Analyzing multi-omics cancer data

Jaron Arbet

Contents

1. TCGA data
2. Exploratory dimension reduction
3. Machine learning for survival data

TCGA

The Cancer Genome Atlas Program

TCGA2STAT R package[1]: available cancer types and omics data

Cancer name	Acronym	RNASeq V2	RNASeq	miRNASeq	CNA_SNP	CNV_SNP	CNA_CGH	Methylation (27K)	Methylation (450K)	Mutation	mRNA_Array	miRNA_Array
Adrenocortical carcinoma	ACC	Y		Y	Y	Y			Y	Y		
Bladder urothelial carcinoma	BLCA	Y	Y	Y	Y	Y			Y	Y		
Breast invasive carcinoma	BRCA	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Cervical and endocervical cancers	CESC	Y		Y	Y	Y			Y	Y		
Cholangiocarcinoma	CHOL	Y		Y	Y	Y			Y			
Colon adenocarcinoma	COAD	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Colorectal adenocarcinoma	COADREAD	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	Y		Y	Y	Y			Y			
Esophageal carcinoma	ESCA		Y	Y	Y	Y			Y			
FFPE Pilot Phase II	FPPP			Y								
Glioblastoma multiforme	GBM	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y
Glioma	GBMLGG	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y
Head and Neck squamous cell carcinoma	HNSC	Y	Y	Y	Y	Y			Y	Y		
Kidney Chromophobe	KICH	Y		Y	Y	Y			Y	Y		
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Kidney renal clear cell carcinoma	KIRC	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Kidney renal papillary cell carcinoma	KIRP	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Acute Myeloid Leukemia	LAML	Y	Y	Y	Y	Y		Y	Y	Y		
Brain Lower Grade Glioma	LGG	Y		Y	Y	Y			Y	Y	Y	

- ~90% of kidney cancers are renal cell carcinoma (RCC)[2]
- ~70% of RCC are “clear cell” type[2]

[1] Wan et al "TCGA2STAT: simple TCGA data access for integrated statistical analysis in R." Bioinformatics 32.6 (2016): 952-954. http://www.liuzlab.org/TCGA2STAT/

[2] https://www.cancer.org/cancer/kidney-cancer/about/what-is-kidney-cancer.html

RNA-Seq Gene Expression

1. Download raw counts:

   TCGA2STAT::getTCGA(disease="KIRC", data.type="RNASeq", type="count", clinical=T)

2. Remove low expressed genes: edgeR::filterByExpr()

3. Normalize using **TMM** (trimmed mean of M-values) and log-CPM (edgeR package)

   207 samples, **16518** genes

Why TMM? Comparison of 7 normalization methods[1] found only TMM and DESeq robust to heterogeneity in library size and composition; controlled FPR while maintaining power

[1] Dillies et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA seq..." Briefings in bioinformatics 14.6 (2013): 671-683.

DNA Methylation

1. Download beta values:

   TCGA2STAT:: getTCGA(disease==“KIRC”, data.type="Methylation", type="27K")

