Comparing different multi-omics data integration tools

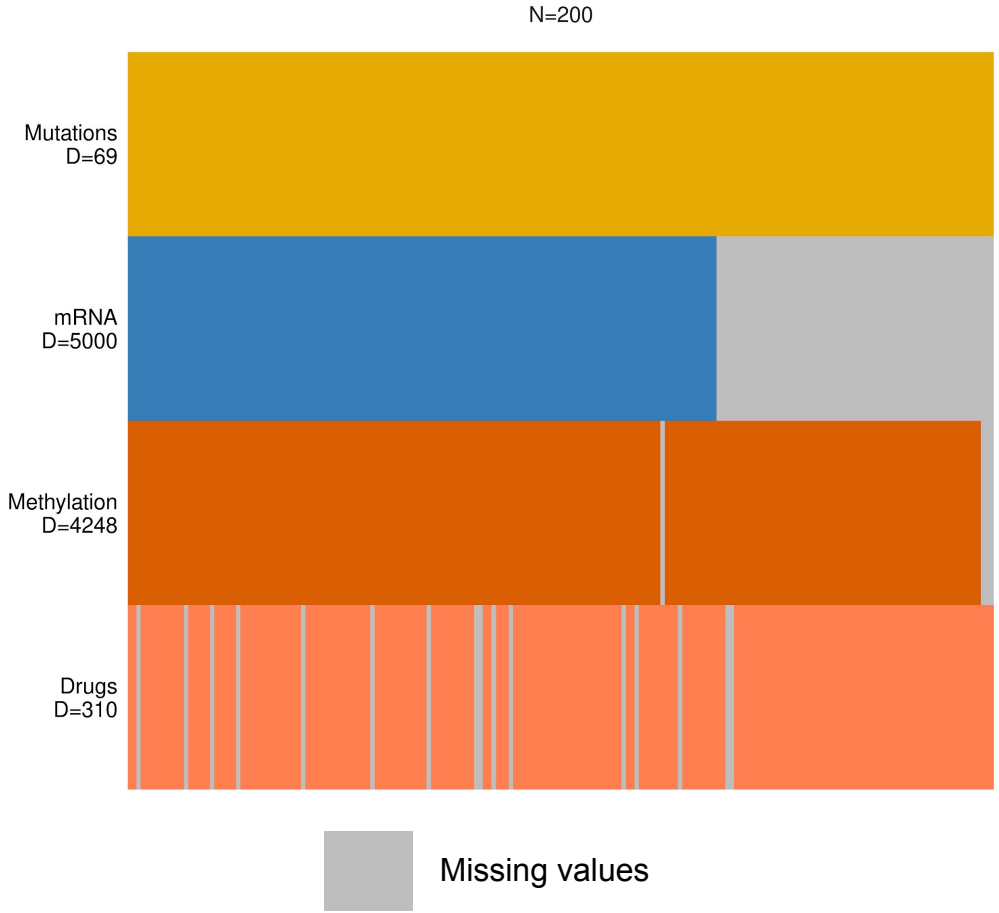
Mentor: Manik Garg

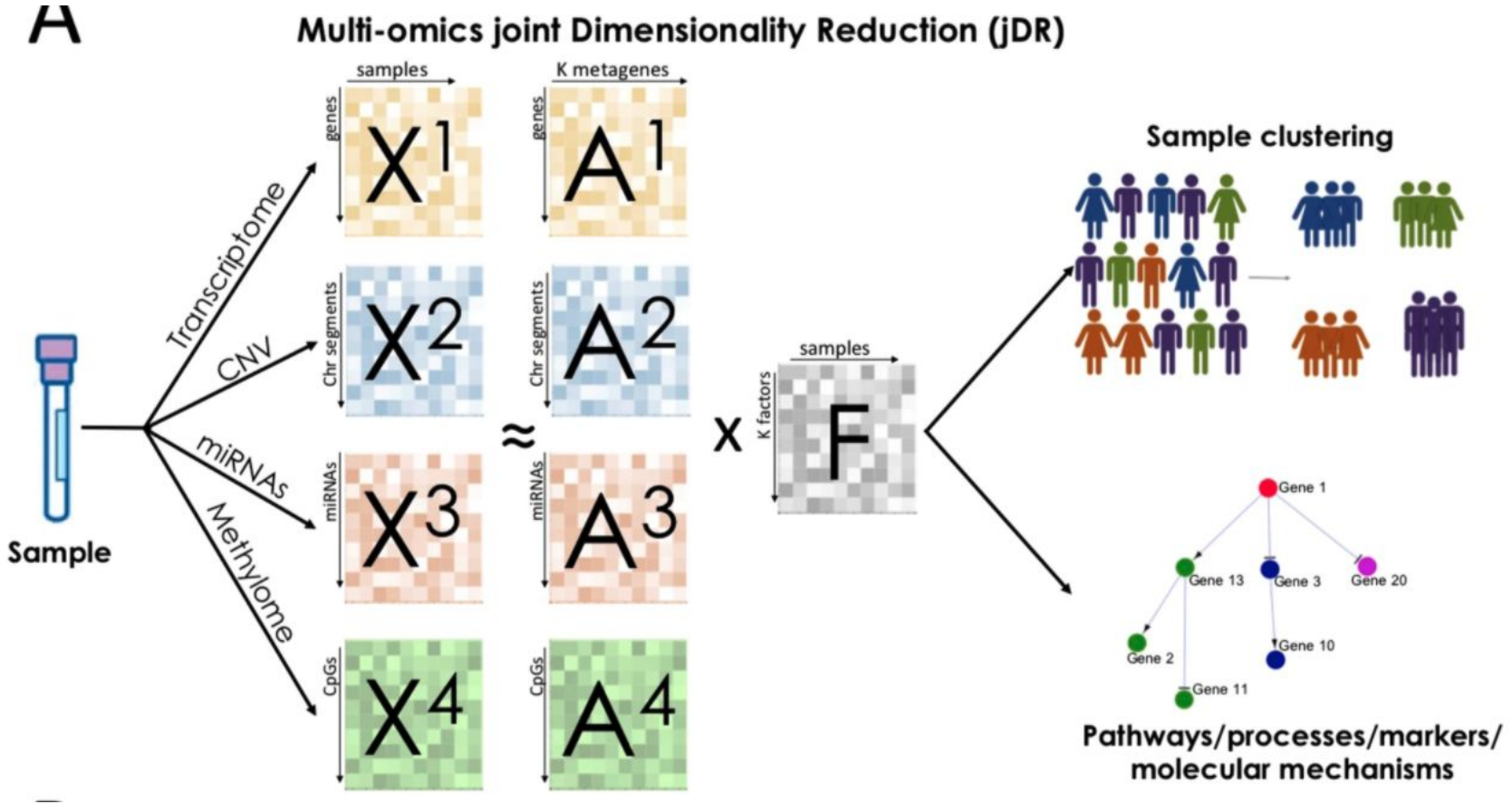
Group members: Aurora Savino, Eirini Christodoulaki, Jonas Hagenberg, Ngoc Nguyen, Savvas Kourtis, Tu Hu, Marta Popeda and Cláudia Raposo de Magalhães

Aim of the project: benchmark three multi-omics data-integration methods:

- Multi-omics factor analysis (MOFA)
- Joint and Individual Variation Explained (JIVE)
- Multiple co-inertia analysis (MCIA)

On a multiomics data set including 3 data modalities: DNA methylation, RNA-seq, somatic mutations and drug response data from blood for N=200 patients with Chronic Lymphocytic Leukemia (CLL) from (Dietrich *et al*, 2018)




A
Multi-omics joint Dimensionality Reduction (jDR)
samples
K metagenes
genes
X1
A1
Transcriptome
CNV
Chr segments
X2
A2
miRNAs
X3
A3
Methylome
CpGs
X4
A4
≈
Sample
X
samples
K factors
F
Sample clustering
Gene 1
Gene 13
Gene 3
Gene 20
Gene 2
Gene 10
Gene 11
Pathways/processes/markers/
molecular mechanisms

Benchmarking methods

- Different imputation methods
- Do multiomics methods capture factors that can’t be retrieved with single modalities?
- Performance in clustering samples based on IGHV and trisomy12 status
- Ability to capture clinical variations / biological functions
- Selectivity in capturing clinical variations / biological functions
- Ability to capture survival
- Validation with an independent dataset
- Integration with PPI networks

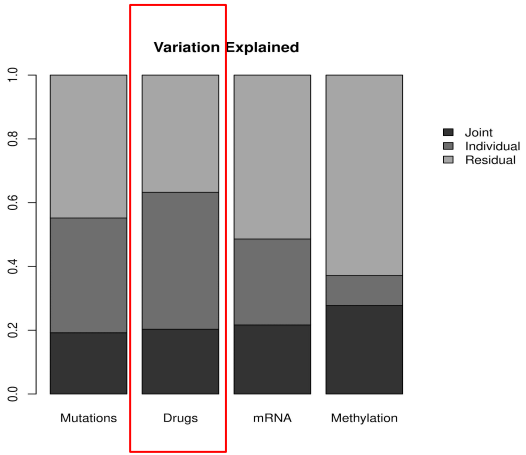
Effect of the imputation method

Missing data is a pervasive problem in biomedical research. Common approaches to face the problem: excluding them, mean/median value imputation, Model-based imputation

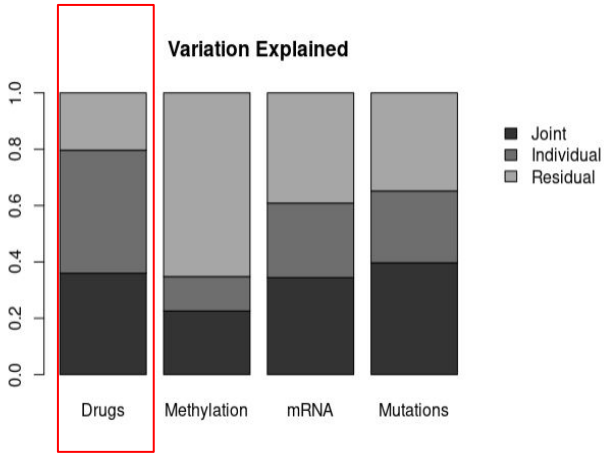
- Replacing missing values with mean/median gives the same results (cons: distorts the histogram and underestimate the variance).

