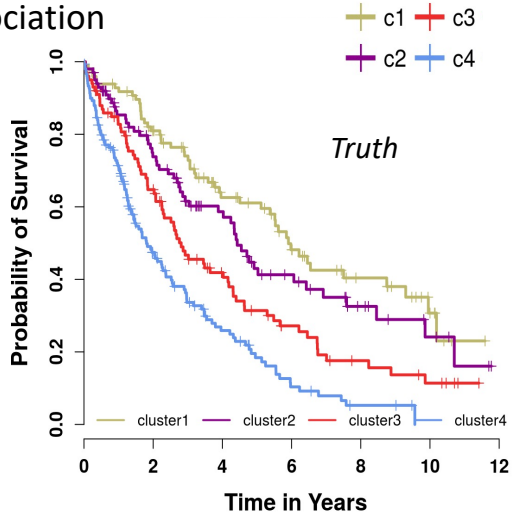
strong clusters, weak survival association

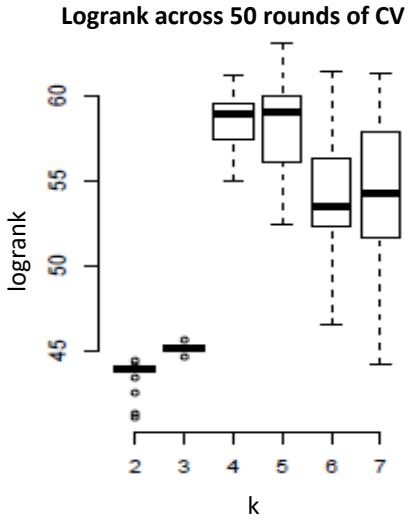


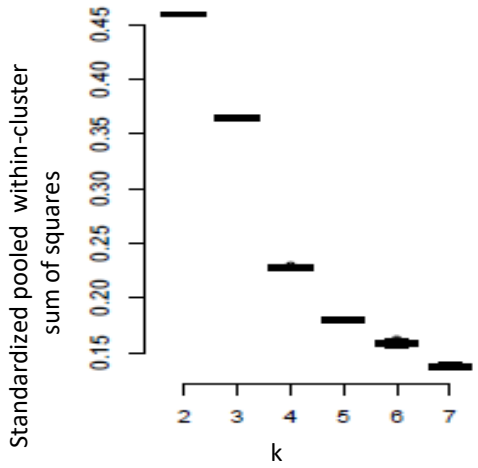
data type 2 –

weak clusters, strong survival association





