Statistical methods for multi-omic data integration

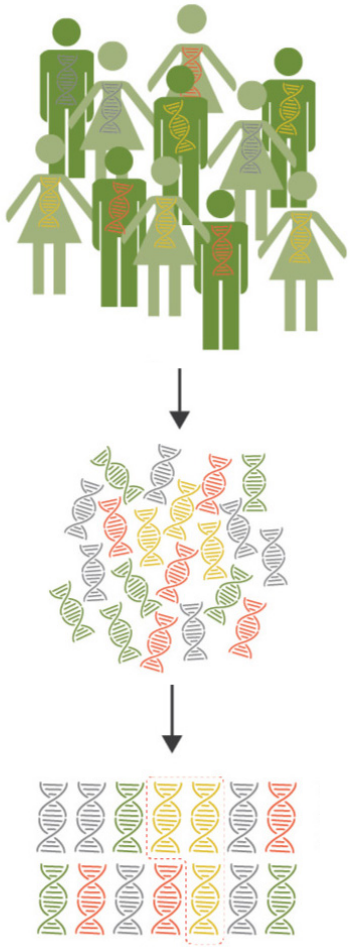
Alessandra Cabassi

ISBA

30 June 2021

Motivation

Clustering within the context of precision medicine.

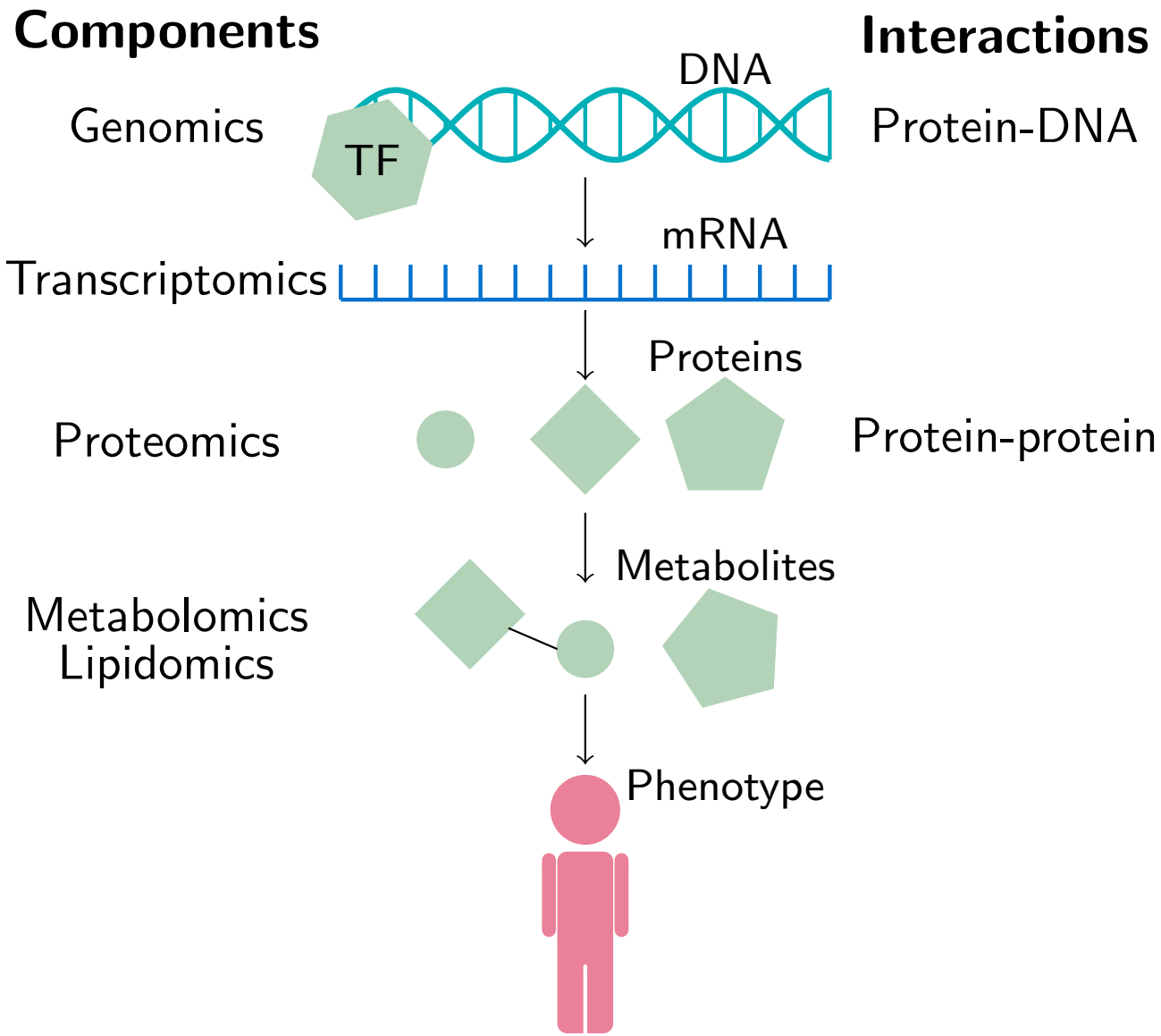


Goal

To use multiple 'omics datasets to find disease subtypes & help clinicians develop more specific treatments.