- Model-Based Imputation (Regression, Bayesian) / Proper Multiple Stochastic Regression: Better than mean/median but require a threshold of missingness.
  samples with 100% NA values in CLL data can not be imputed with any regression method

- Future directions: An approach that accounts for missing values needs to be developed

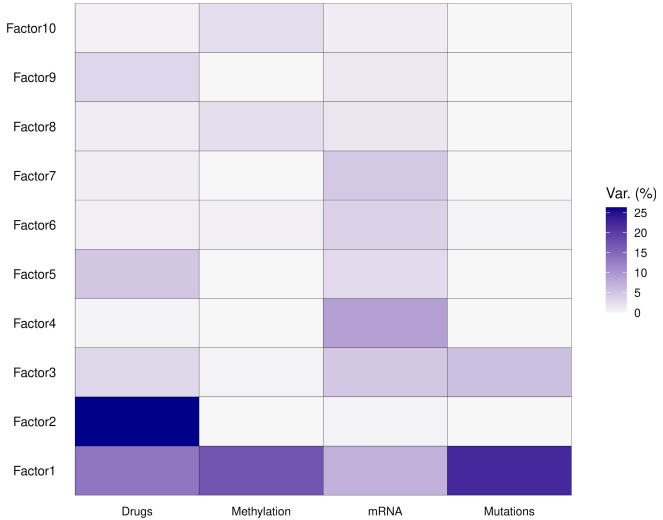


Mean vs regression Imputation

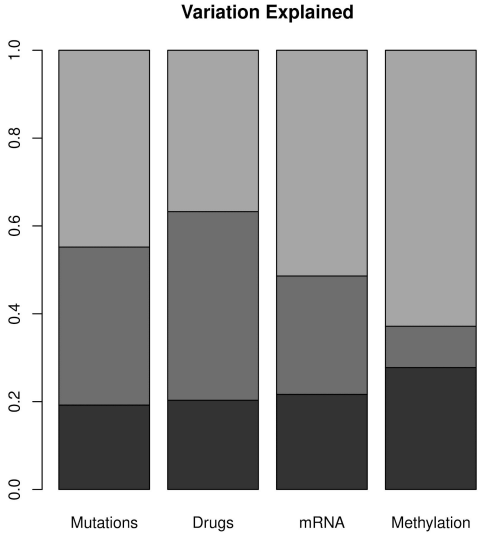


Total variance explained per data modality

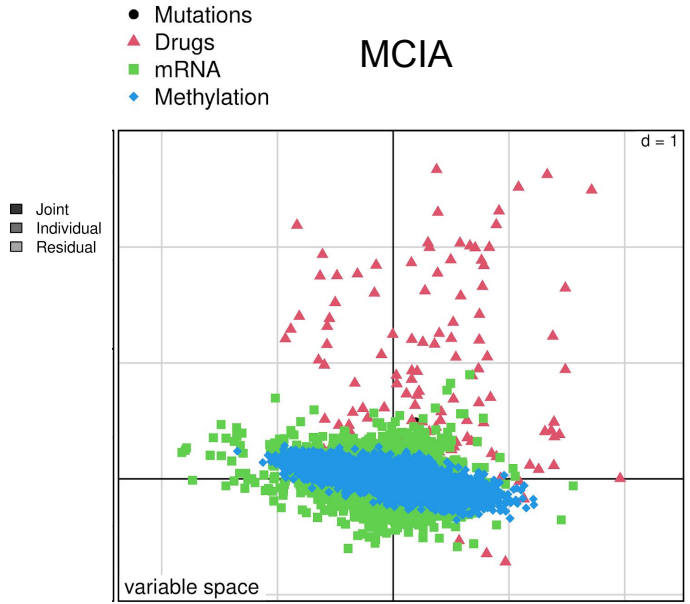
MOFA



JIVE

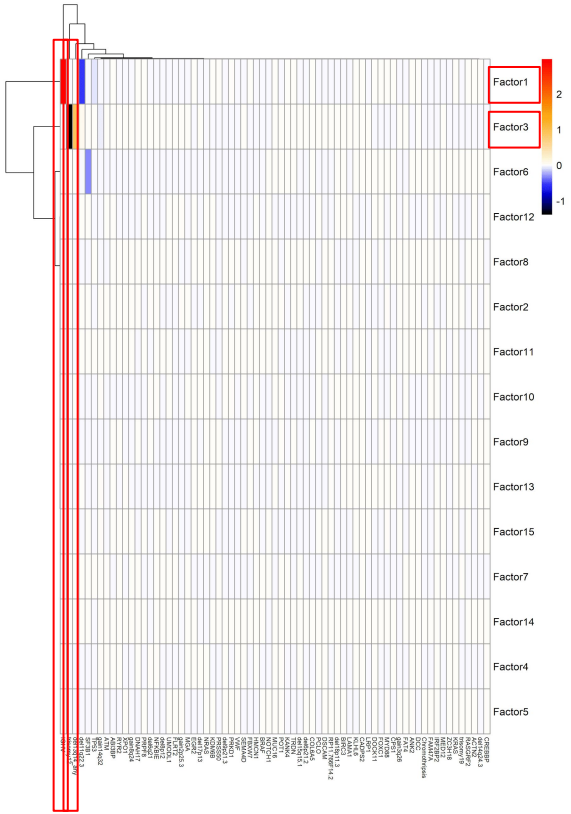


MCIA

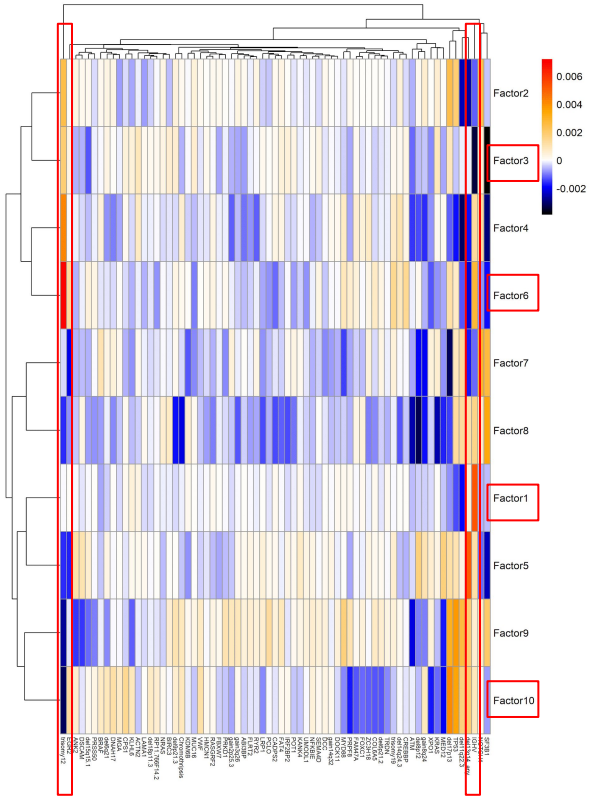


Weights given to each mutation by all the factors for the each method

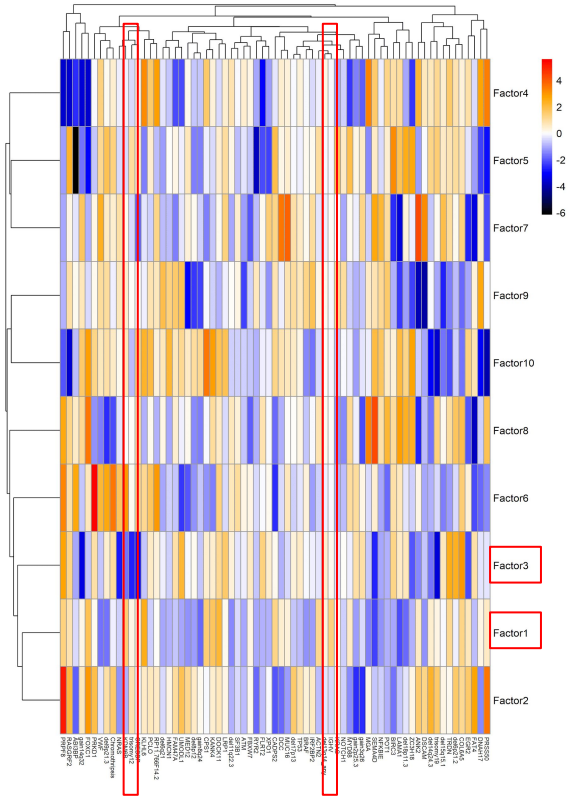
MOFA



JIVE

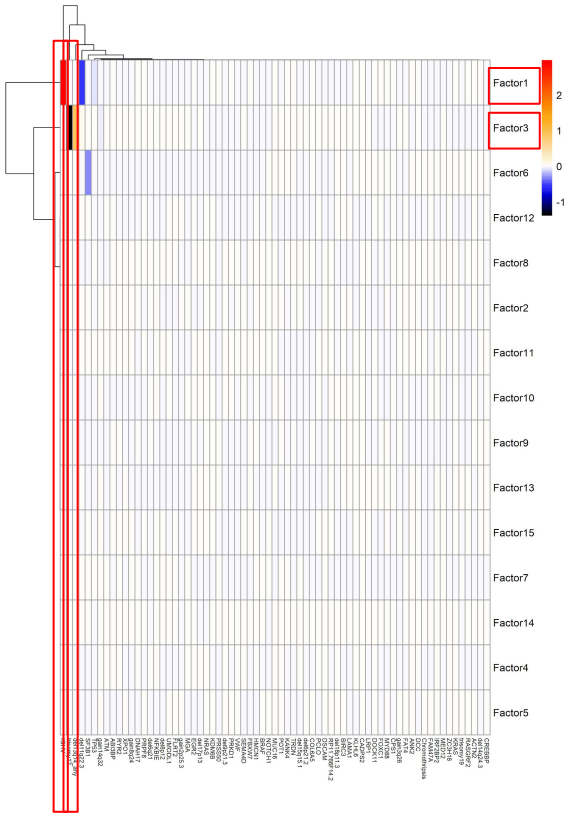


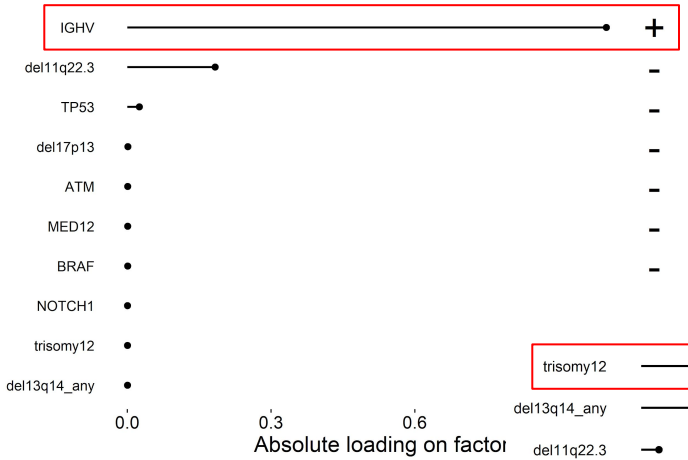
MCIA

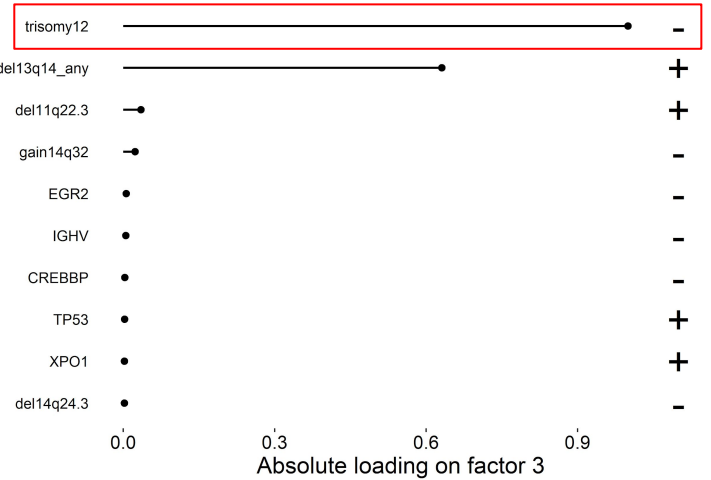