2. Remove features with missing values

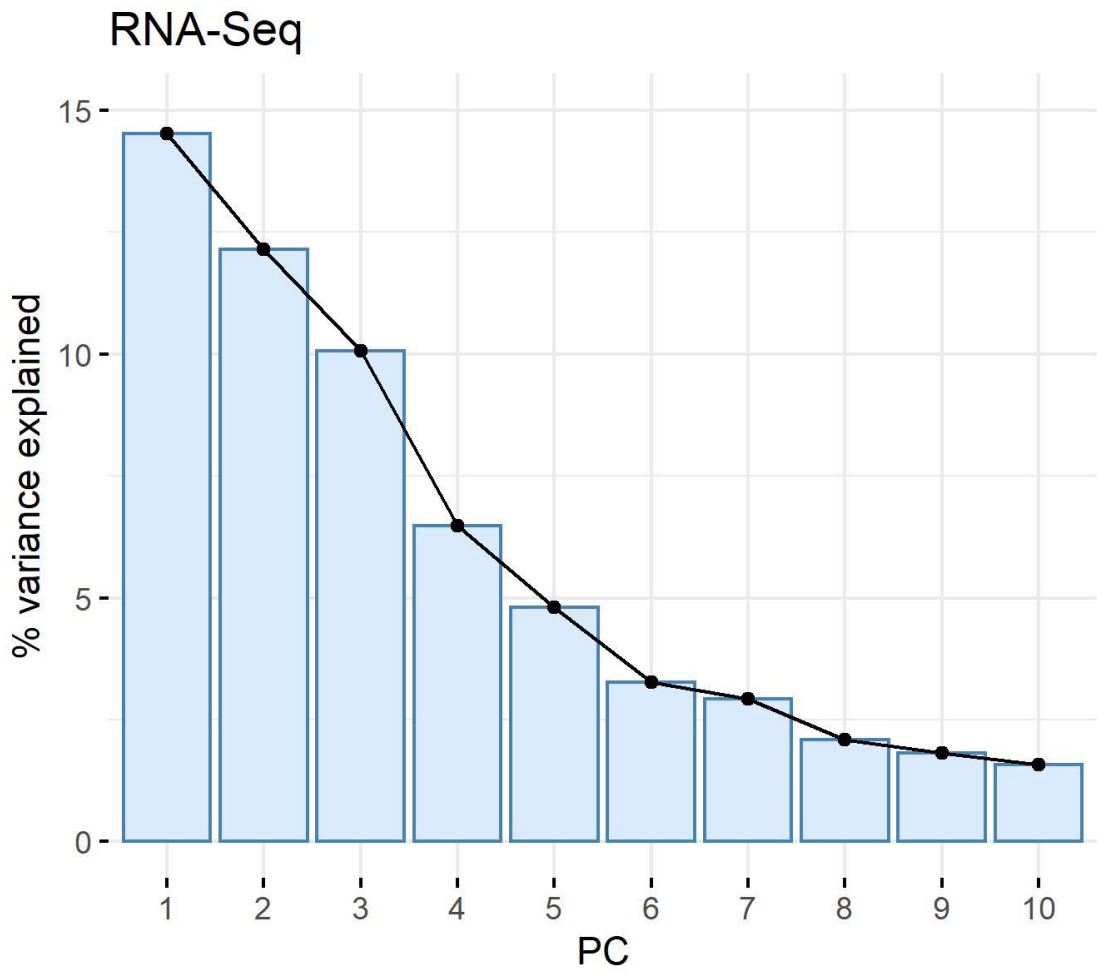
Clinical data

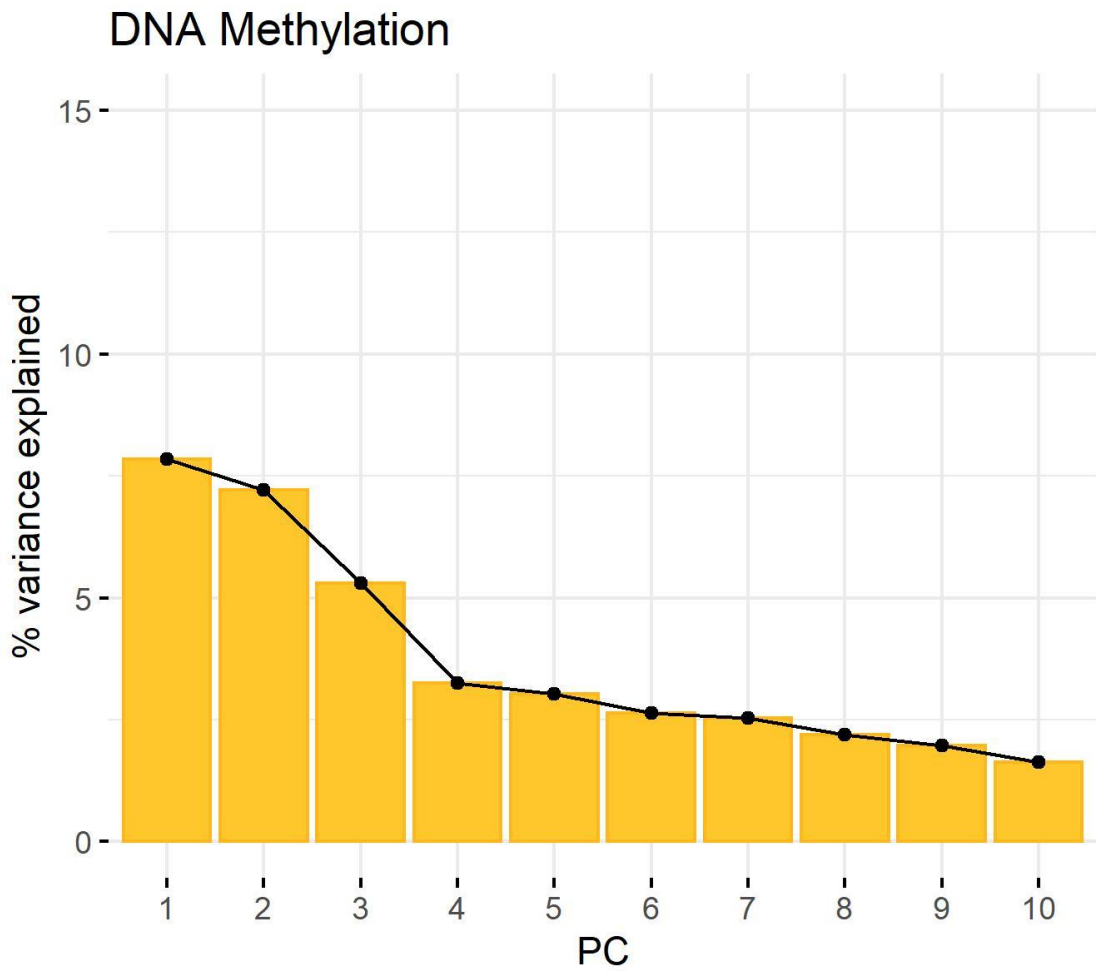
- Age at diagnosis
- Time until death (or censoring)
- Cancer stage (1-4)
- Gender
- Race

Total data:
207 samples with 16518 genes and 23166 CGs (39684 features)

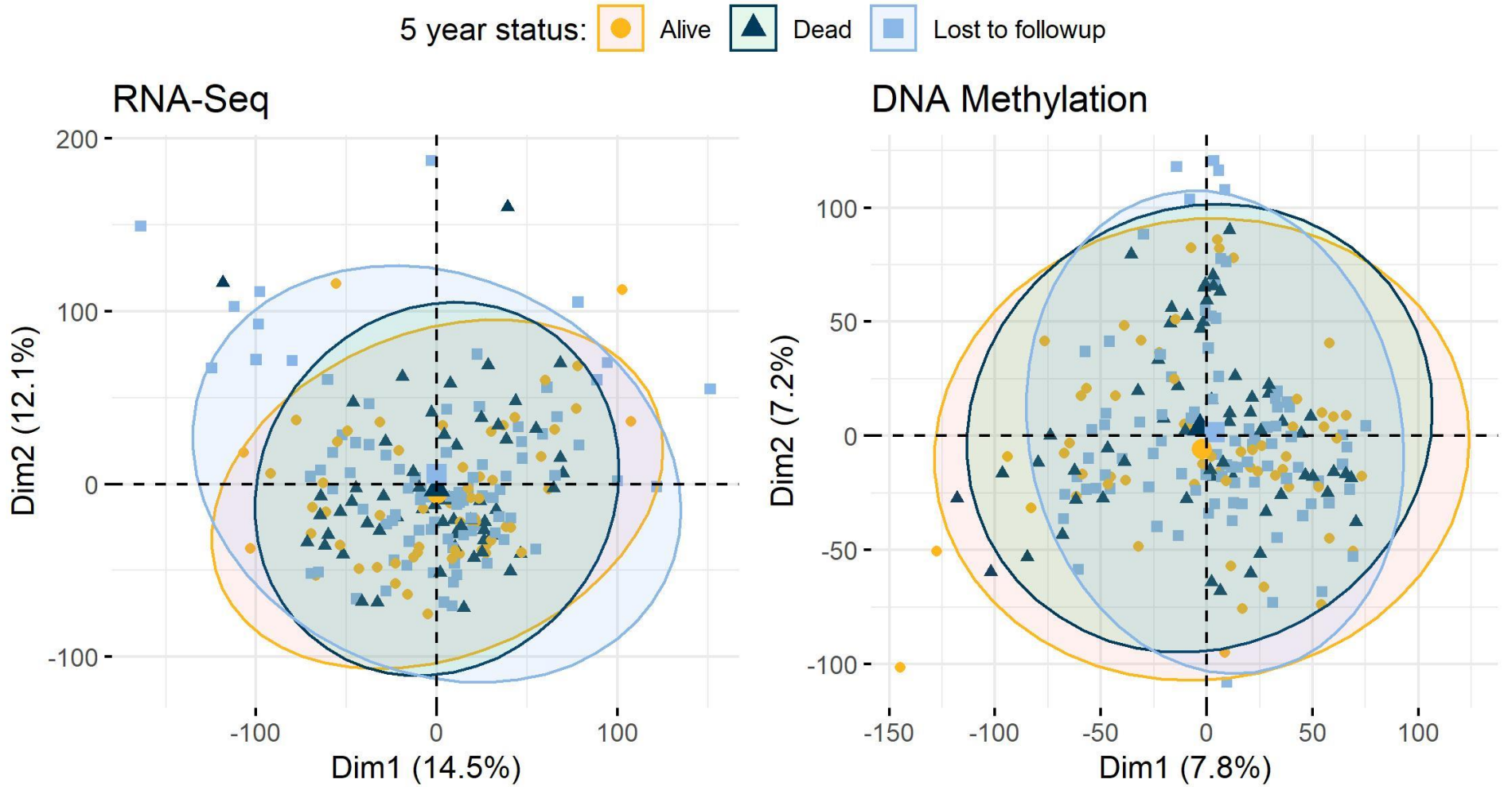
Exploratory dimension reduction

Principal component analysis (PCA)





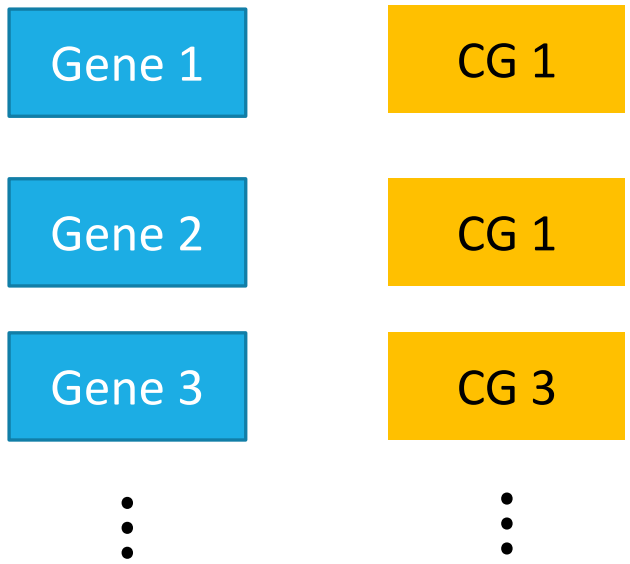
Principal component analysis (PCA)