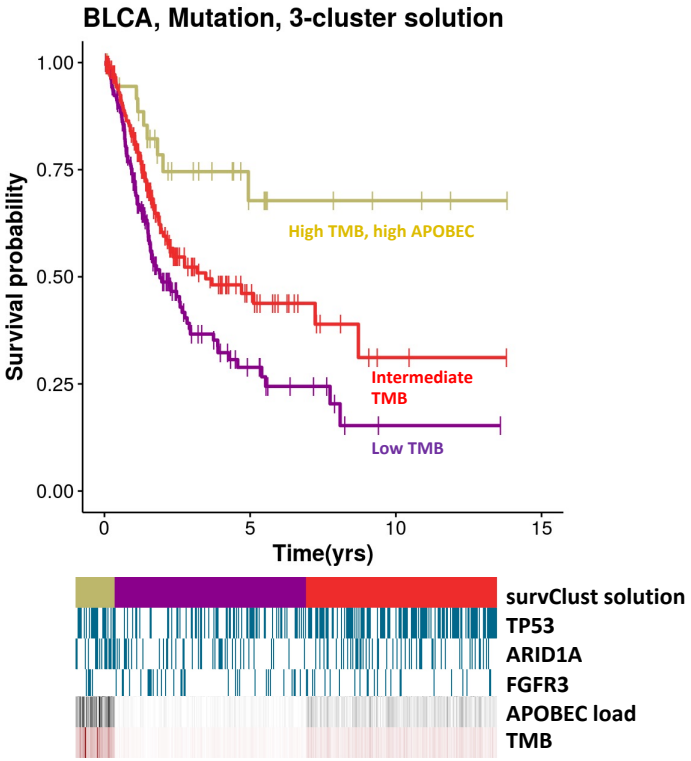




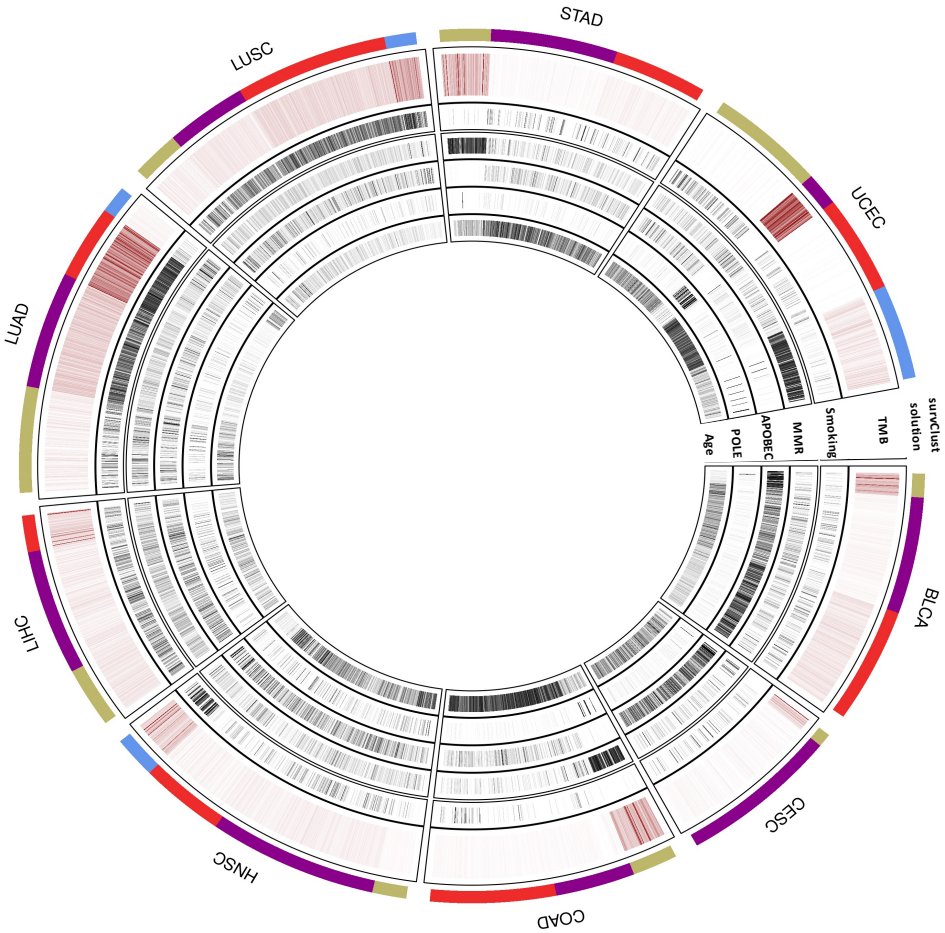
survClust labels	c1	c2	c3	c4
Simulated labels				
Cluster 1	100	0	0	0
Cluster 2	0	100	0	0
Cluster 3	0	0	100	0
Cluster 4	0	0	0	150