Top 10 weighted mutations by a specific factor for the each method

MOFA

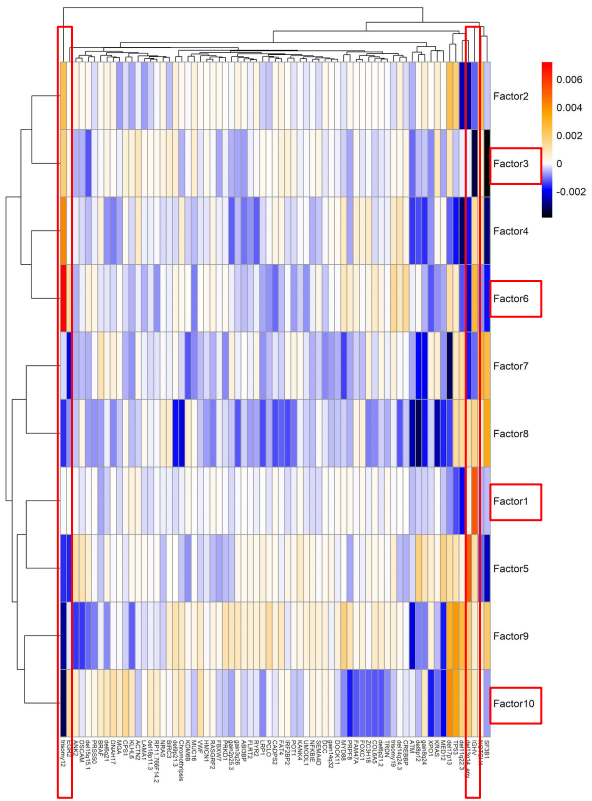


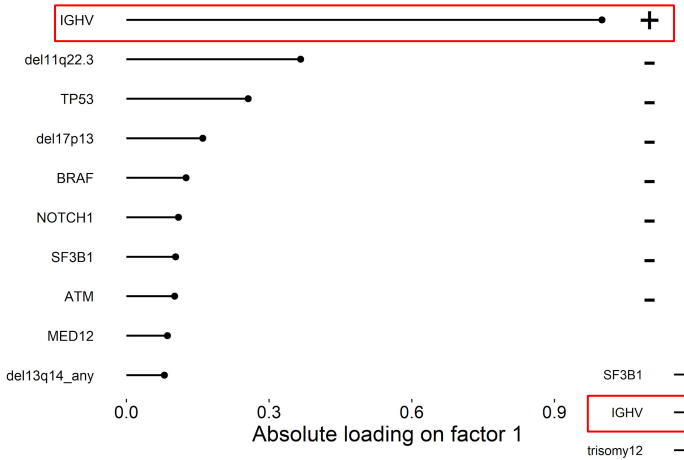


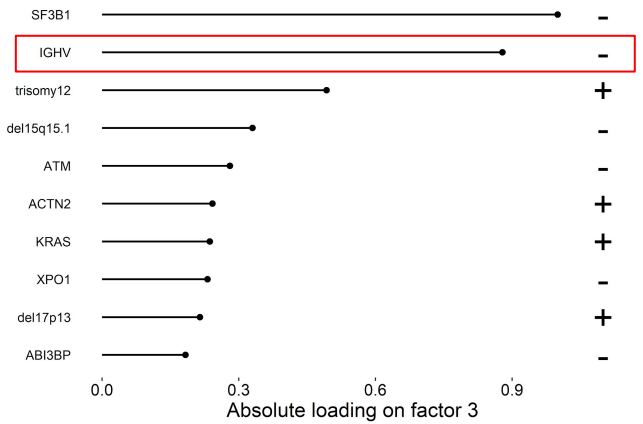


Top 10 weighted mutations by a specific factor for the each method

JIVE



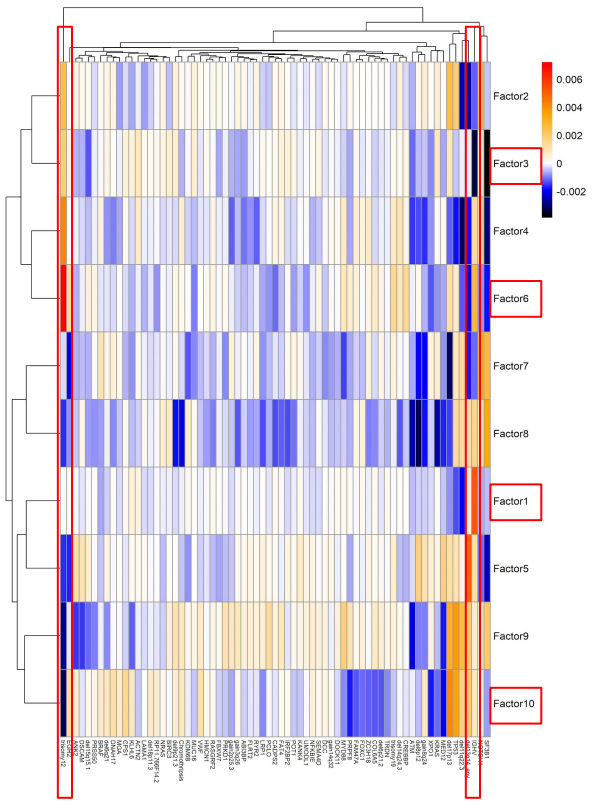


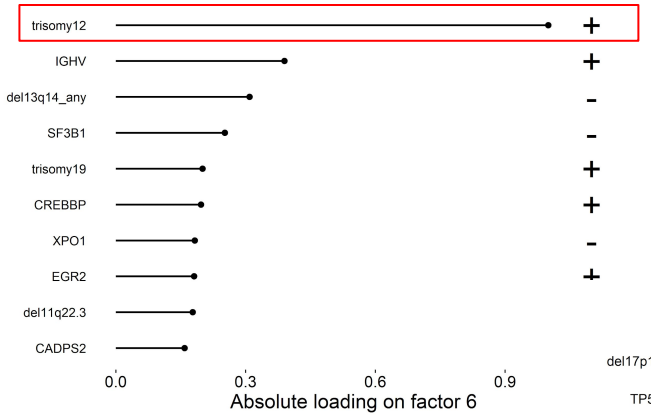


>1 factor is giving the highest weights to **IGHV** status and trisomy12, meaning that the factors are capturing similar clinical information and are not exclusive to each other

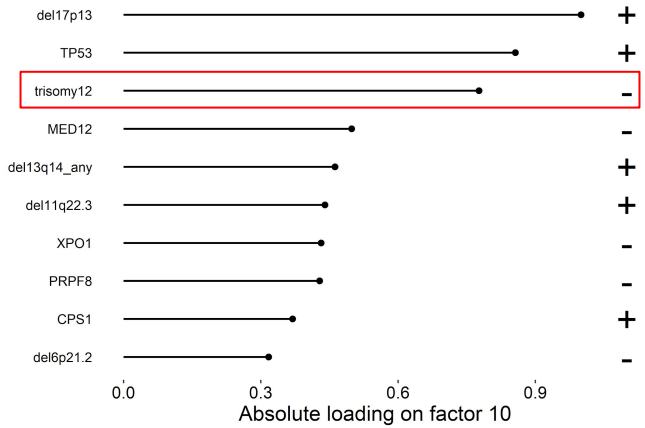
Top 10 weighted mutations by a specific factor for the each method

JIVE



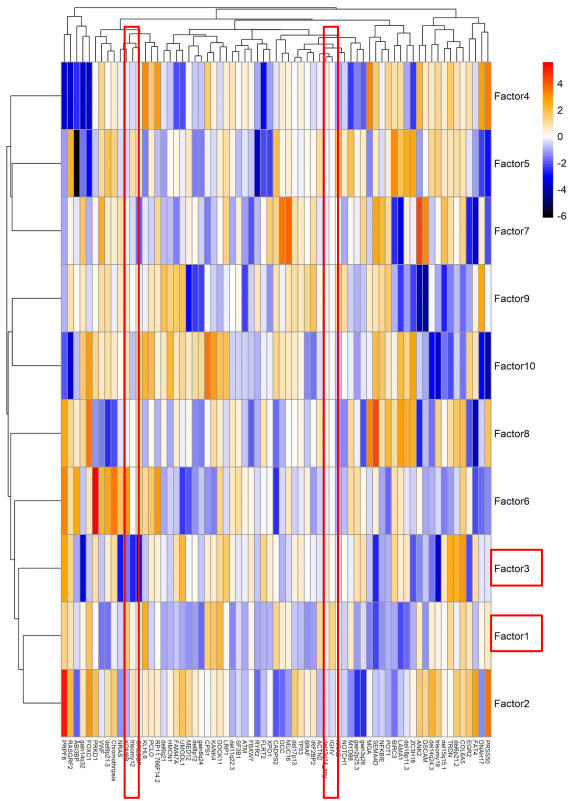


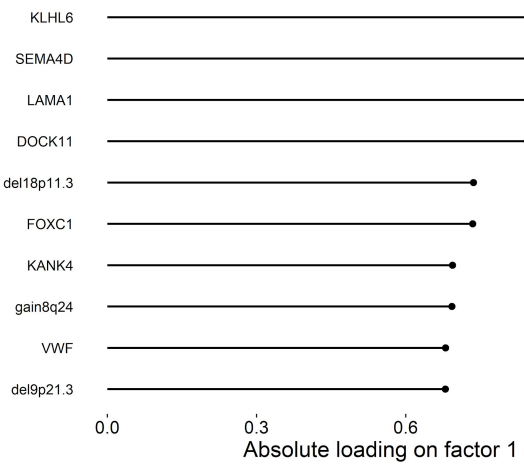
>1 factor is giving the highest weights to IGHV status and **trisomy12**, meaning that the factors are capturing similar clinical information and are not exclusive to each other