Association between genes and CGs

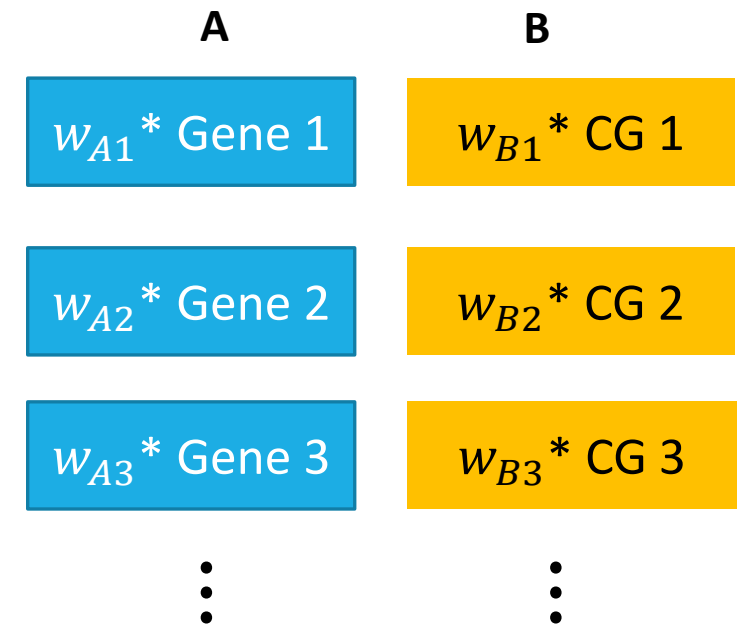
Partial least squares (**PLS**) and Canonical Correlation Analysis (**CCA**)

- Identify correlated sets of features between multi-omics data types
- Derives “latent features” that are linear combinations of original features
  - Features are assigned weights to maximize the covariance (PLS) or correlation (CCA) between new latent features
  - “sparse” versions perform variable selection



Could look at pairwise correlations, not feasible in high dimensions (e.g. 10k genes vs 10k CGs = 100 million correlations between data types)

Create new latent features **A** and **B** (linear combos of original features)

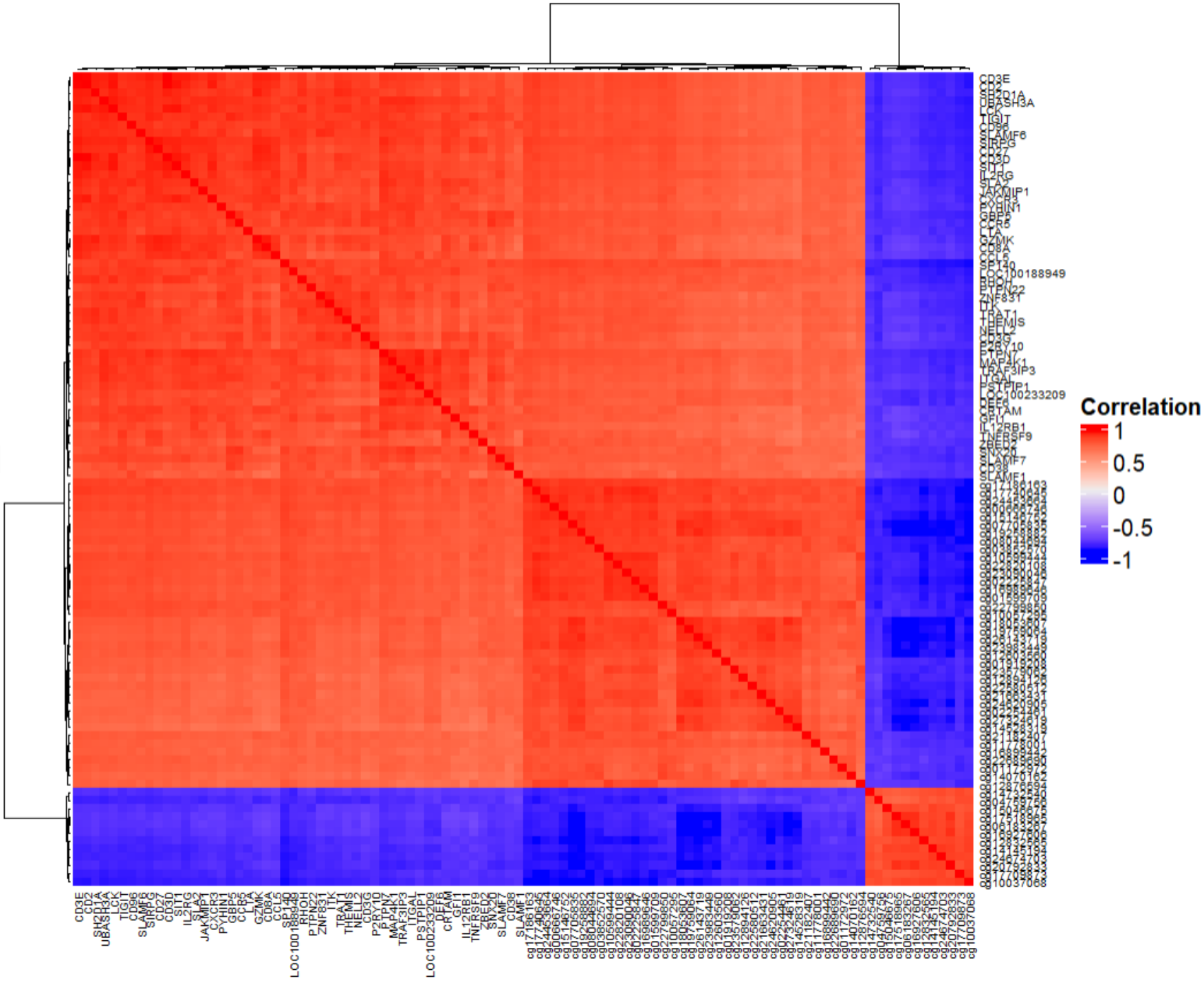


- Weights w_A and w_B are chosen to maximize $Cov(A, B)$ for PLS, or $Corr(A, B)$ for CCA
- "sparse" versions give weights of 0 to unimportant features
- A large $|w_{Aj}|$ means that gene is correlated with many CGs
- A large $|w_{Bj}|$ means that CG is correlated with many genes