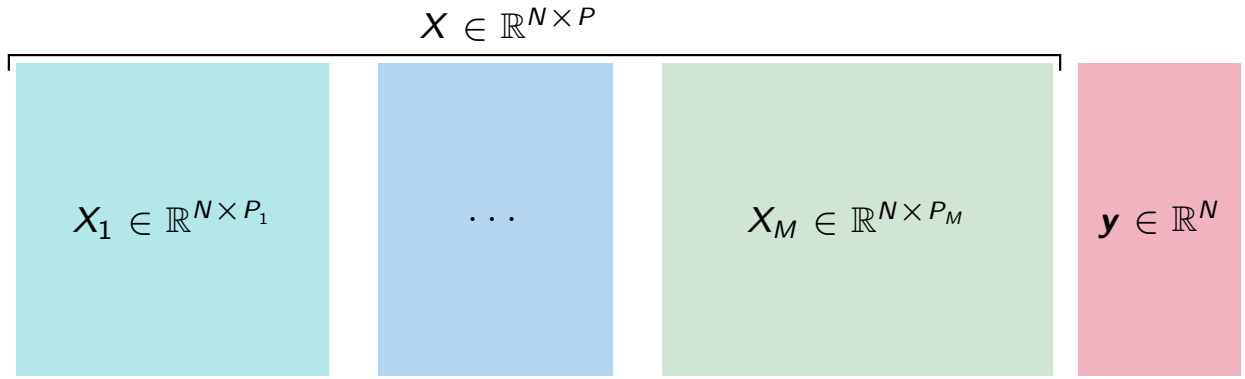
 Illustration: Health Education England's Genomics Education Programme.

Multi-omics


Components
Interactions
DNA
Genomics
TF
Protein-DNA
mRNA
Transcriptomics
Proteins
Proteomics
Protein-protein
Metabolites
Metabolomics
Lipidomics
Phenotype

Statistical setting and challenges

Multiple data sources relative to the same statistical units.

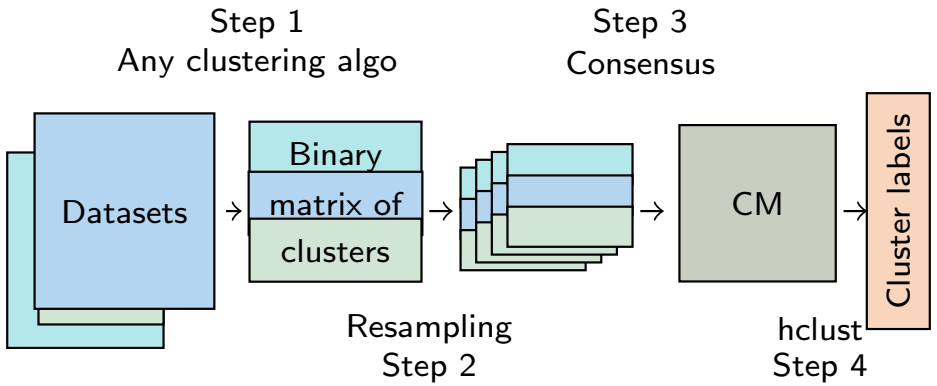


Challenges:

- Different types of data
- Different layer sizes
- Varying levels of noise
- High computational cost
- Large P small N

Cluster-Of-Clusters Analysis

COCA (Cluster-Of-Clusters Analysis)*



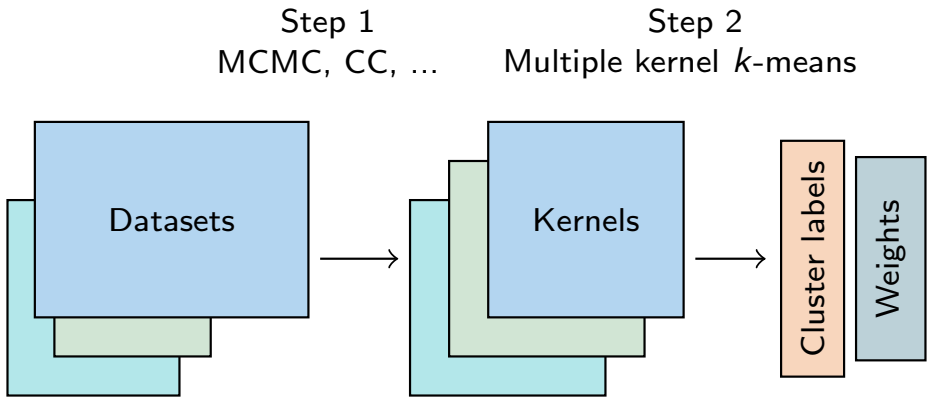
Challenges

- Incorporate uncertainty of cluster allocations.
- Quantify contribution of each data source to the final clustering.