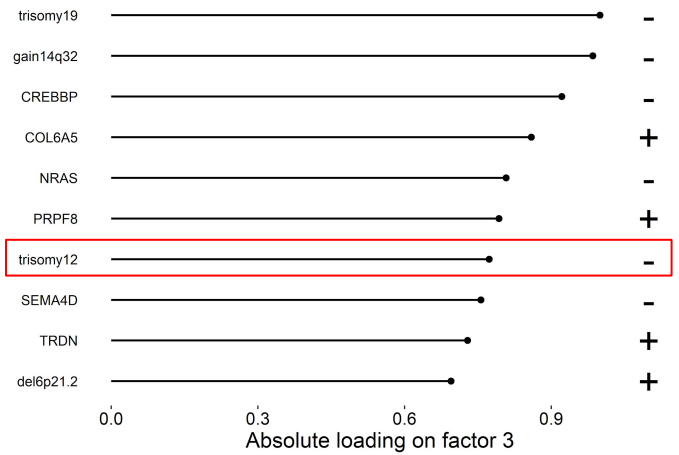


Top 10 weighted mutations by a specific factor for the each method

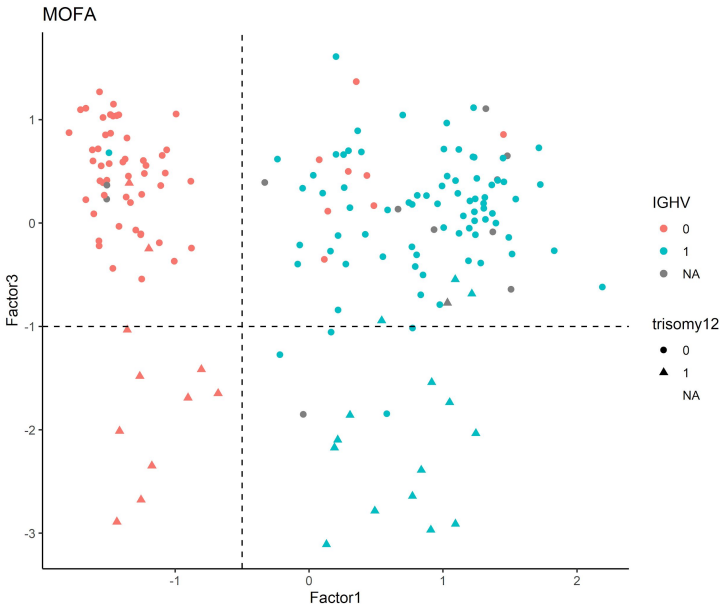
MCIA

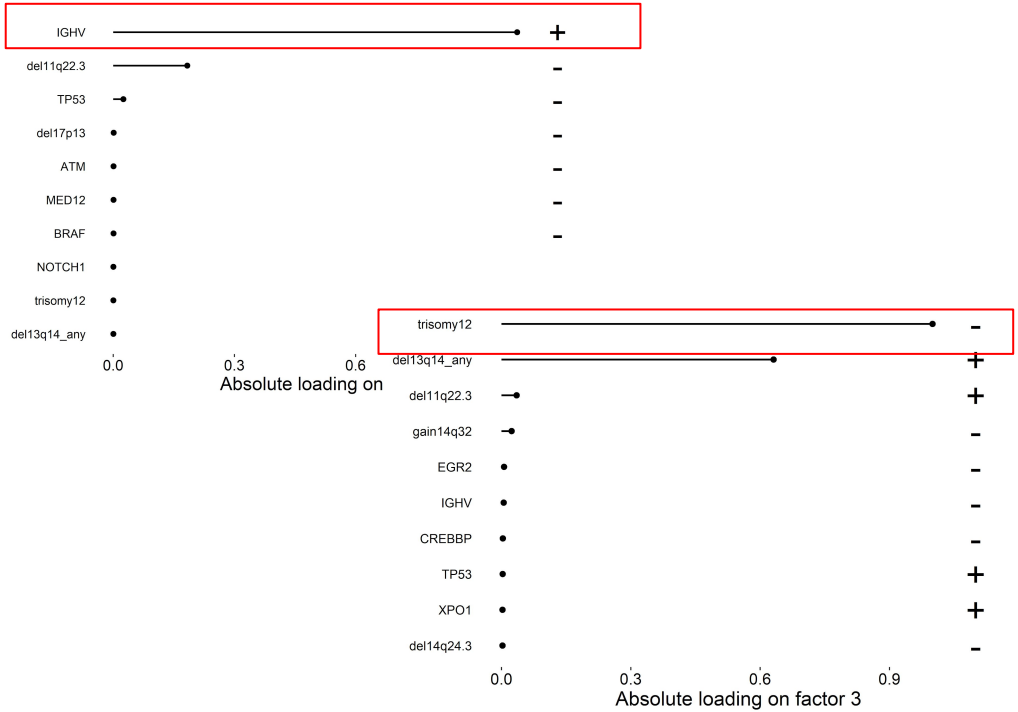




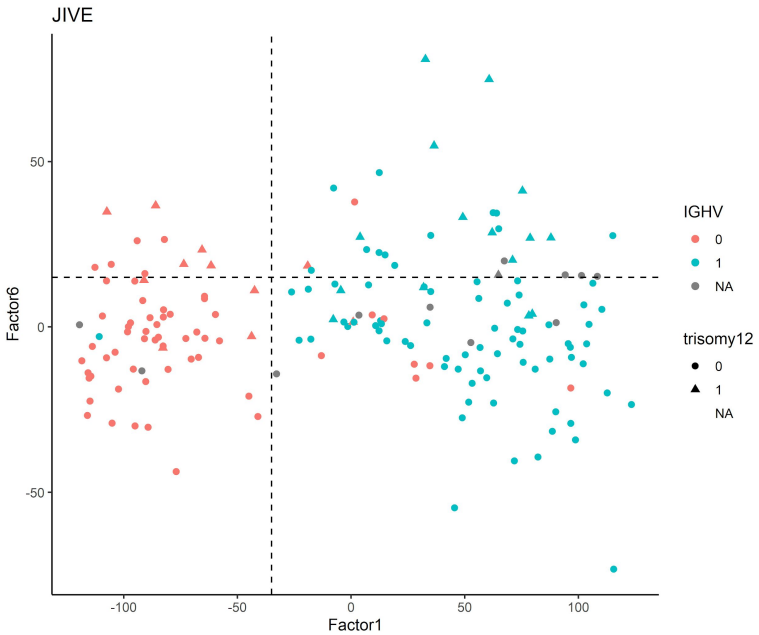


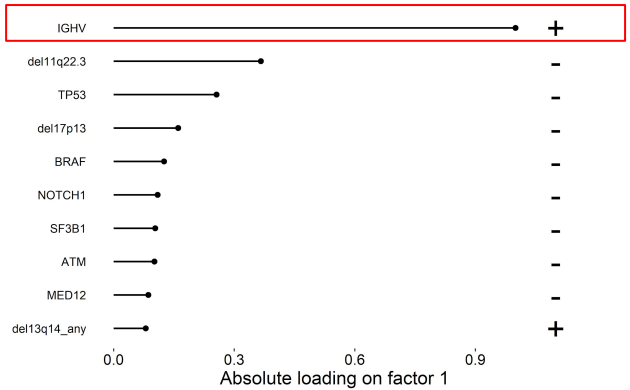
Sample scatter-plot per method: separation between the patients having IGHV status and trisomy12 mutations





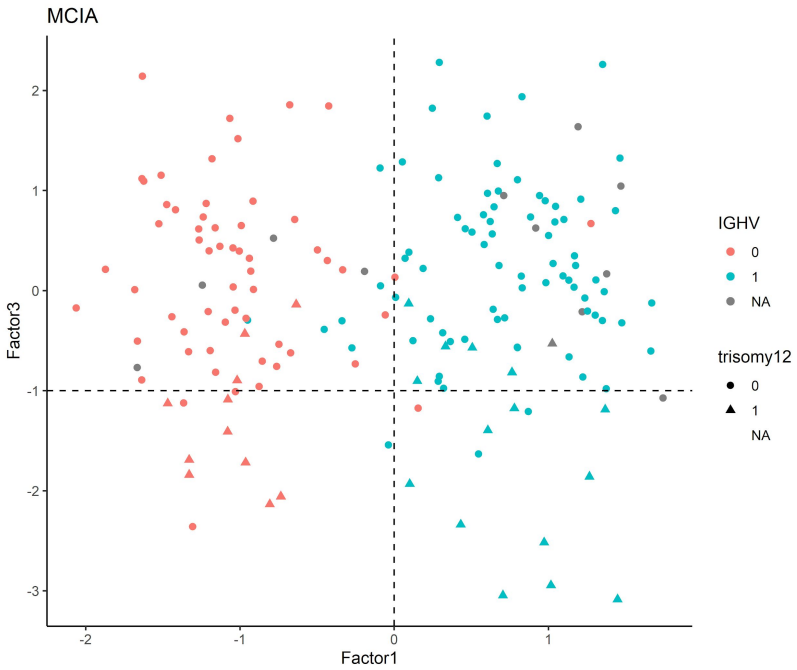
Sample scatter-plot per method: separation between the patients having IGHV status and trisomy12 mutations

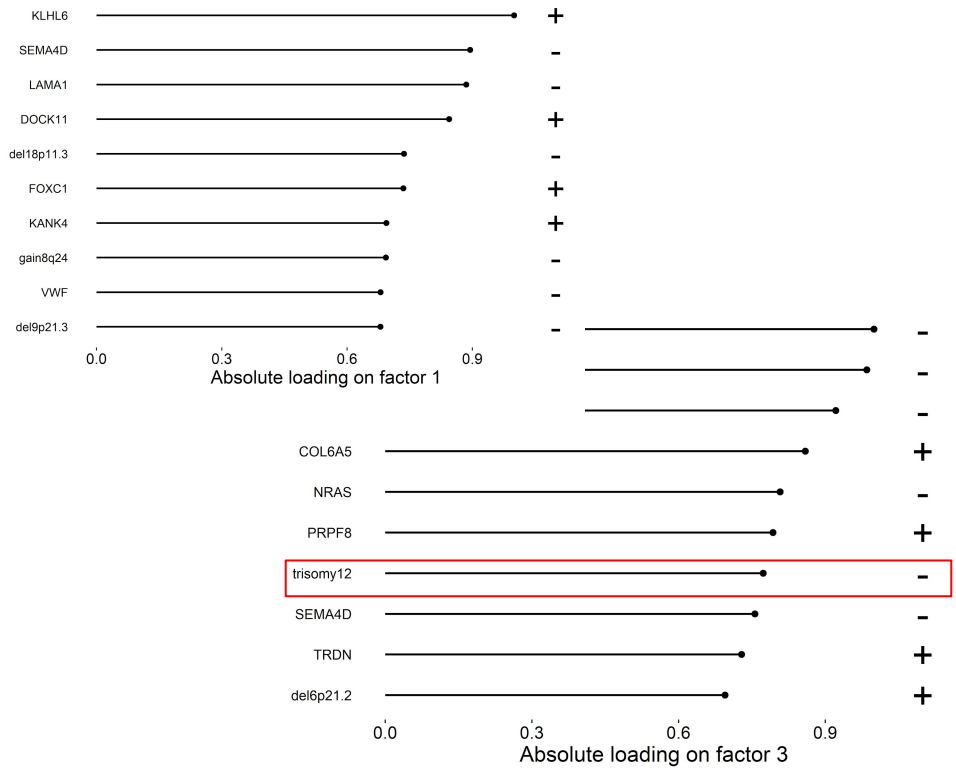




>1 factor is giving the highest weights to **IGHV** status and trisomy12, meaning that the factors are capturing similar clinical information and are not exclusive to each other

Sample scatter-plot per method: separation between the patients having IGHV status and trisomy12 mutations





Selectivity analysis for each method

To quantify the independence of each factor in representing biological information, a selectivity score was devised by Laura et al. (2020) given by:

$$Selectivity\ Score = \left(\frac{N_c + N_f}{2L}\right)$$

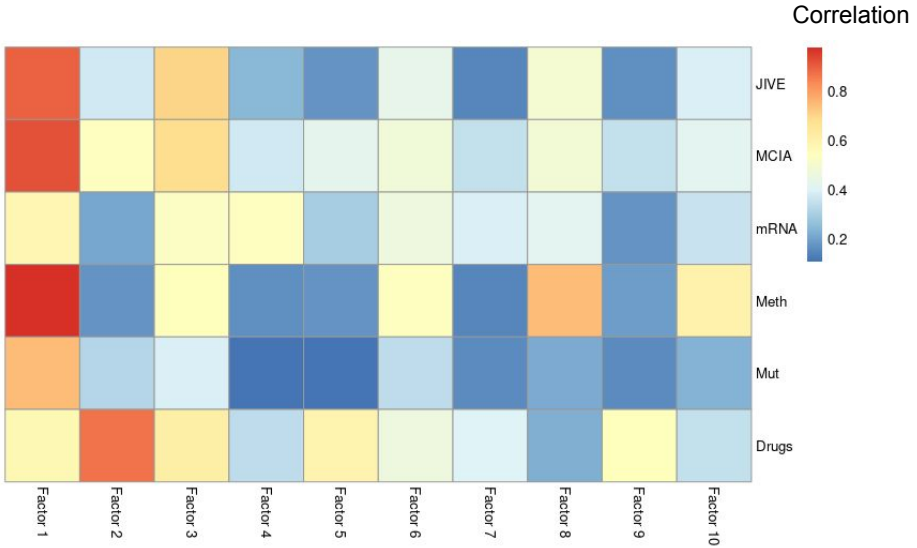
	Selectivity	nonZeroFacs	total_annotations
MOFA	0.3181818	6	8
JIVE	0.3035714	9	8
MCIA	0.2727273	10	8

Comparison of factors

Question:
Do multi-omics approaches retrieve factors that are not retrieved by single modalities?