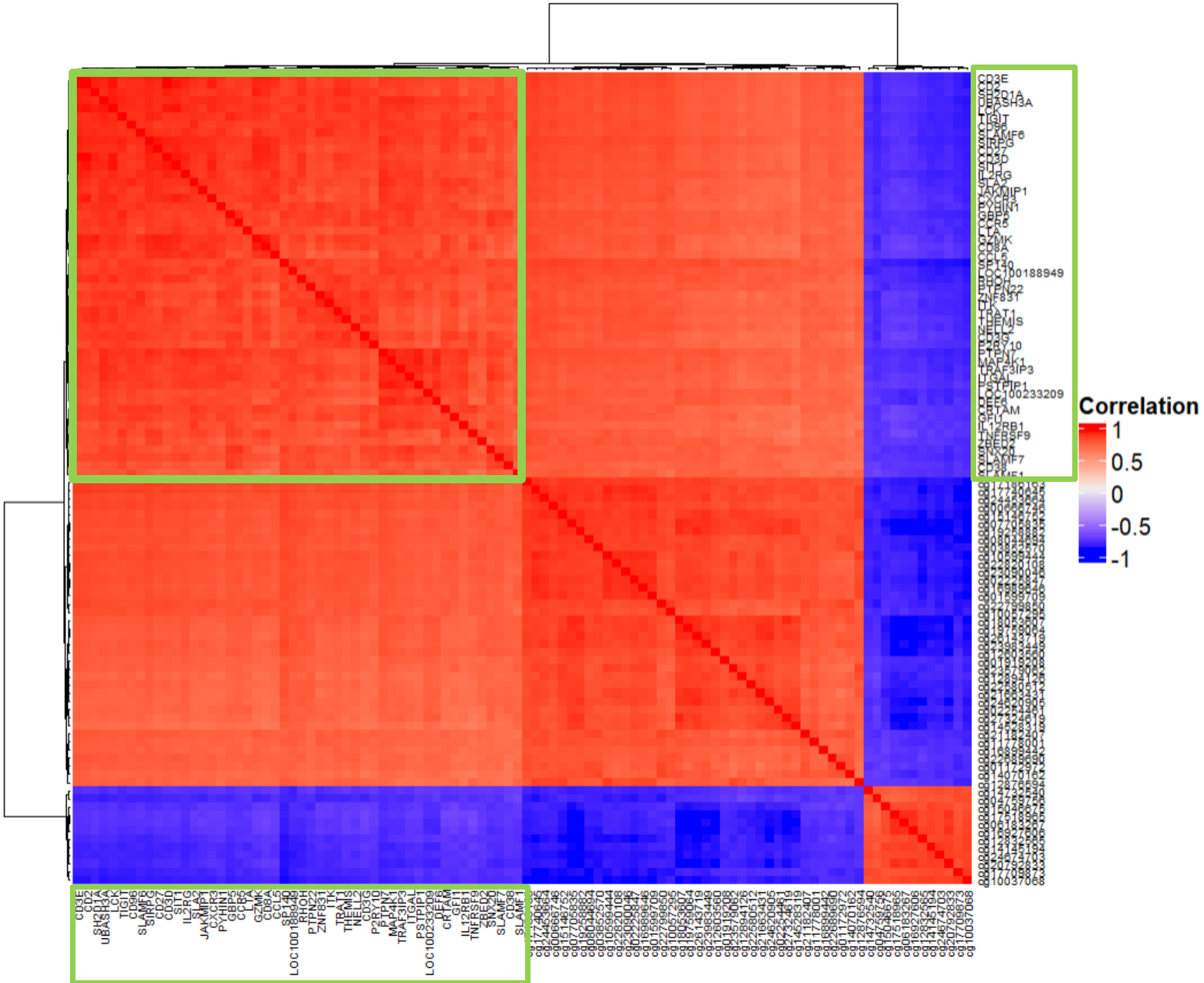
sPLS Results

- mixOmics R package
- Must specify # latent features (LFs) and # variables for each LF
- For exploratory purposes, I chose 1 LF, with 50 genes and 50 CGs

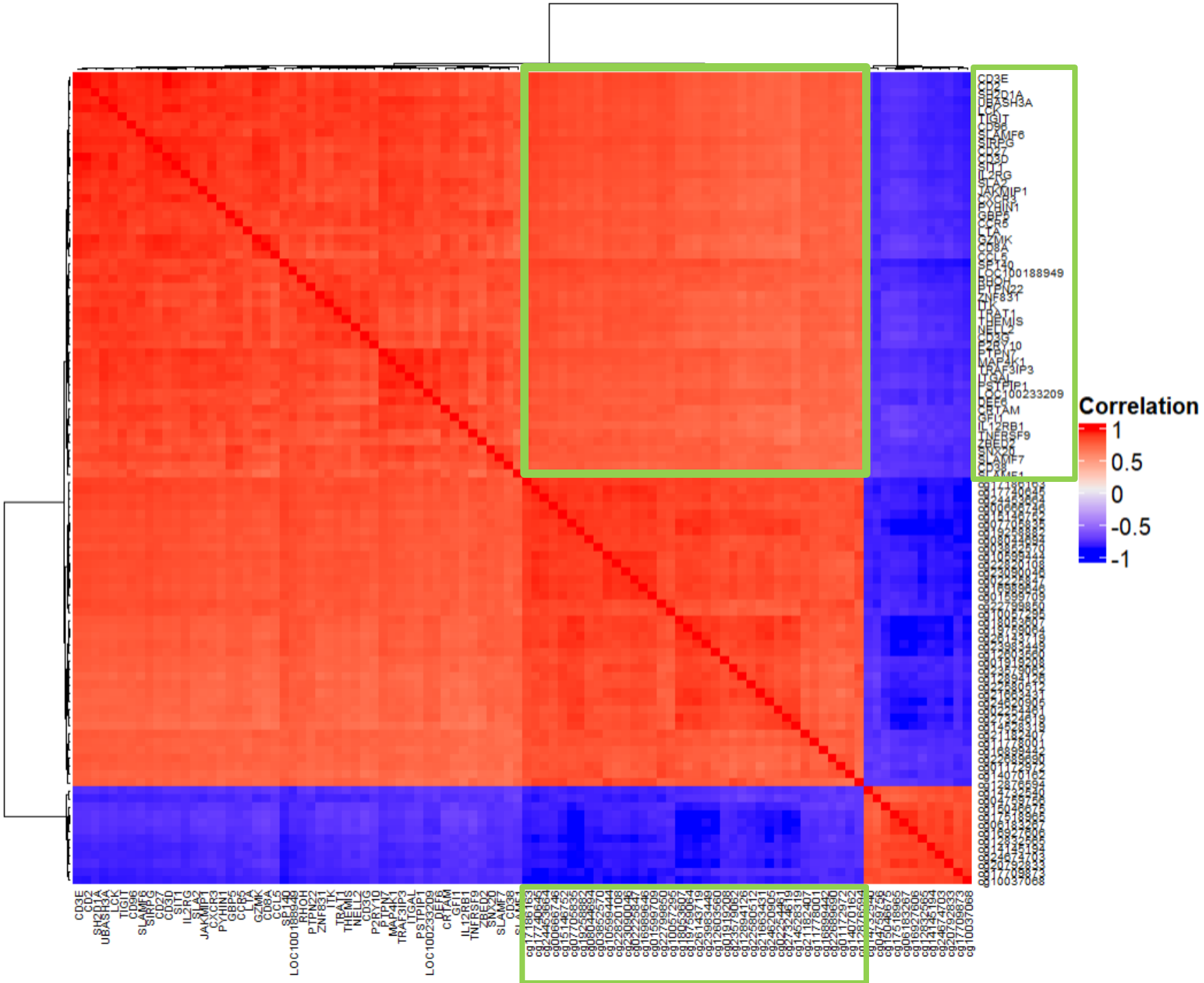
Spearman correlation heatmap of the 100 selected variables:



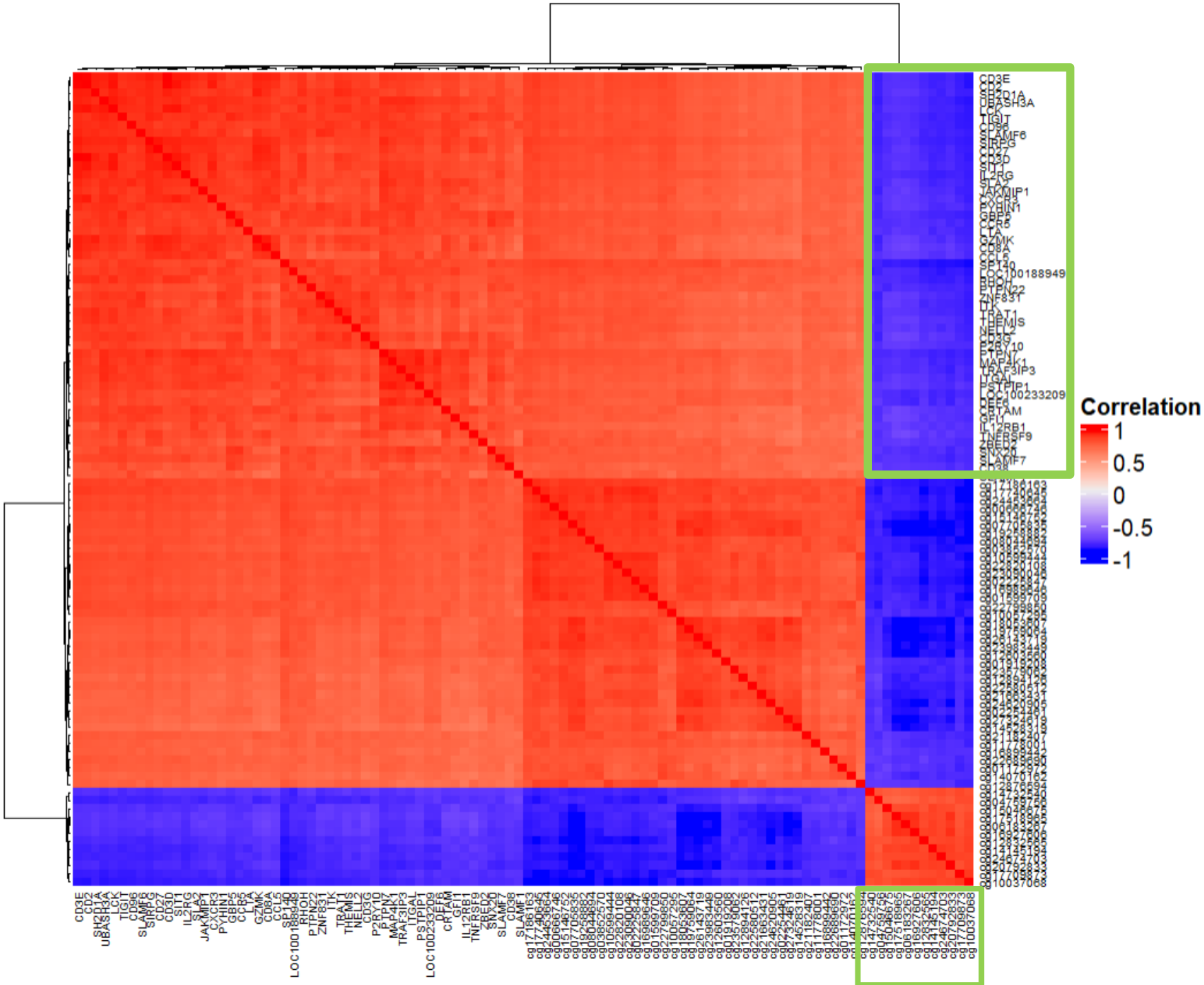
Positively correlated gene expression



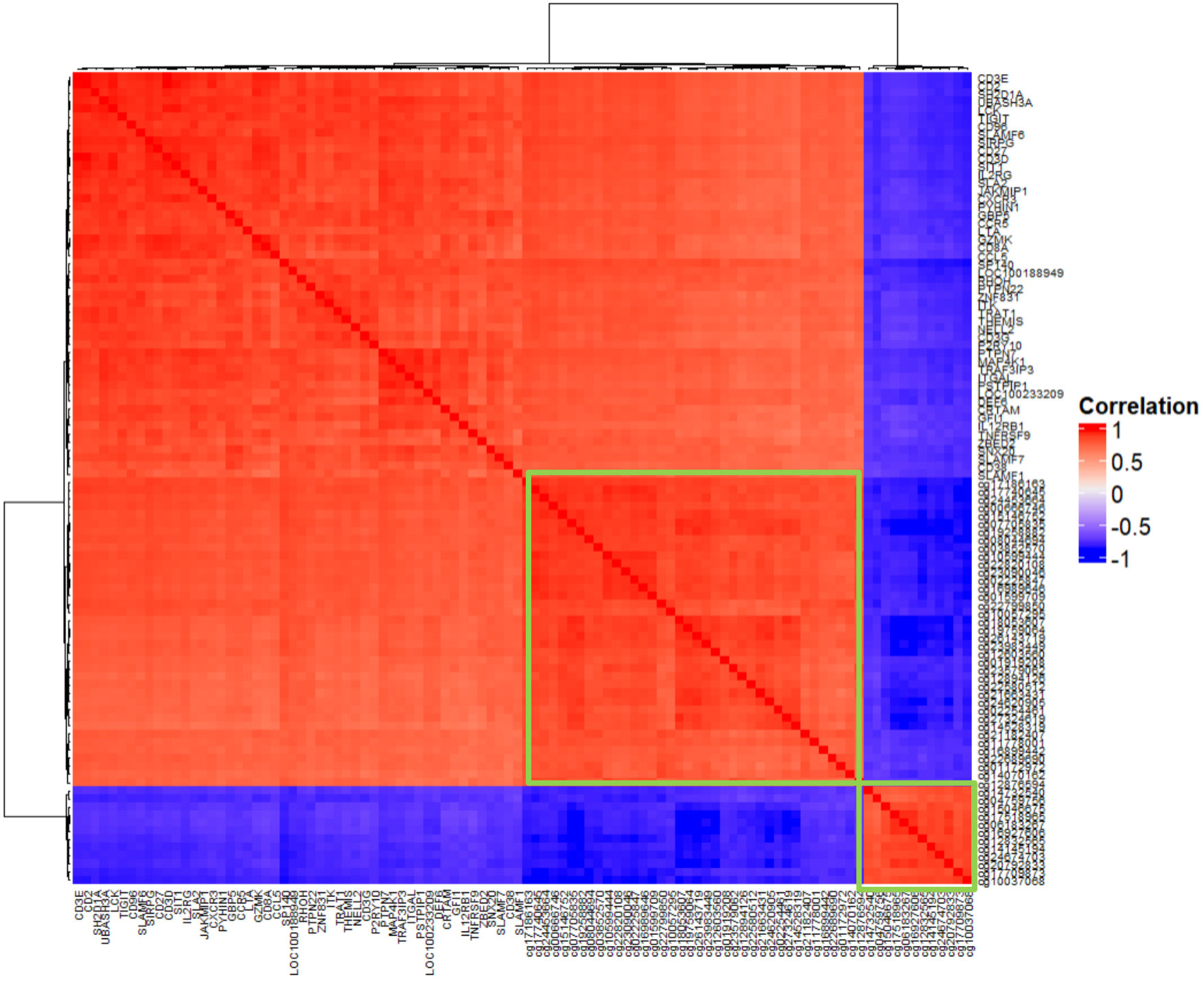
Positively correlated genes-CGs



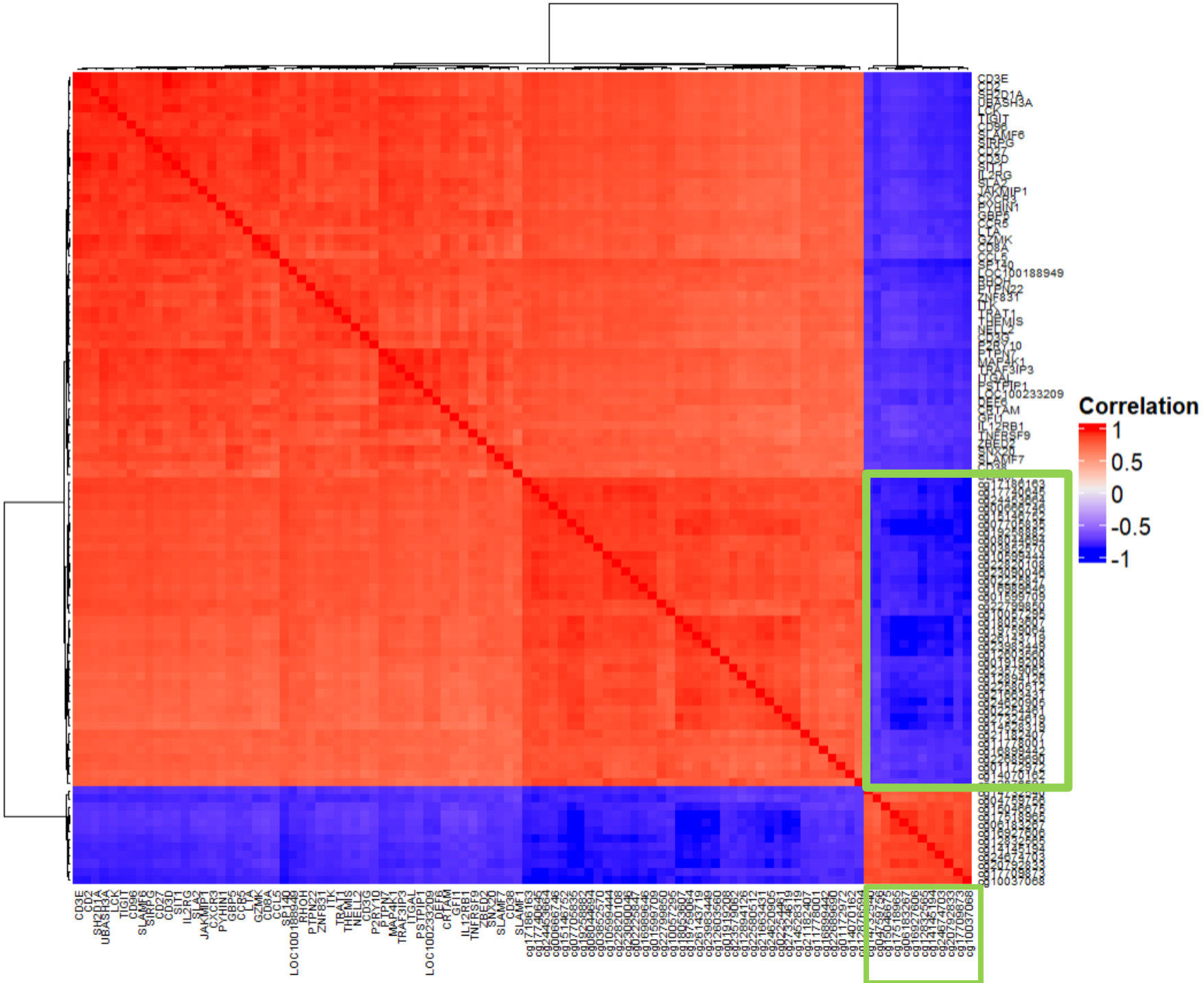
Negatively correlated genes-CGs



Positively correlated CGs



Negatively correlated DM



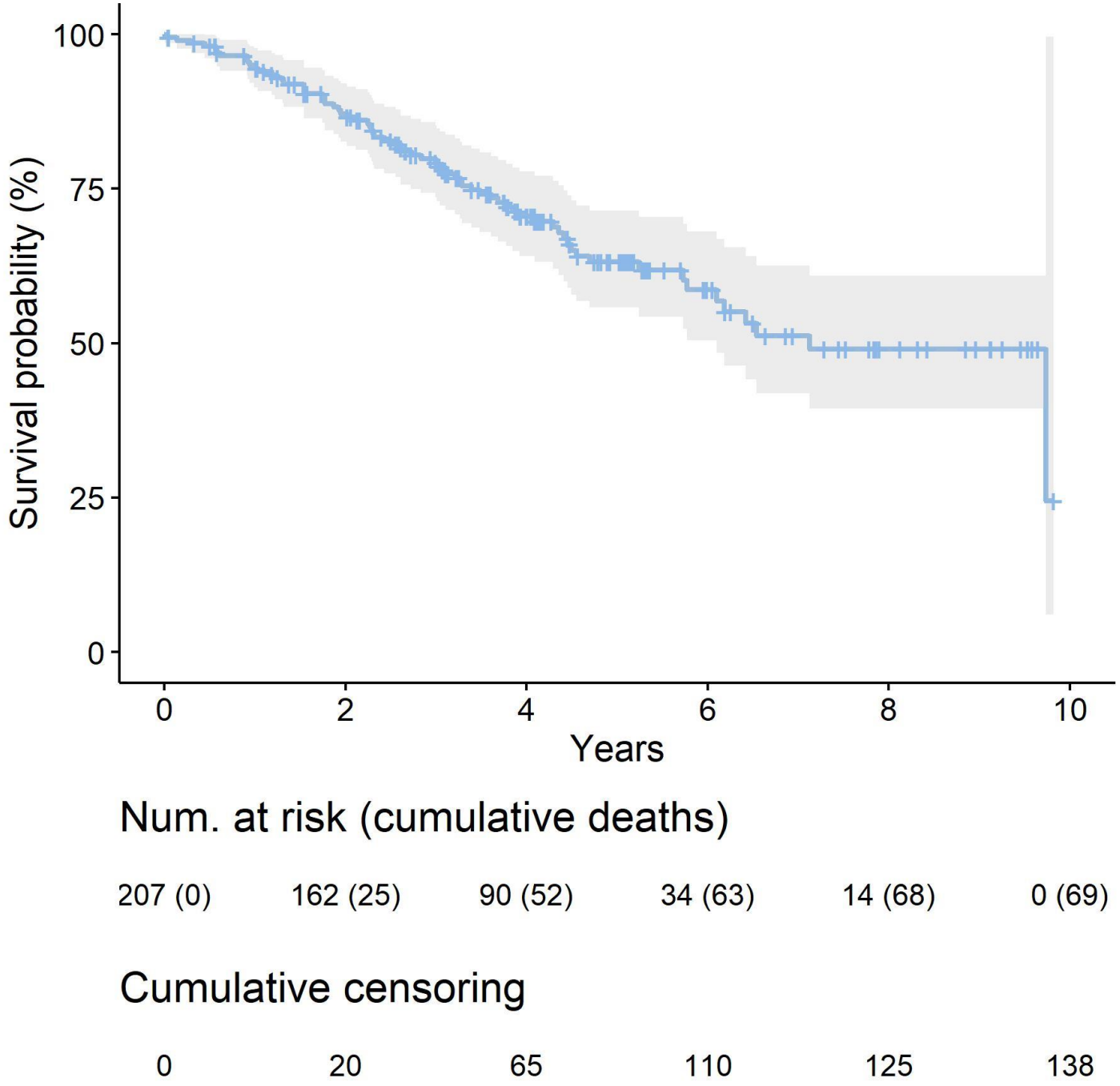
Extensions and other ideas for sPLS/CCA

- Supervised sPLS/CCA: latent feature weights are chosen to maximize covariance/correlation between multi-omics data types *and* with an outcome (mixOmics and PMA R packages)
- Use unsupervised sPLS/CCA as a filtering step before network analysis (e.g. WGCNA)
  - Dimension reduction: only keep top X variables with non-zero weights
- Pathway enrichment analysis to assess functional importance of any “modules”

Machine learning with survival data

Kaplan-Meier survival curve

- 5 year survival rate = 63%



2 ML methods for survival analysis

Regularized Cox model[1]

- Extends Cox model with variable selection
- R packages: glmnet, penalized

Pros

- Easy interpretation (hazard ratios)
- LASSO/Elastic-net identify important variables (logHR ≠ 0) vs unimportant variables (logHR=0)

Cons

- Proportional hazards assumption
- Assumes linear and additive effects

Random survival forests[2]

- Extends random forests for survival outcome
- R packages: randomforestSRC, ranger, party

Pros

- Nonparametric (less assumptions)
- Automatically allow for non-linear and interaction effects
- Variable importance scores
- Some software allows for missing values
- Bootstrap gives estimate of generalization error (cross-validation not necessary)