*TCGA (2012). "Comprehensive molecular portraits of human breast tumours." Nature 487.7407, pp. 61–70.

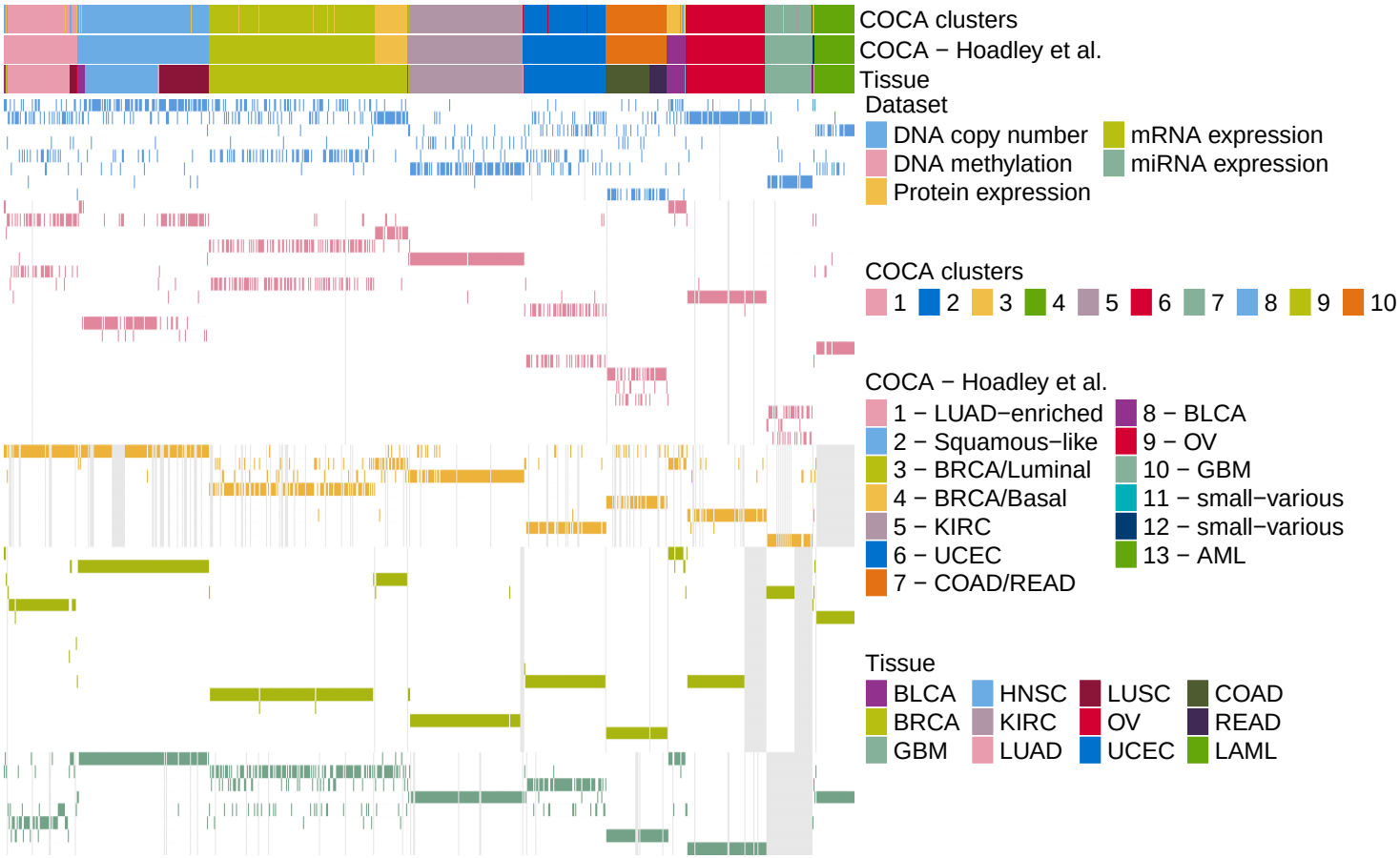
Multiple kernel learning for clustering

Unsupervised KLIC (Kernel Learning Integrative Clustering)