Comparison of factors identified by each method

MOFA factors vs other methods



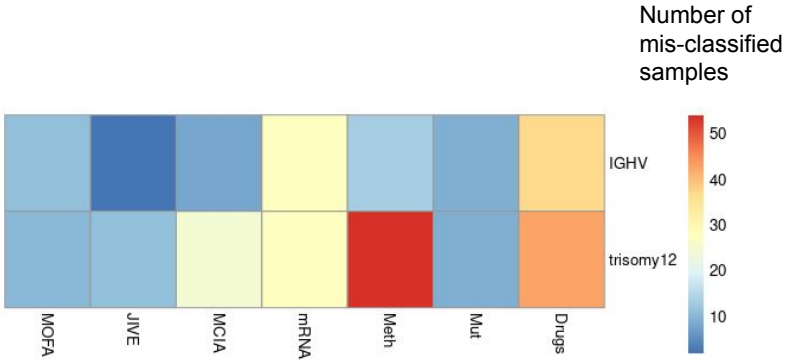
The first factor is captured by all methods. Some factors can't be captured by single modalities

Factors that aren't captured by any modality:
MOFA -> 1
JIVE -> 6
MCIA -> 2

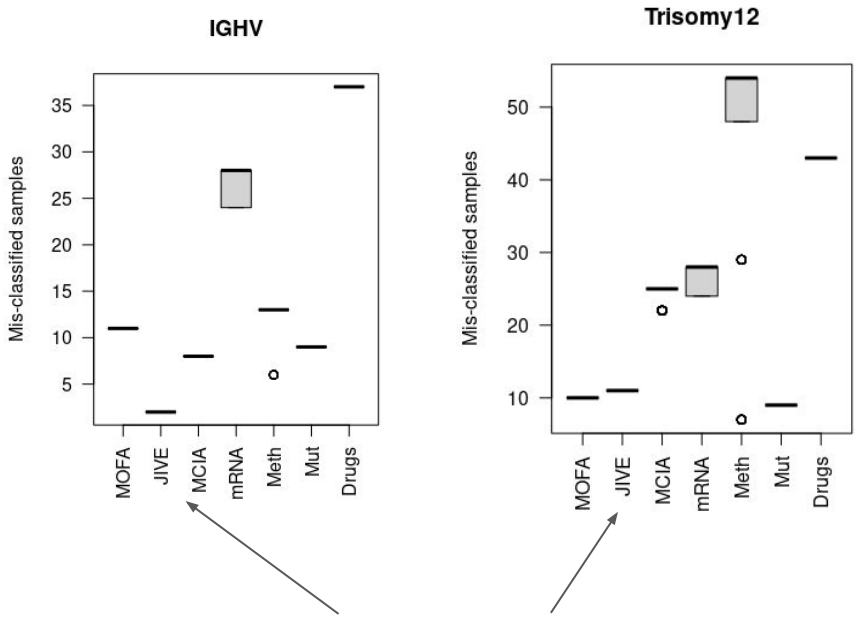
Clustering samples based on IGHV and trisomy12 status

K-means clustering (k=2) based on one identified factor

For each method, the factor with the lowest number of mis-classified samples is shown



1000 iterations of k-means clustering

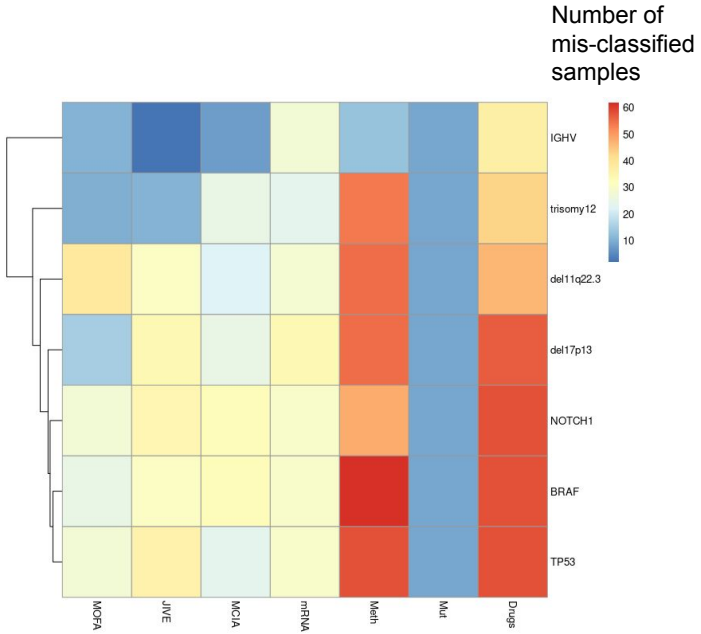


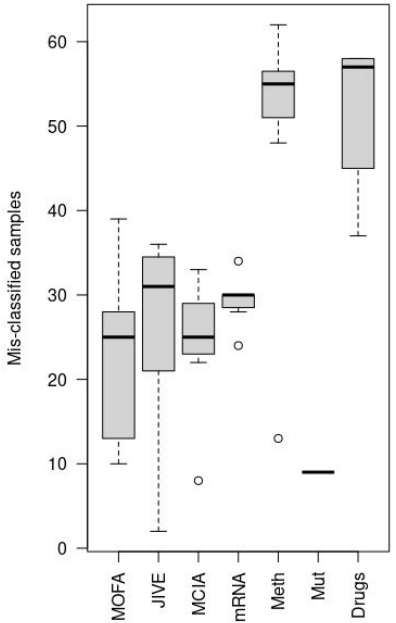
- JIVE has the best performance
- Single modalities are worse

Clustering samples based on prognostic mutations/CNA status

K-means clustering (k=2) based on one identified factor

For each method, the factor with the lowest number of mis-classified samples is shown





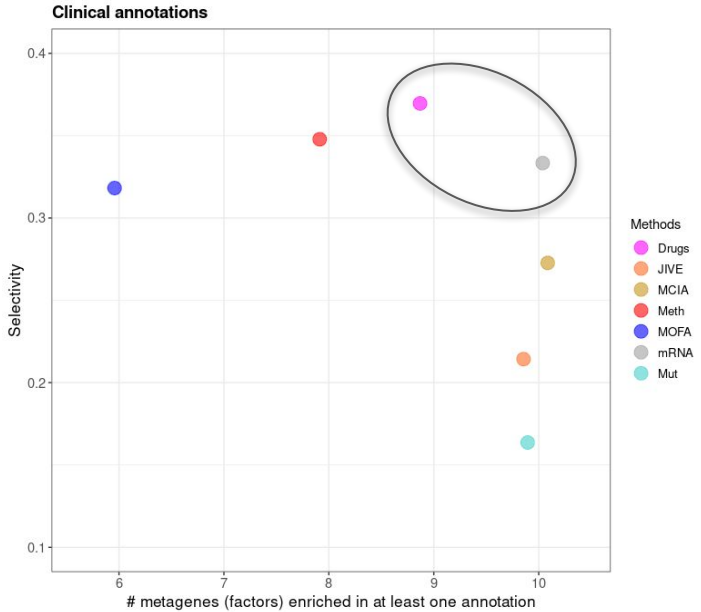
Extending the evaluation to other mutations/CNAs:

- Multiomics methods have the same performance
- mRNA alone is enough

Capturing clinical / biological variation

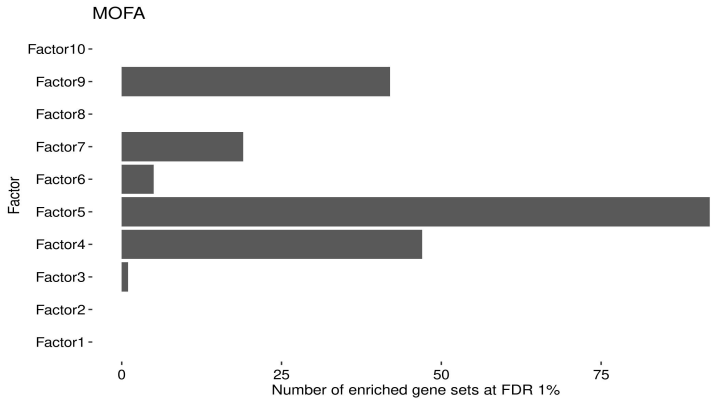
To quantify the independence of each factor in representing biological information, a selectivity score was devised by Laura et al. (2021) given by:

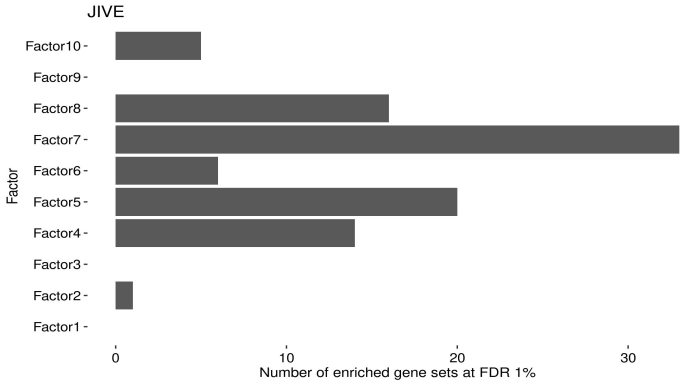
$$S = \frac{Nc + N_f}{2L}$$

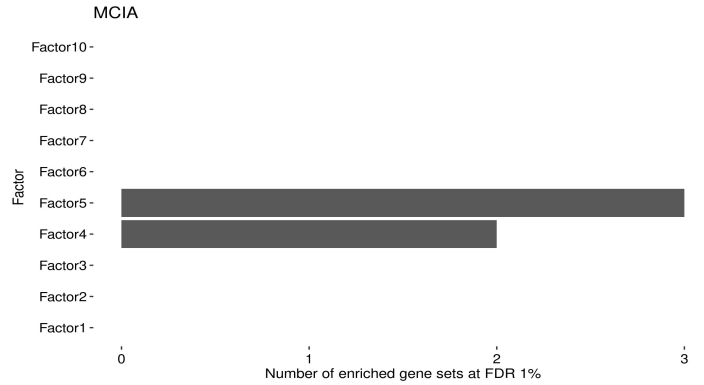


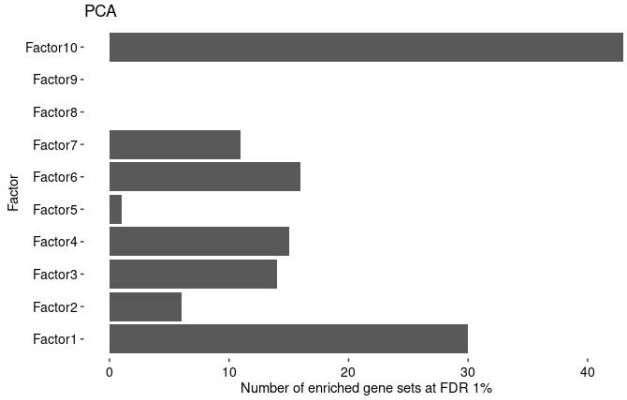
Single modalities capture many specific clinical variables