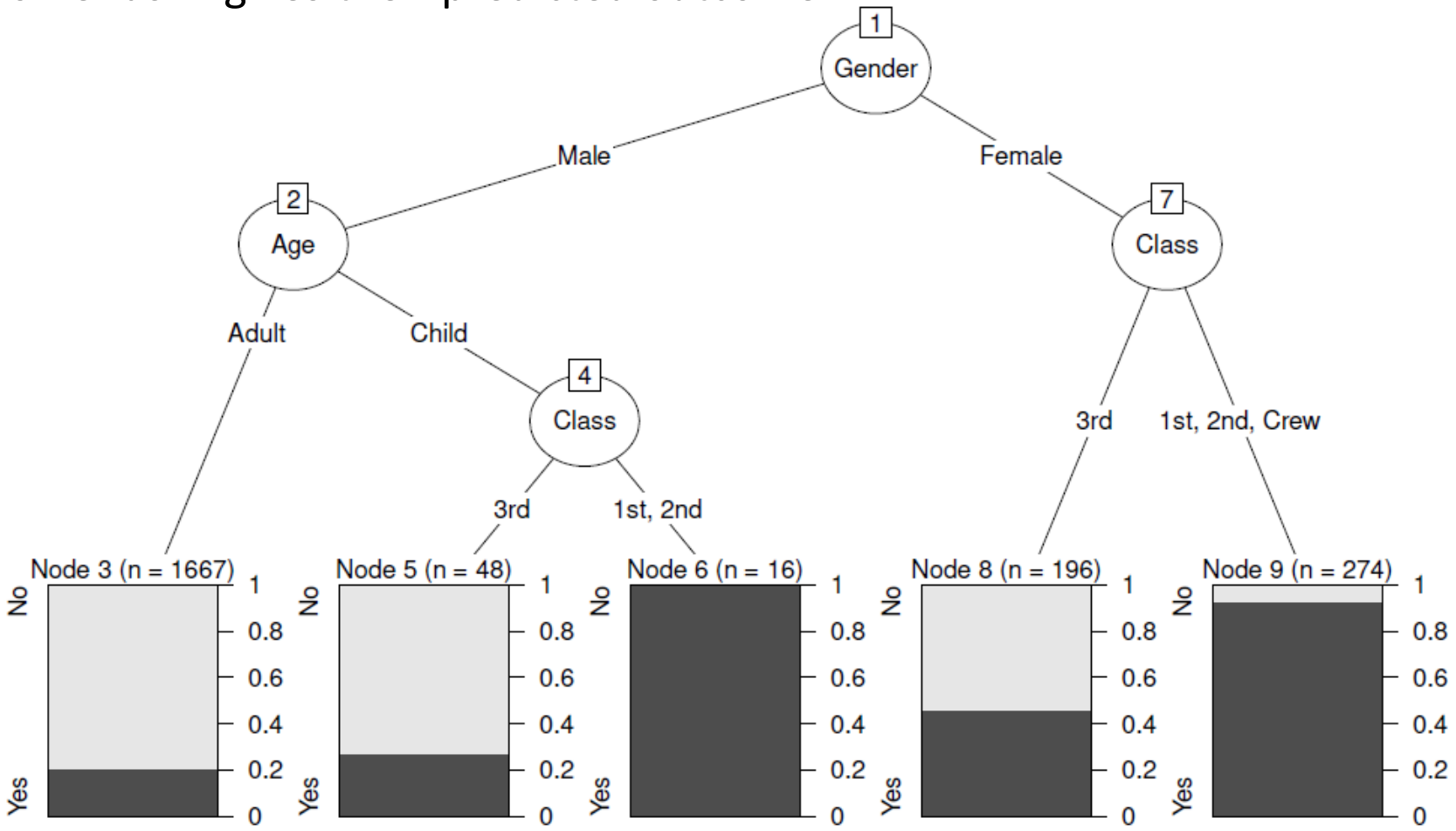
Cons: harder to interpret

1 Simon et al. "Regularization paths for Cox's proportional hazards model via coordinate descent." *Journal of statistical software* 39.5 (2011): 1.
2 Bou-Hamad, Imad, Denis Larocque, and Hatem Ben-Ameur. "A review of survival trees." Statistics surveys 5 (2011): 44-71.

Decision trees and forests?

Example[1]: want to know the probability that a particular person survived the Titanic:

- Starting at the top of the tree, sequentially ask each question… whatever final "leaf" the person ends in gives their predicted outcome



[1] https://cran.r-project.org/web/packages/partykit/vignettes/constparty.pdf

Random forest

A single tree is often unstable (high variance); random forests average predictions across *heterogeneous* trees to reduce variance

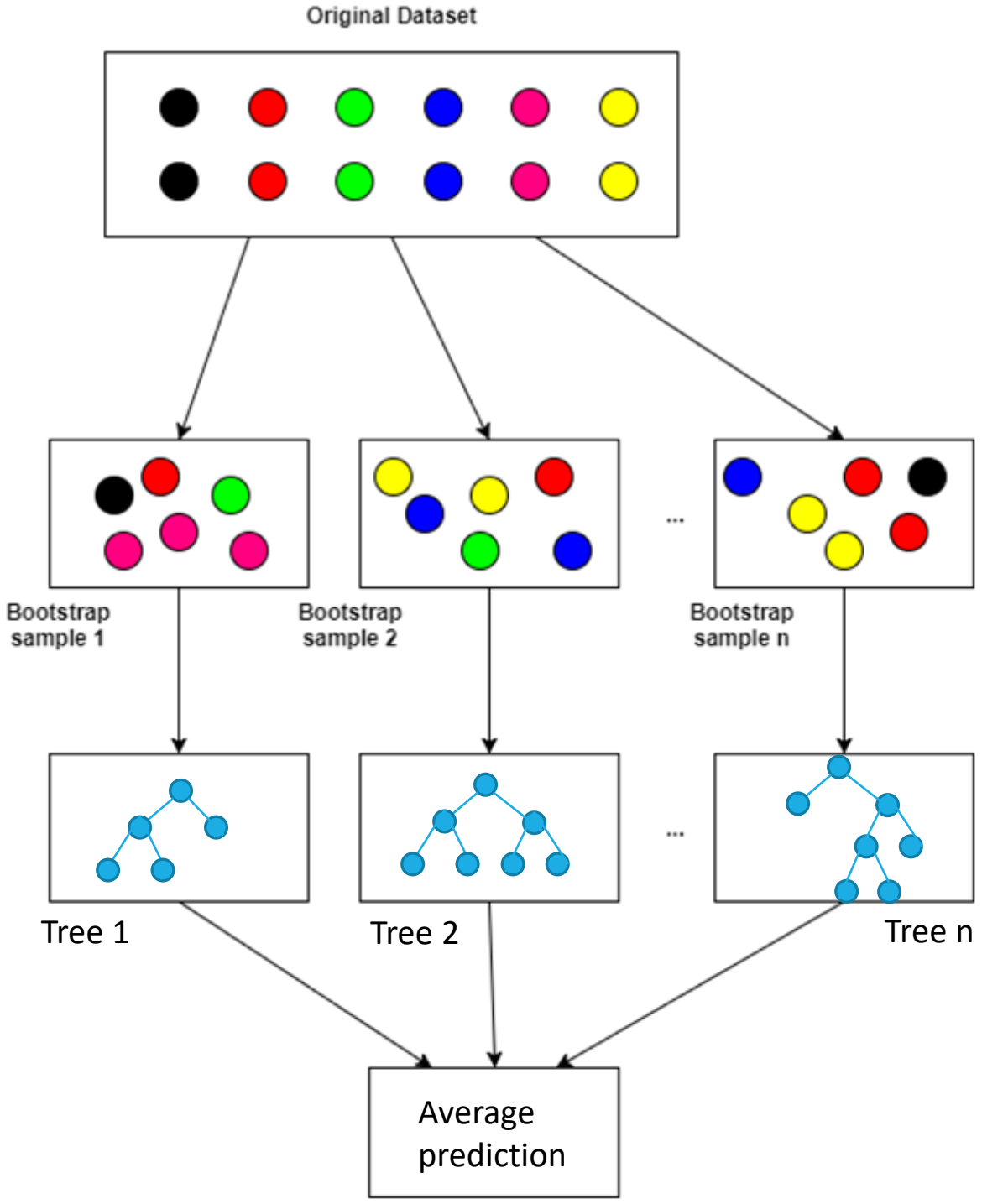


Figure modified from: https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-2-2b1fcd03e342

Prediction accuracy for survival data

Harrell's concordance index (**C-index**)[1]

- "probability that, in a randomly selected pair of cases, the case that fails first had a worse predicted outcome"[2]
- C-index = AUC for classification problems

Model	Average C-index	Method
LASSO	**0.70**	10 fold CV with 5 repeats
Random Forest	**0.64**	Bootstrap 3000 times

[1] Harrell, Frank E., et al. "Evaluating the yield of medical tests." *Jama* 247.18 (1982): 2543-2546