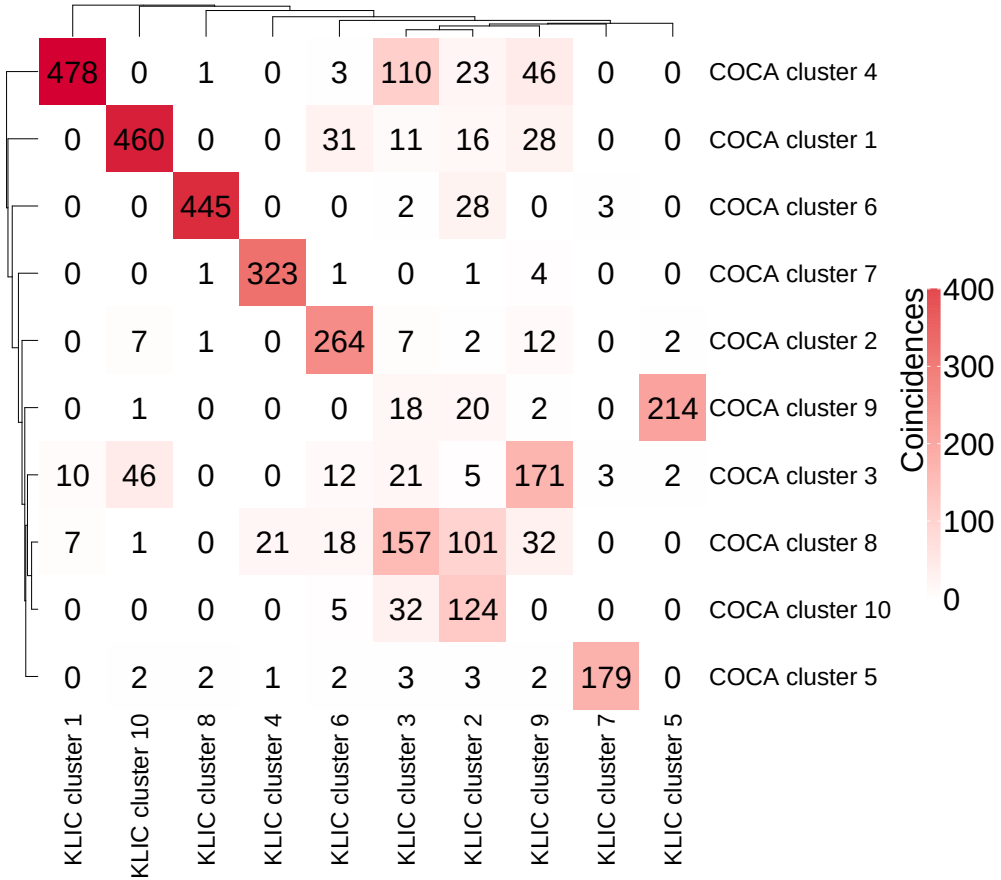
Multiplatform analysis of 12 cancer types

Data: 5 'omic layers, 12 tumour types, 3,527 patients.
Goal: identify cancer subtypes/patients w/ similar molecular profiles.



Hoadley et al. (2014). "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin". Cell 158.4, pp. 929–944.

Multiplatform analysis of 12 cancer types



Challenge

Extract the most relevant clustering structure.

Summarising/combining PSMs

Posterior similarity matrices are valid kernels.

How to:

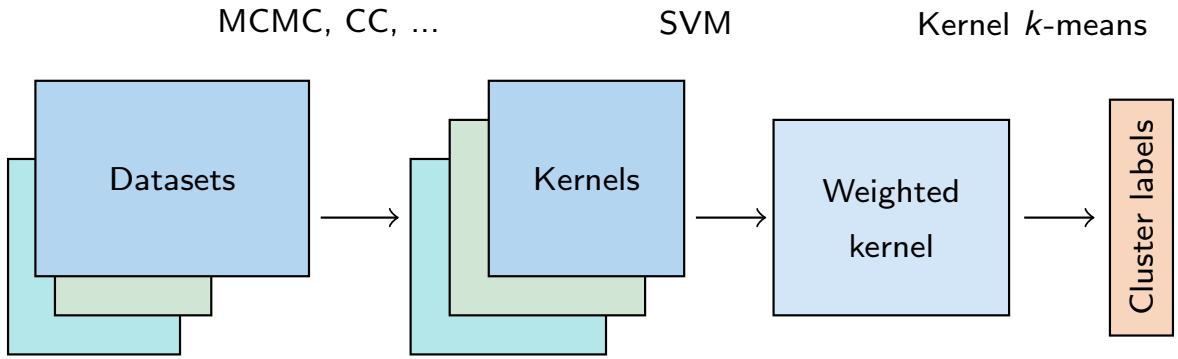
- summarise PSMs? $\rightarrow$ Kernel k-means.
- combine multiple PSMs? $\rightarrow$ Unsupervised KLIC.
- find most relevant clustering structure? $\rightarrow$ Outcome-guided KLIC.