Enriched pathways in the factors - I

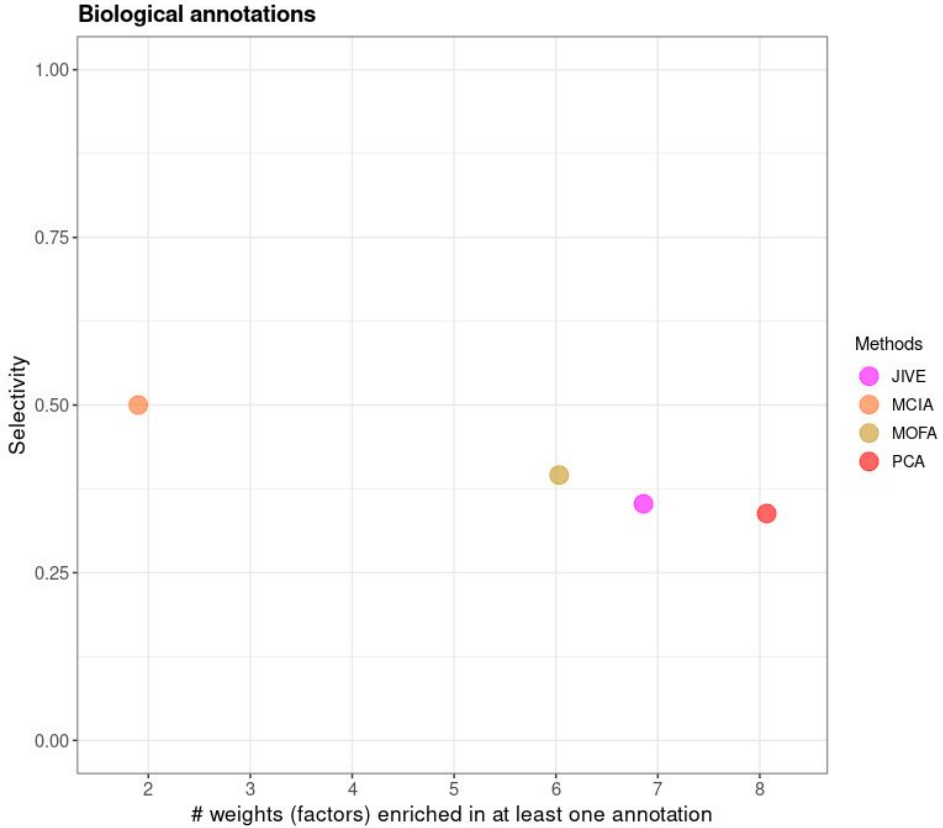




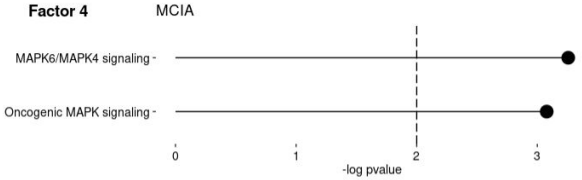


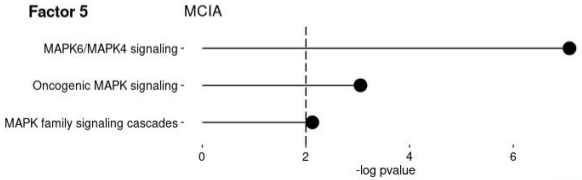


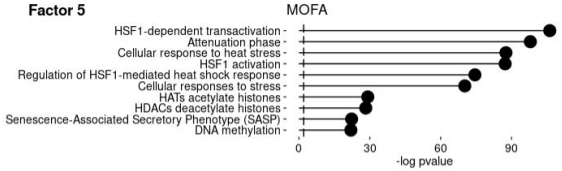
Enriched pathways in the factors - II

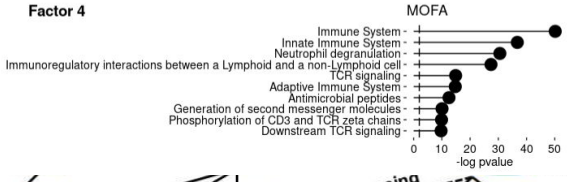


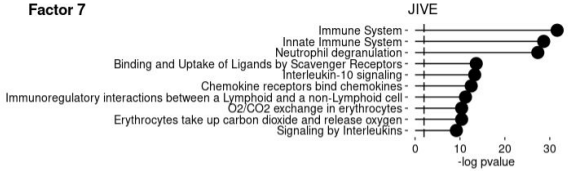
Enriched pathways in the factors - III

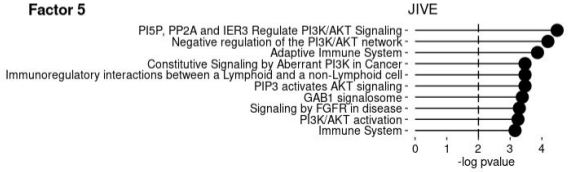


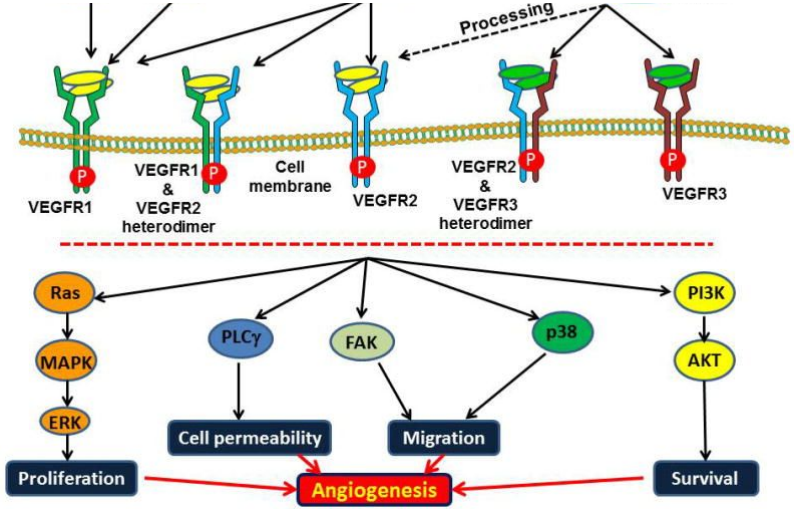


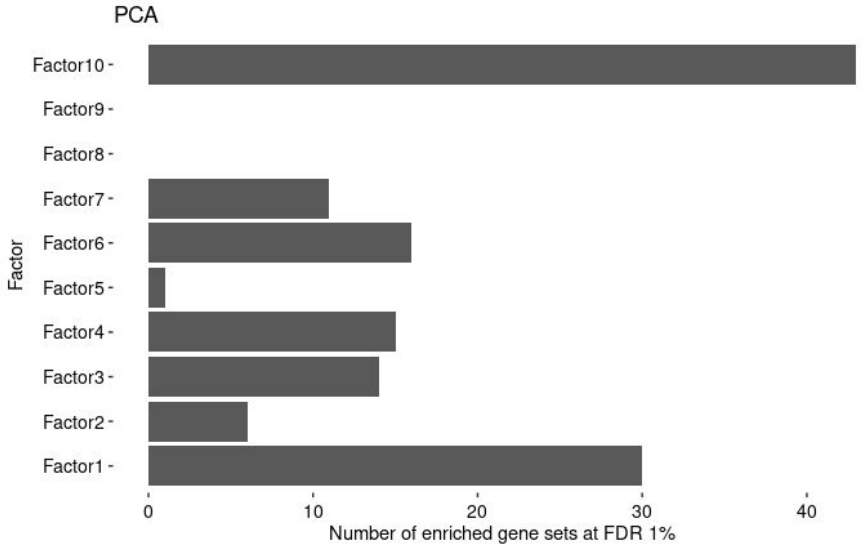


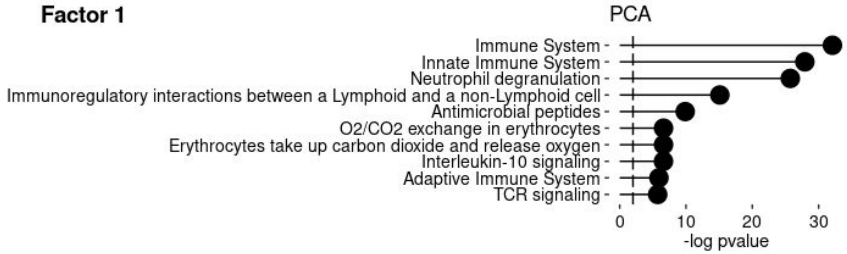


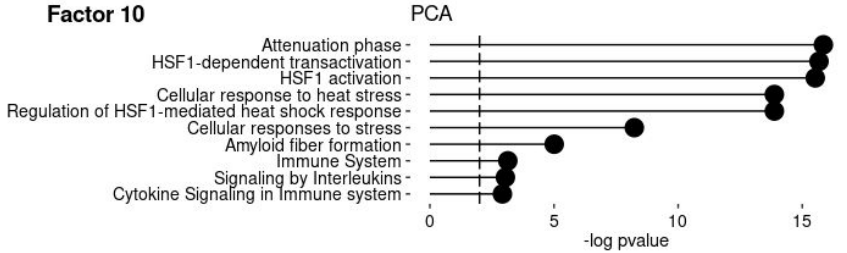




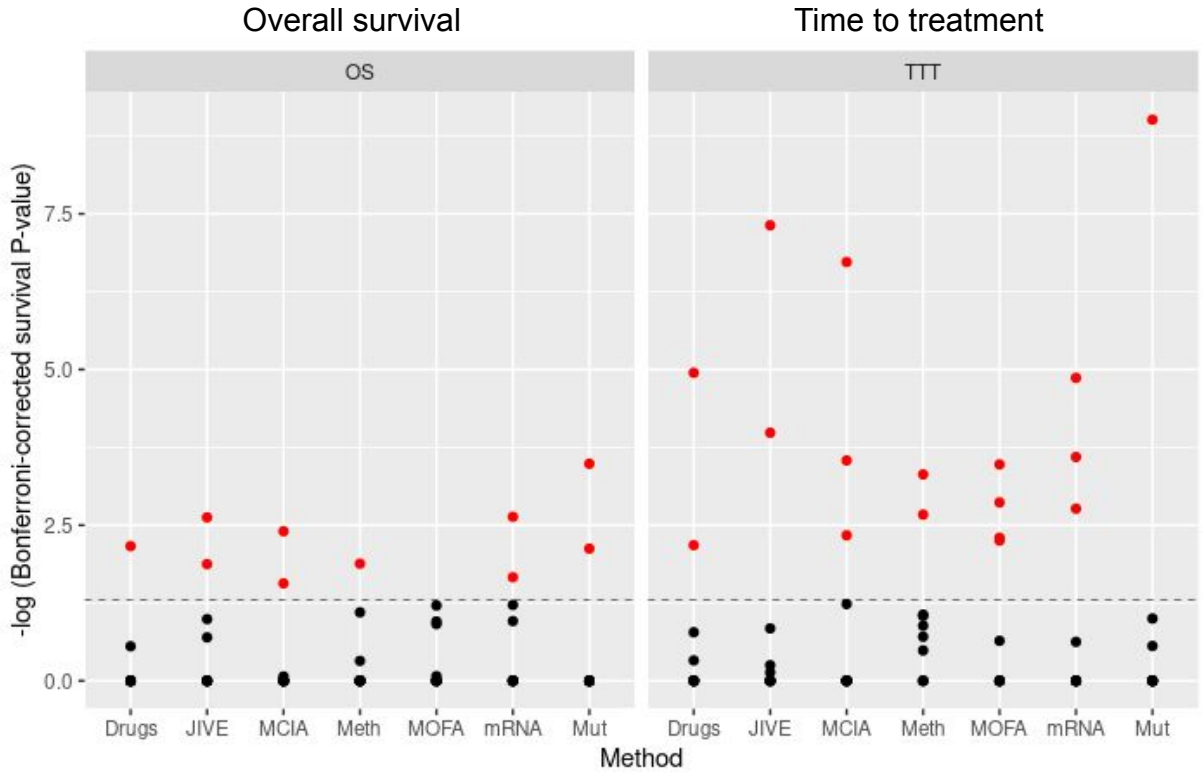



PCA
Factor10
Factor9
Factor8
Factor7
Factor6
Factor5
Factor4
Factor3
Factor2
Factor1
Factor
0
10
20
30
40
Number of enriched gene sets at FDR 1%


Factor 1
PCA
Immune System
Innate Immune System