Mutation based stratification using survClust in TCGA datasets

a **Bladder Cancer (BLCA)**



b **A circomap showing survClust identified TMB patterns across cancer types**

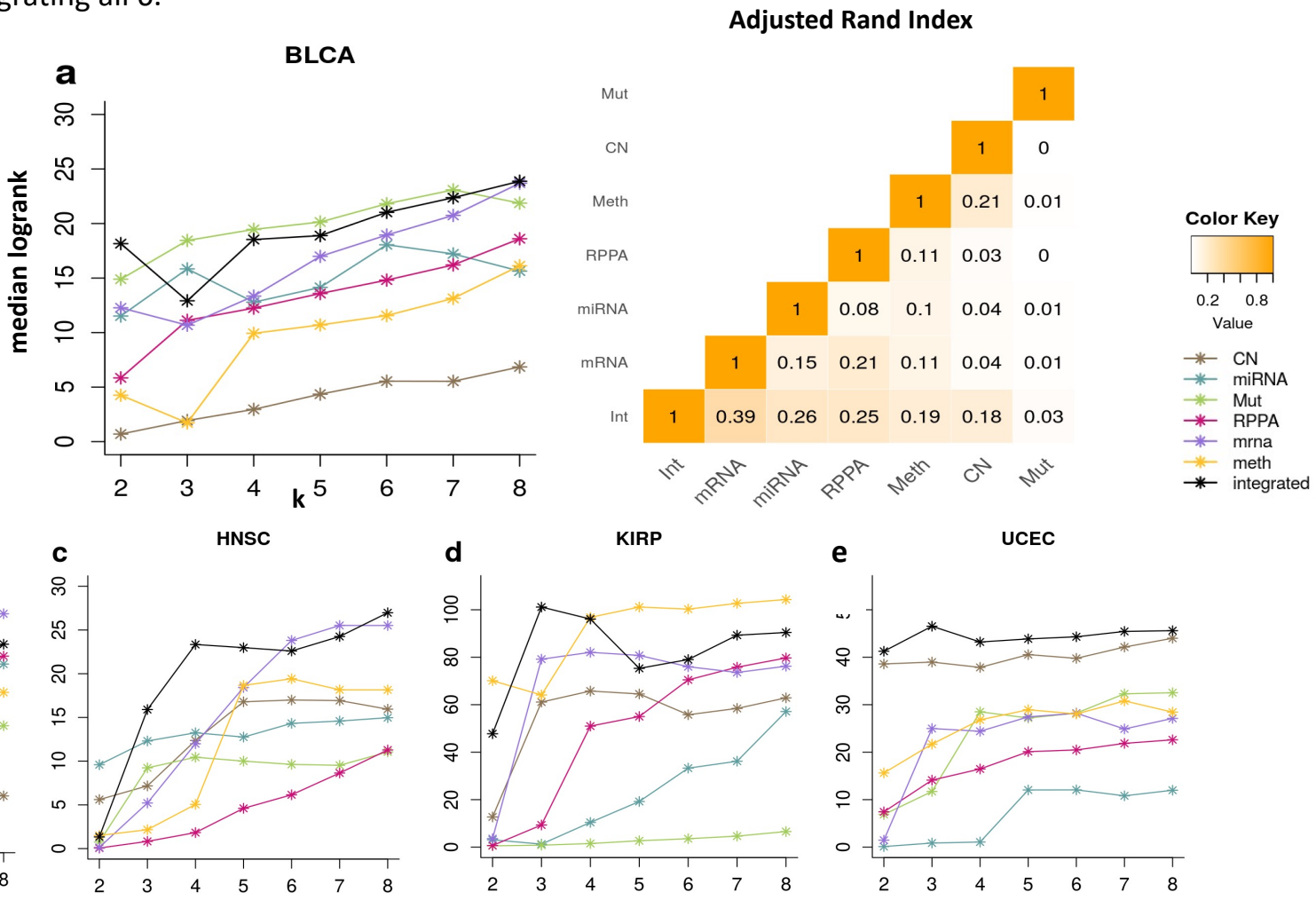


c1 c3 c5
c2 c4 c6

Memorial Sloan Kettering Cancer Center

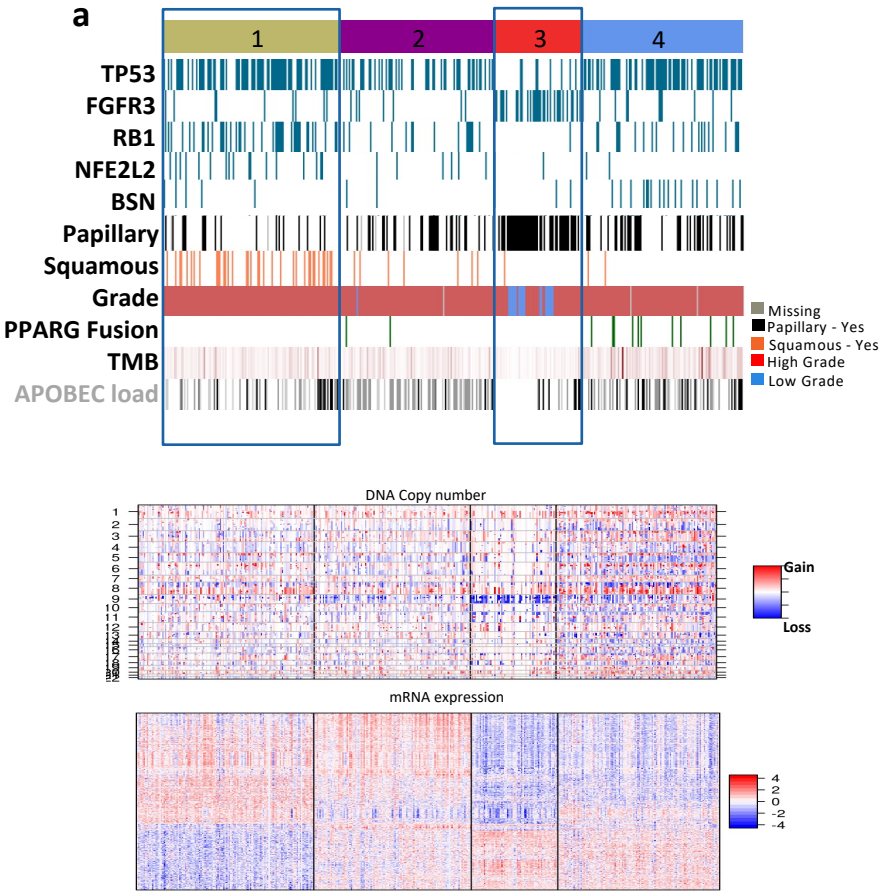
Integrative analysis of multiple platforms