Outcome-guided KLIC

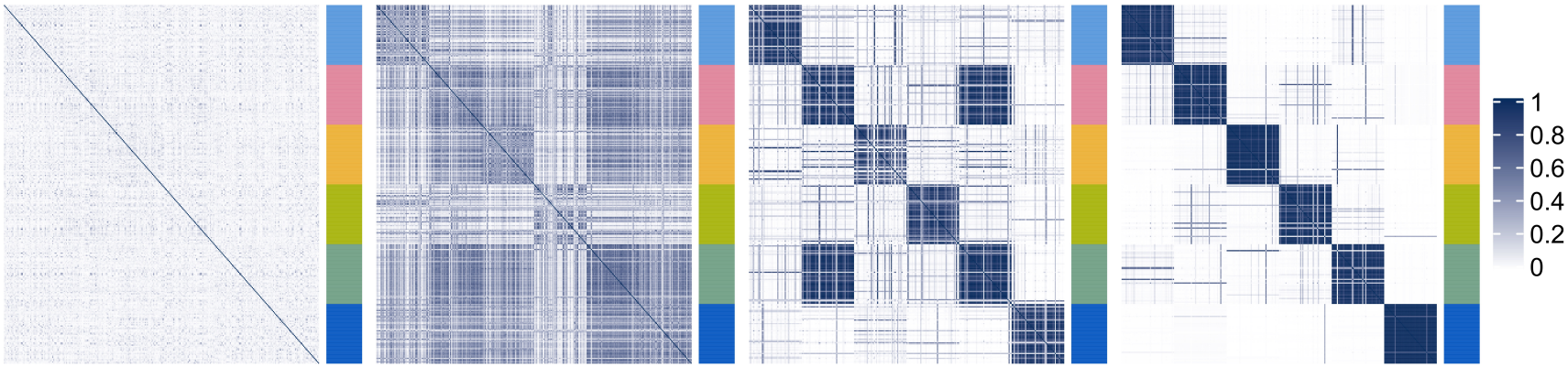
Goal

Use a variable related to the outcome of interest to determine the kernel weights.

Outcome-guided KLIC (Kernel Learning Integrative Clustering)