Neutrophil degranulation
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Antimicrobial peptides
O2/CO2 exchange in erythrocytes
Erythrocytes take up carbon dioxide and release oxygen
Interleukin-10 signaling
Adaptive Immune System
TCR signaling
0
10
20
30
-log pvalue

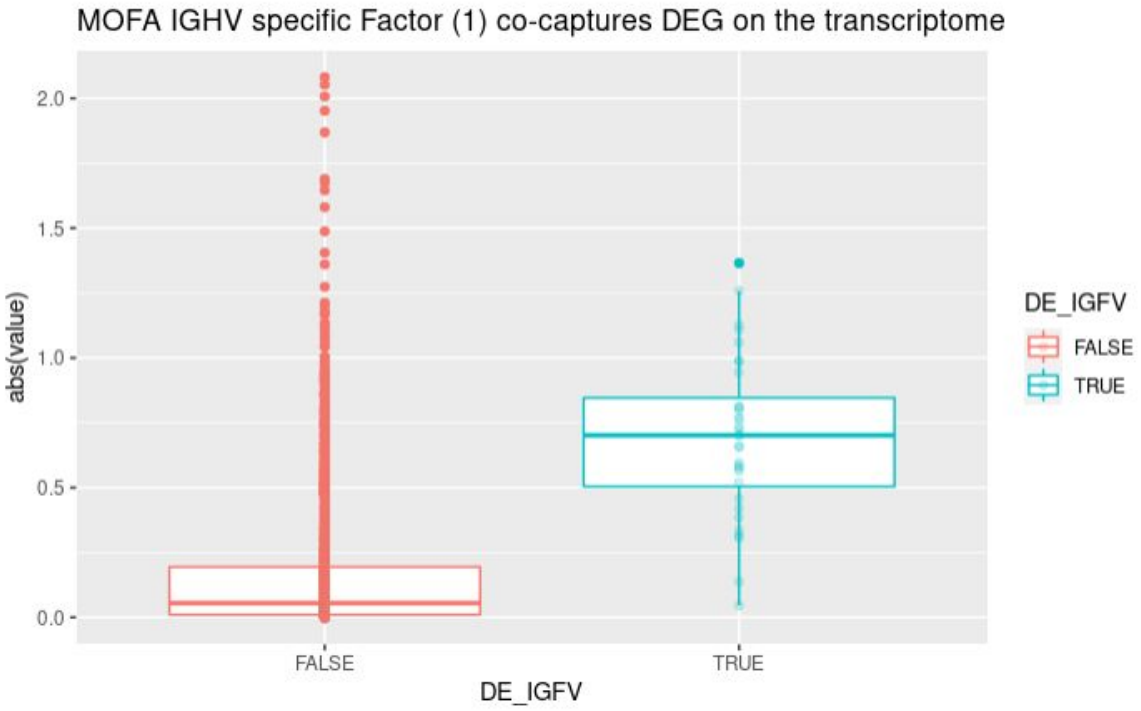

Factor 10
PCA
Attenuation phase
HSF1-dependent transactivation
HSF1 activation
Cellular response to heat stress
Regulation of HSF1-mediated heat shock response
Cellular responses to stress
Amyloid fiber formation
Immune System
Signaling by Interleukins
Cytokine Signaling in Immune system
0
5
10
15
-log pvalue

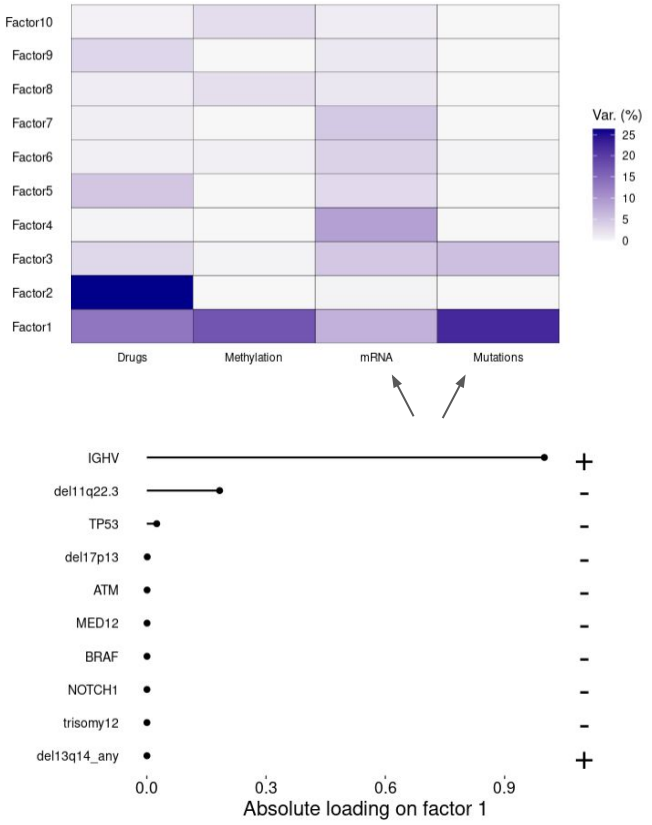
Capturing survival



All methods except for MOFA capture 1-2 factors correlated with overall survival.
True also for the simple PCA with one modality