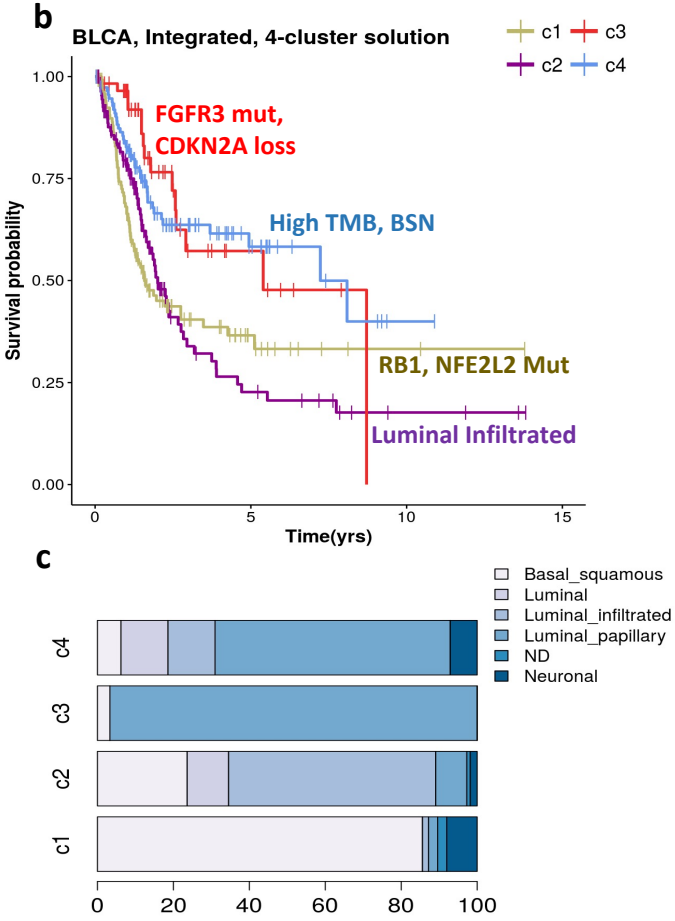
survClust was run on each of the available 6 molecular platforms on each cancer type – Mutation, copy number, DNA Methylation, mRNA expression, miRNA expression and protein assay (RPPA), and integrating all 6.



Memorial Sloan Kettering Cancer Center

Integrated solution identified by survClust on TCGA BLCA cohort





Memorial Sloan Kettering Cancer Center

Conclusion

- Developed a supervised learning approach for survival outcome-weighted molecular stratification

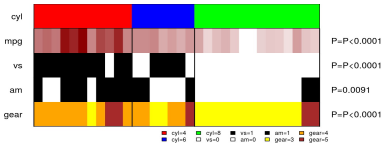
- Application to somatic mutation data led to stratifications associated with mutational burden and hyper-mutation signatures corresponding to distinct mutagenic processes

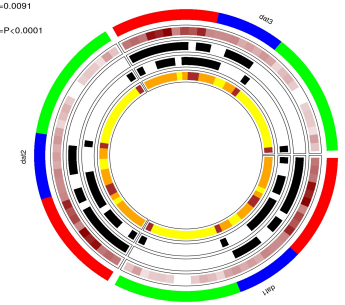
- The integration of multiple data platforms led to more refined outcome stratifications than individual platform derived clustering results in the majority of the cancer types in our analysis

- Developed annotation tools (*circomap, panelmap*) to visualize the association of molecular and clinical information with the subtypes

  R package *panelmap* and function *circomap* –

  found here – https://github.com/arorarshi/panelmap





- *survClust* – developmental version

  Happy to talk! - email – arshiaurora@gmail.com

  check my Github repository when it's published!

Memorial Sloan Kettering Cancer Center

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 classical multi-dimensional 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.
- Gu, Z. circlize implements and enhances circular visualization in R. Bioinformatics 2014

Thank you!