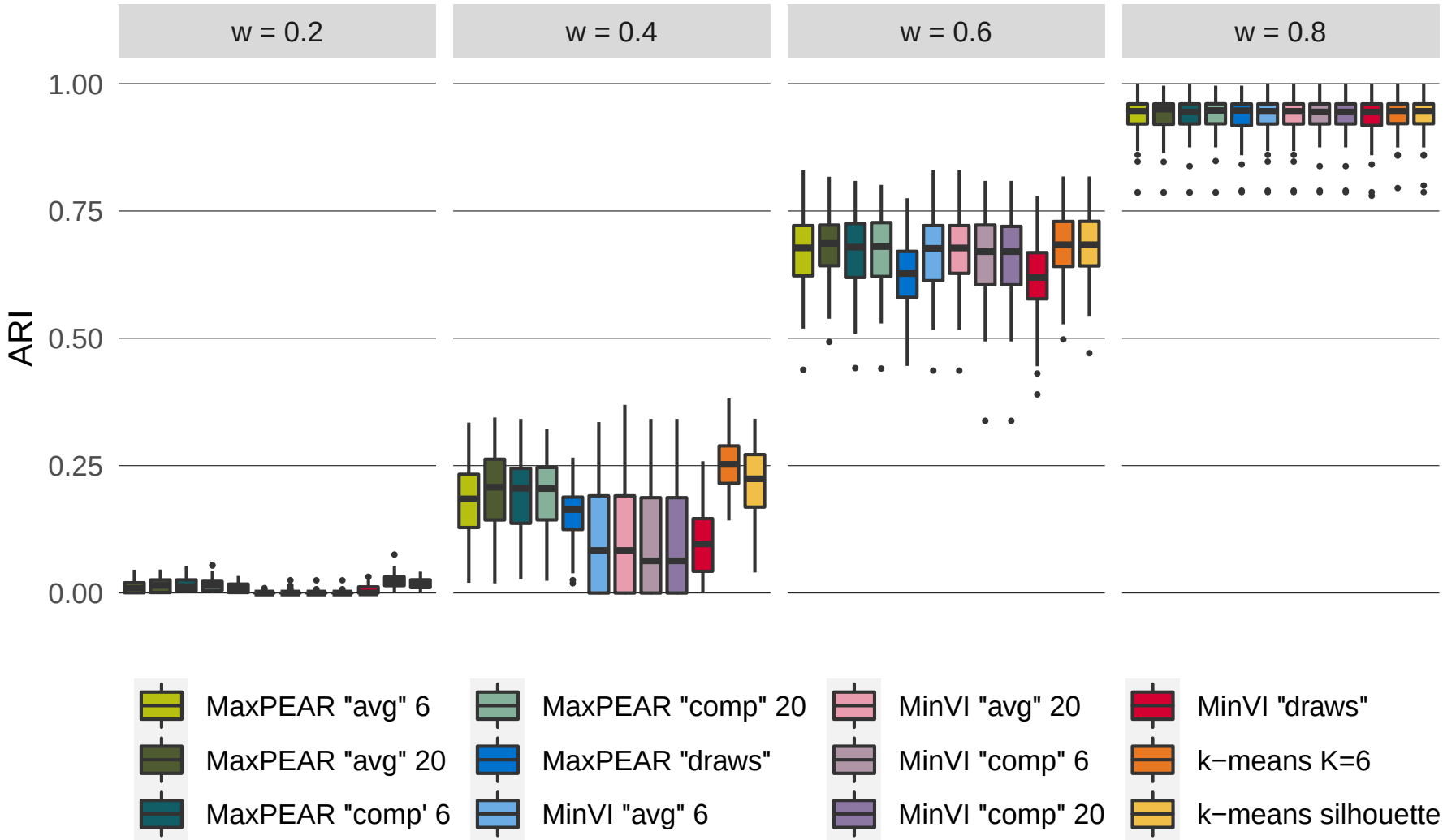


Simulation studies

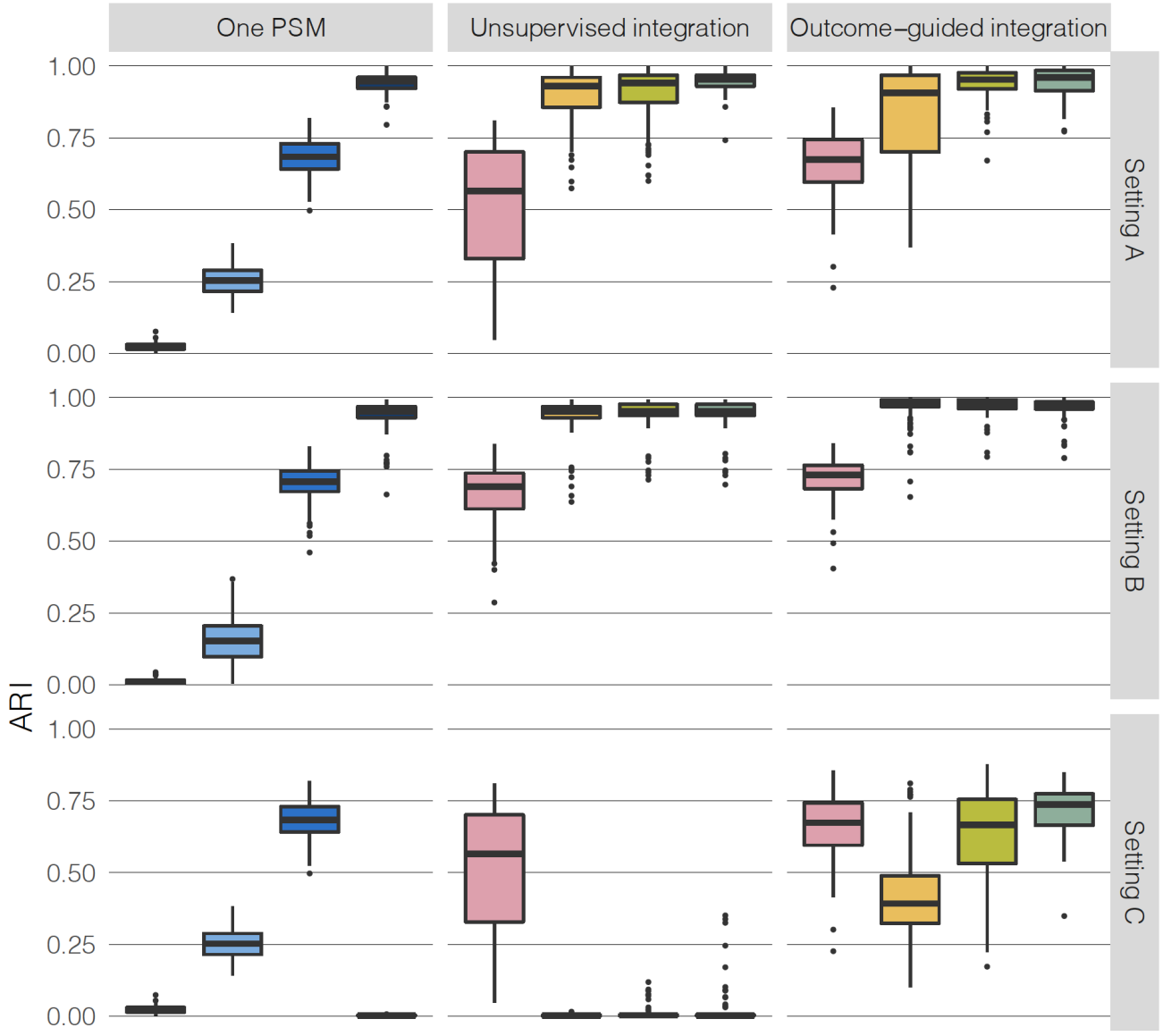


PSMs of 4 datasets with different cluster separability.
Row & columns = statistical units.
Coloured bar on the right = true clustering.