MOFA Factors capturing Shared Variation

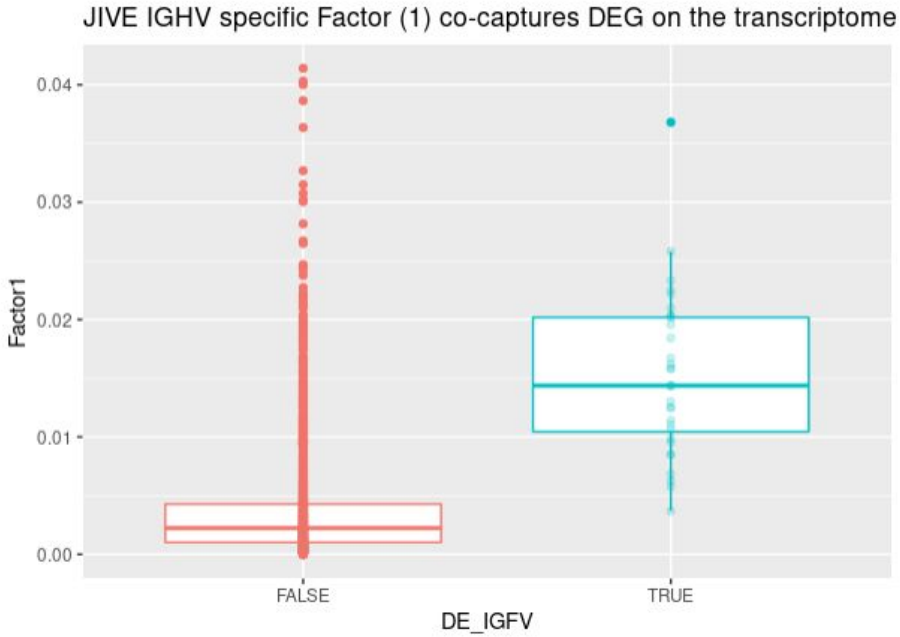


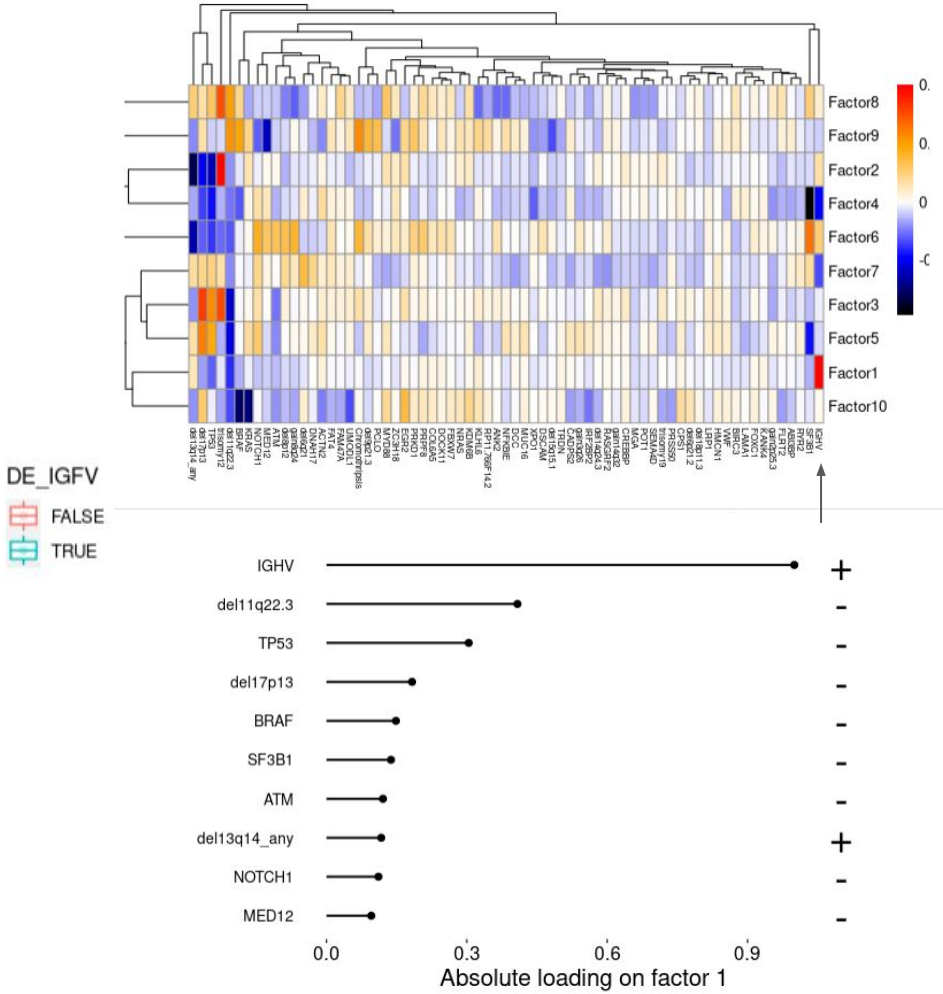


Enrichment Score < 0.001

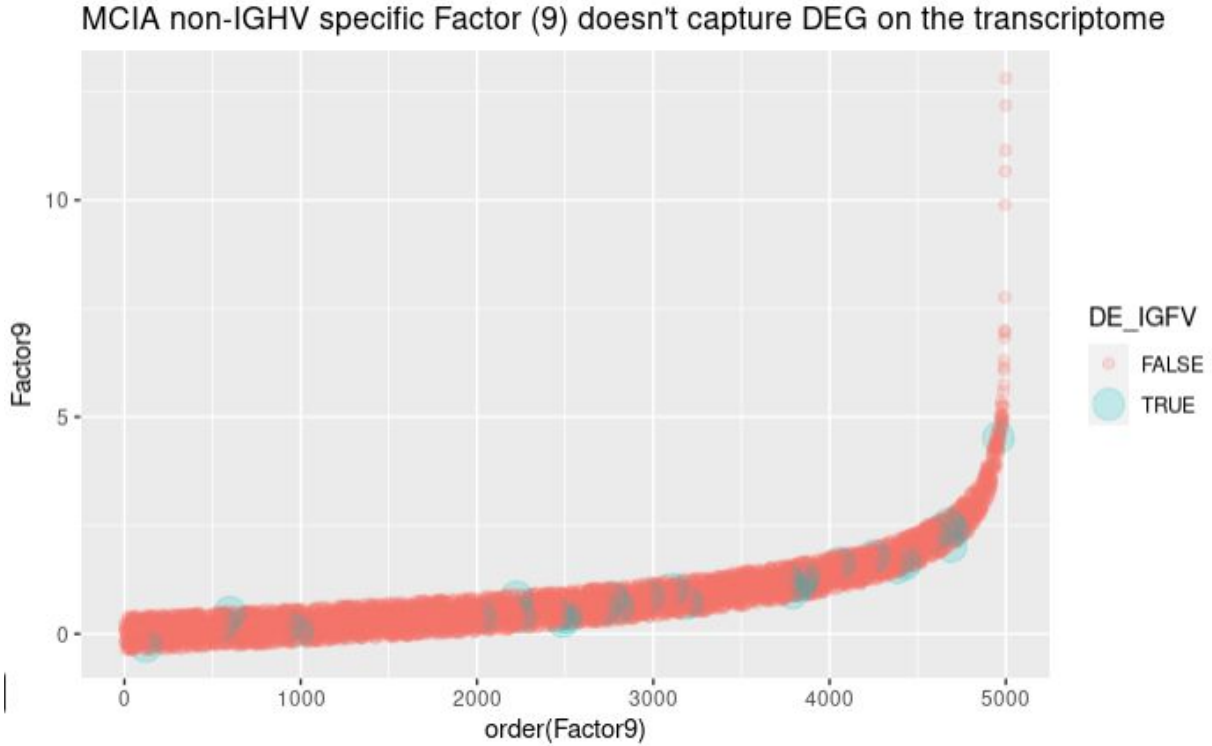
Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia Fereira et. al 2019

JIVE Factors capturing Shared Variation



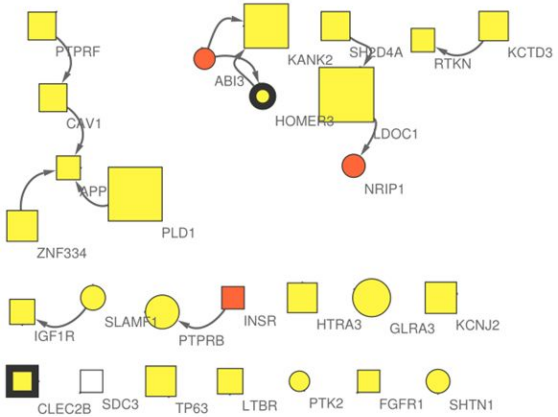


Enrichment Score < 0.001

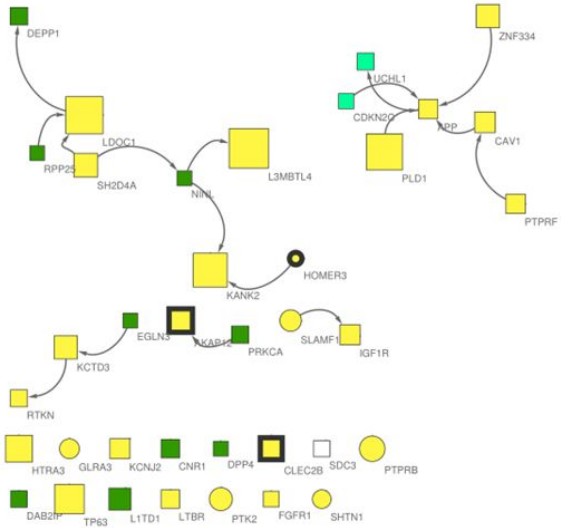

MCIA non-IGHV specific Factor (9) doesn't capture DEG on the transcriptome
Factor9
10
5
0
0
1000
2000
3000
4000
5000
order(Factor9)
DE_IGFV
FALSE
TRUE

PPI interactions for protein products of transcripts with largest weight on Factor 1

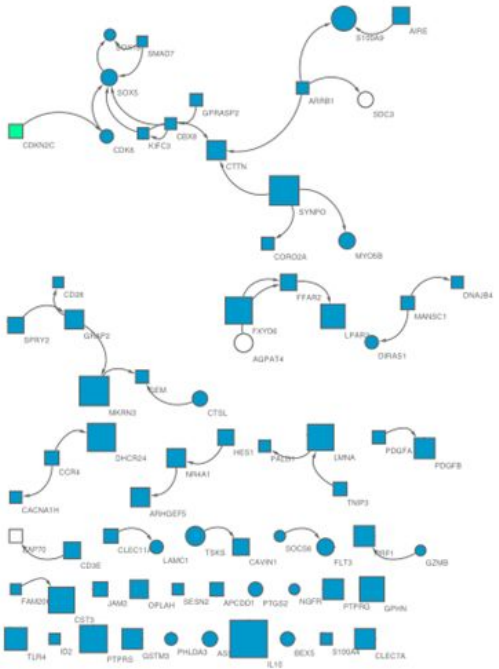
MOFA



JIVE



MCIA



MOFA
JIVE
MCIA
MOFA&JIVE
MCIA&JIVE
MOFA&JIVE&MCIA

IGHV-associated DEGs
marked with thicker border

WEIGHT:
MINUS
PLUS

IntAct networks for top-weighted transcripts on Factor 1 (absolute weight > 0.3)

Conclusions

- The evaluation strongly depends on the chosen metric
- Multiomics methods capture some variation that can’t be captured with single modalities
- All methods and PCA on mRNA perform similarly in clustering based on mutations. JIVE outperforms in clustering for IGHV and trisomy12 status.
- Simple PCAs with single modalities are enough to selectively retrieve clinical and biological variations
- Results obtained with MOFA and JIVE fit with independent datasets

Future plans

Additional benchmarkings:

- Do drugs with similar weights belong to the same class?

Explore relationships between modalities:

- Do corresponding mRNA and methylation have opposite weights on the same factor?
- Weights for drugs and their targets
- Can these methods inform on gene expression/methylation/mutations predictive of drug response?

Appendix

Model fitting and technical details

MOFAobject_unprepared <- **create_mofa**(data_list)
MOFAobject_prepared <- **prepare_mofa**(MOFAobject, **data_options**, **model_options**, **training_options**)
MOFAmodel <- **run_mofa**(MOFAobject, outfile = "path/file_name.hdf5")

data_options
- **scale_views**: whether the variance between views should be scalded (to unit variance).
- **scale_groups**: whether the variance between groups should be scalded (to unit variance).

model_options
- **likelihoods**: assumption of data distribution ("gaussian" for continuous data, "bernoulli" for binary, "poisson" for count) by default, inferred automatically
- **num_factors**: number of factors to be fitted

training_options
- **convergence_mode**: fast, medium, or slow
- **seed:** randomization seed
- gpu_mode: support gpu accelerating

MOFA2 supports gpu accelerating.

JIVEmodel <- jive(imputed_data, rankJ, rankA, method, cov, maxiter, showProgress, scale, center)

imputed_data: **should we always feed JIVE with imputed data?**

Jive … **replaces the missing values** using the SVDmiss function if necessary.

Function that completes a data matrix using iterative svd as described in Fuentes et. al. (2006). The function iterates between computing the svd for the matrix and replacing the missing values by linear regression of the columns onto the first ncomp svd components. As initial replacement for the missing values regression on the column averages are used. The function *will fail* if entire rows and/or columns are missing from the data matrix.

rankJ: number of factors to fit

method: use "given" to specify number of factors

jive does not support parallel computing natively, i.e. however powerful your computer (cluster) is, jive only uses one core.

mcia(df.list, cia.nf, cia.scan, nsc, svd)

df.list: a list of dataframe (matrix, ExpressionSet) containing multi-omics data

cia.nf: **number of factors to be fitted**

cia.scan: keeping default value is good

nsc:keeping default value is highly recommended

svd:keeping default value is good

	MOFA2	jive	mcia
Package	Bioconductor MOFA2	CRAN r.jive	Bioconductor omicade4
Model training function	create_MOFA prepare_MOFA run_MOFA	jive	mcia
Accept missing values	Yes	No	No
Missing values handling (default method)	Ignore NA[1]	SpatioTemporal::SVDmiss Error if unable to complete. [2]	
Allow specifying distribution	Yes	No	No
Predict function	predict	jive.predict	No
Parallel computing [3]	GPU supported	No	No

[1] Developers said, MOFA2 is robust to ignore missing values
[2] SVDmiss can't accept too many missing values. Our data can't be imputed.
[3] Factorization methods support parallel computing: iCluster, RGCCA

Selectivity score. We define the selectivity as:

$$S = \frac{Nc + N_f}{2L} \tag{13}$$

where Nc is the total number of clinical annotations associated with at least a factor, Nf the total number of factors associated with at least a clinical annotation, and L the total number of associations between clinical annotations and factors. S has a maximum value of 1 when each factor is associated with one and only one clinical/biological annotation, and a minimum of 0 in the opposite case. An optimal method should thus maximize its number of factors associated with clinical/biological annotations without having a too low selectivity.