Summarising posterior similarity matrices

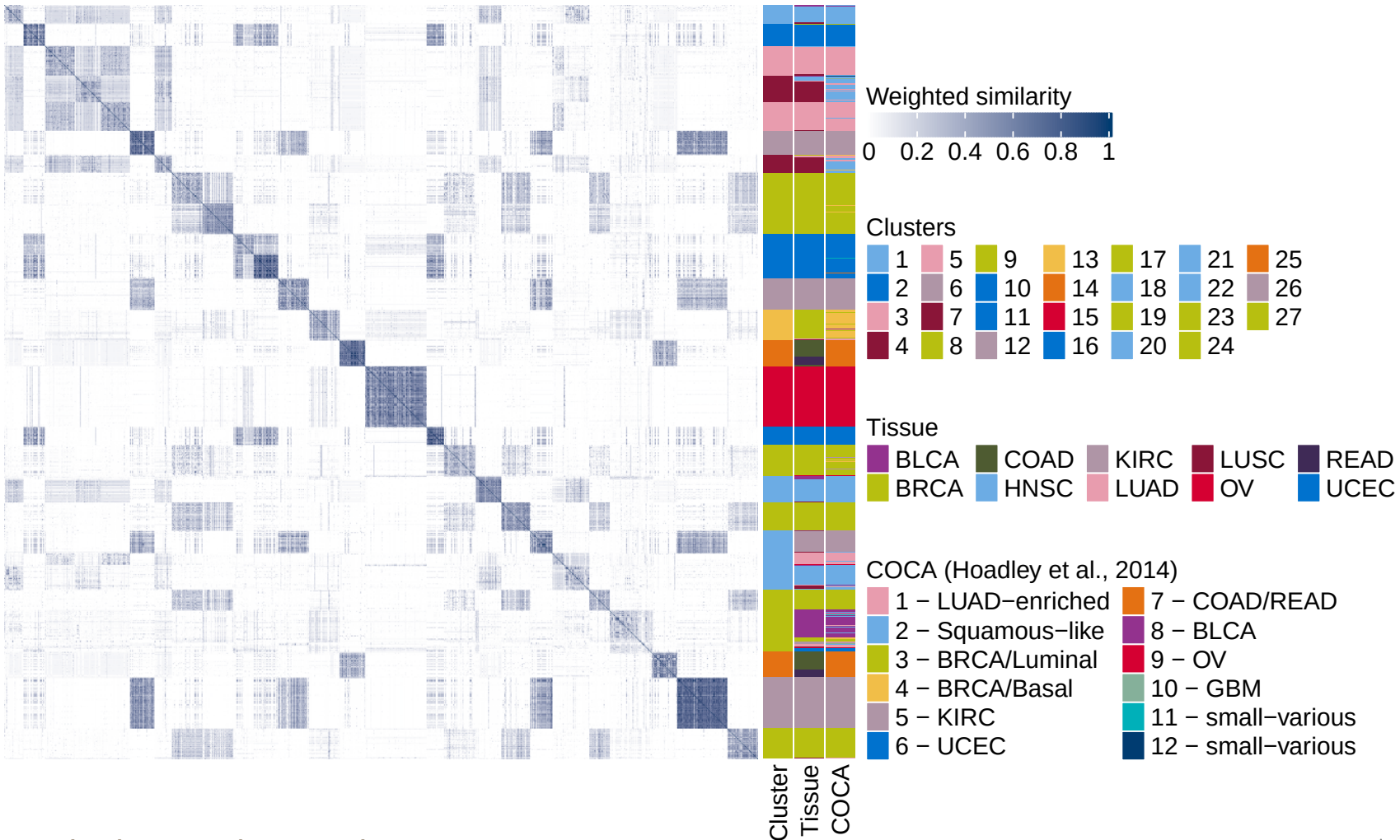


Integrative clustering

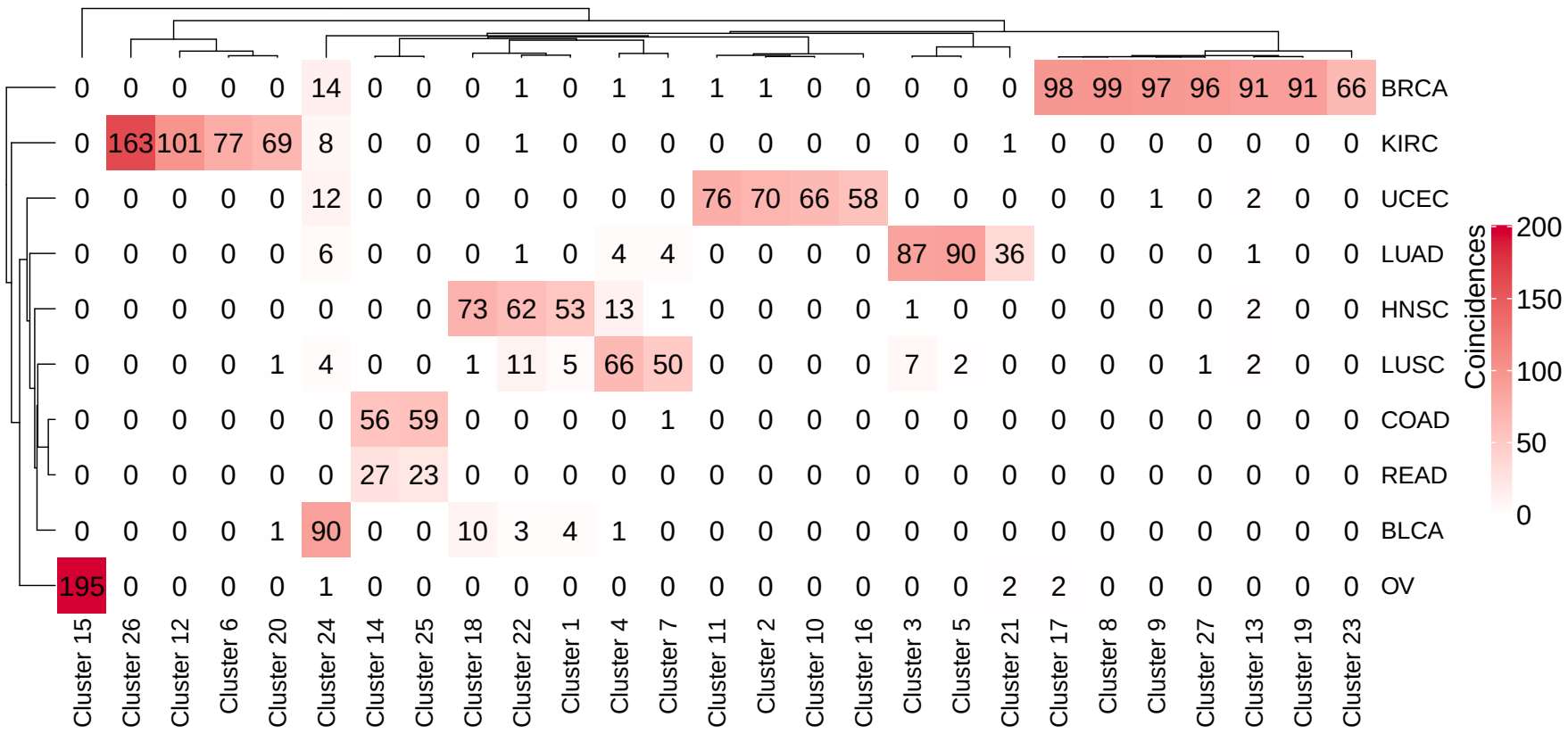


Multiplatform analysis of 10 cancer types

- Unsupervised KLIC: 9 clusters.
- Outcome-guided KLIC: 27 clusters.



Multiplatform analysis of 10 cancer types



Main findings and contributions

- Common to all approaches:
  - Ability to monitor the influence of each layer on the final outcome.
  - Low computational cost thanks to two-step approaches.

Main findings and contributions

- Common to all approaches:
  - Ability to monitor the influence of each layer on the final outcome.
  - Low computational cost thanks to two-step approaches.
- Unsupervised integration
  - Down-weight noisy datasets and noisy variables.
  - Ability to define kernels based on heuristic/model-based clustering.
  - Ability to combine multiple PSMs.

Main findings and contributions

- Common to all approaches:
  - Ability to monitor the influence of each layer on the final outcome.
  - Low computational cost thanks to two-step approaches.
- Unsupervised integration
  - Down-weight noisy datasets and noisy variables.
  - Ability to define kernels based on heuristic/model-based clustering.
  - Ability to combine multiple PSMs.
- Outcome-guided integration
  - All of the above.
  - Up-weight most relevant layers.
  - Ability to uncover more refined partitions of the data.

Main findings and contributions

- Common to all approaches:
  - Ability to monitor the influence of each layer on the final outcome.
  - Low computational cost thanks to two-step approaches.
- Unsupervised integration
  - Down-weight noisy datasets and noisy variables.
  - Ability to define kernels based on heuristic/model-based clustering.
  - Ability to combine multiple PSMs.
- Outcome-guided integration
  - All of the above.
  - Up-weight most relevant layers.
  - Ability to uncover more refined partitions of the data.
- Real data applications
  - ! Importance of variable selection.
  - ! Need to deal with missing values.

Further research areas

- Model extensions
  - Handling missing values (for regression models and DPMMs)
  - Handling continuous and survival outcomes
- Evaluation and comparison of clustering results
  - Assessment of the similarity of two partitions
  - Choice of the number of clusters
  - Assessment of cluster quality

Publications

- **Cabassi, A.** and Kirk, P.D.W. (2020). "Multiple kernel learning for integrative consensus clustering of 'omic datasets"
  Bioinformatics, btaa593.
- R packages `klic` and `coca` available on CRAN.
- **Cabassi, A.**, Richardson, S., and Kirk, P.D.W. (2020). "Kernel learning approaches for summarising and combining posterior similarity matrices"
  arχiv preprint, 2009.12852.

Thanks for listening!

MRC | Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

alessandra.cabassi@mrc-bsu.cam.ac.uk

alessandracabassi.com

@SandyCabassi