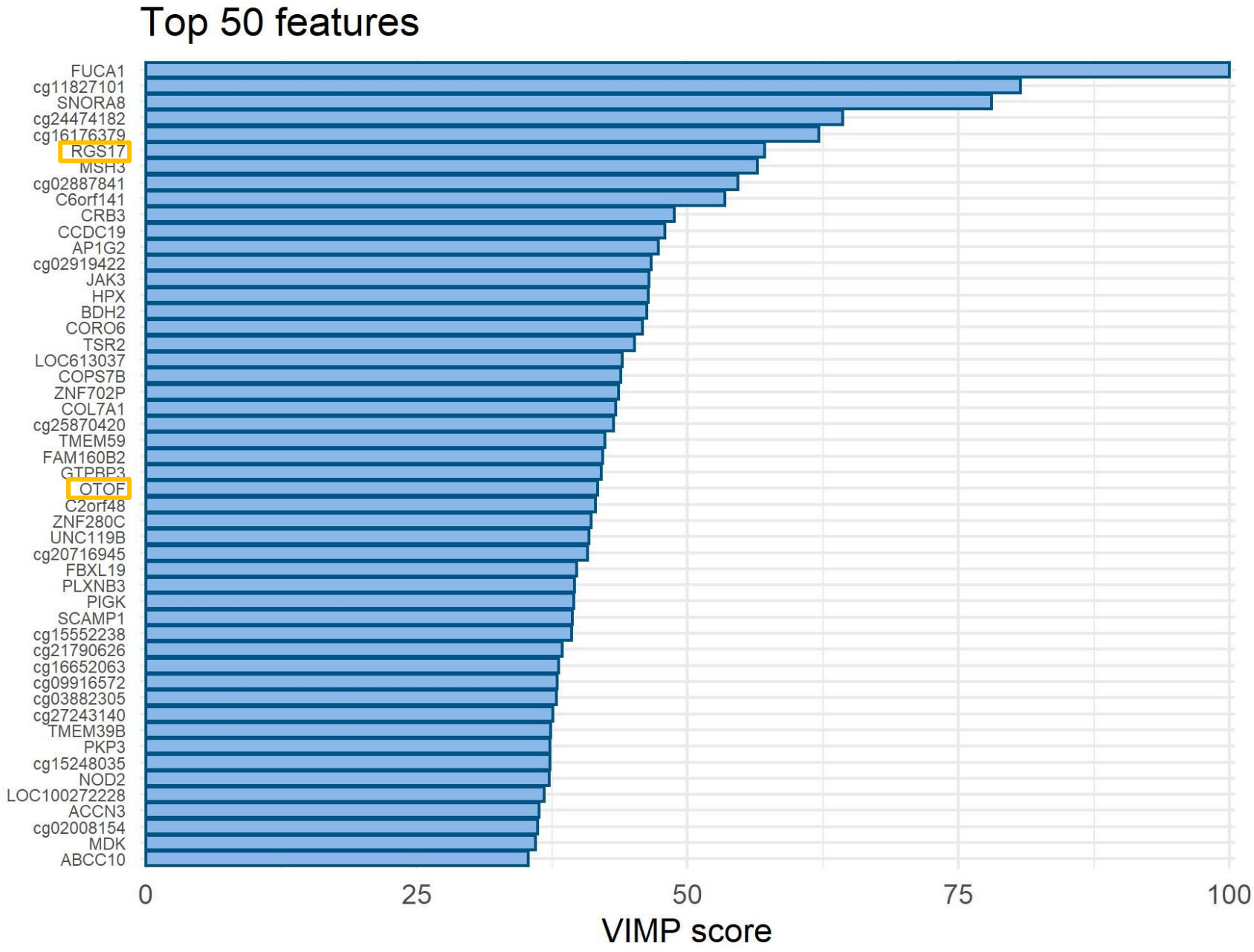
[2] Ishwaran, Hemant, et al. "Random survival forests." The annals of applied statistics 2.3 (2008): 841-860.

LASSO selected 3 variables

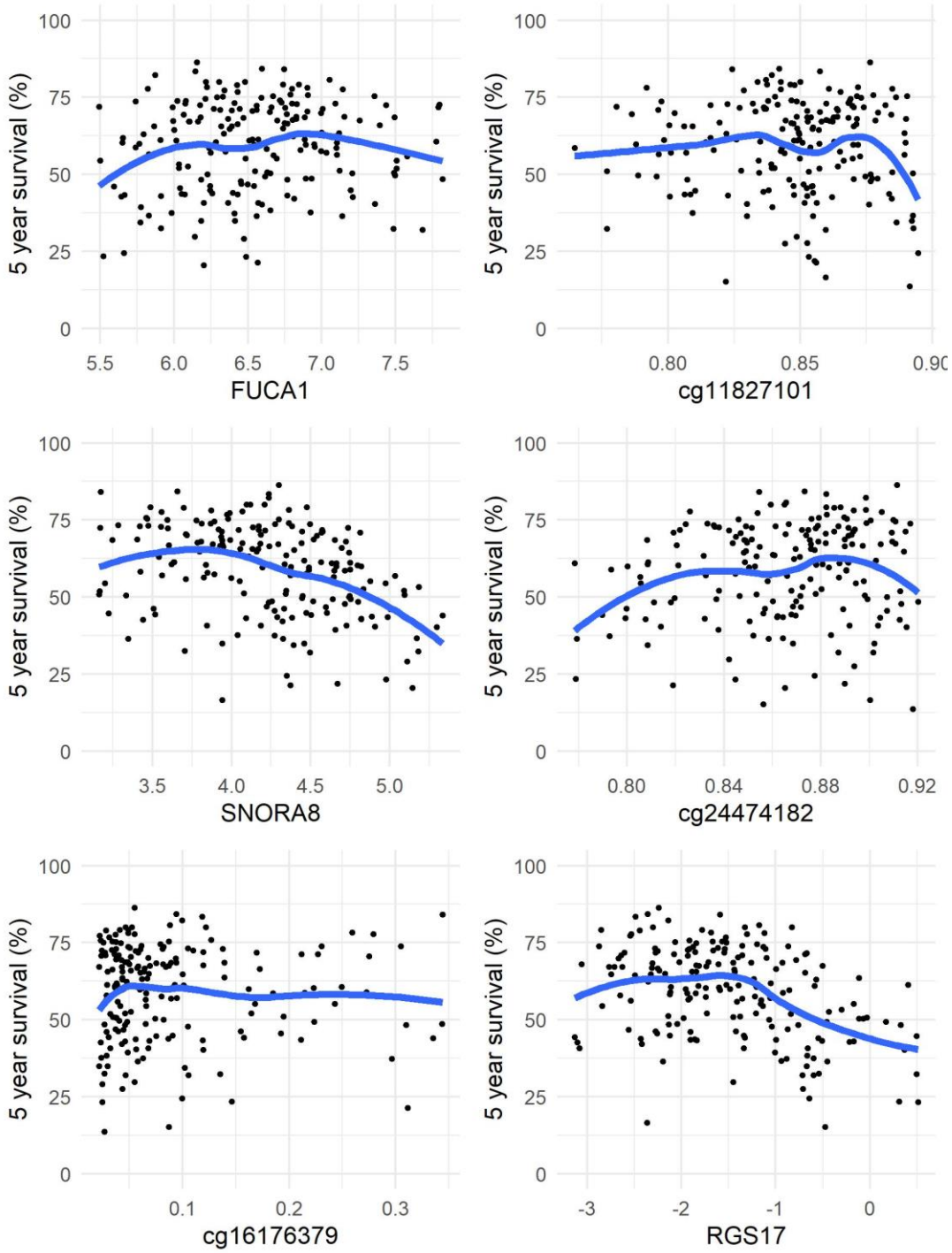
Variable	Hazard Ratio for 1 SD increase
OTOF	1.16
RGS17	1.14
PINK1	0.99

Random Forest Variable Importance (VIMP)

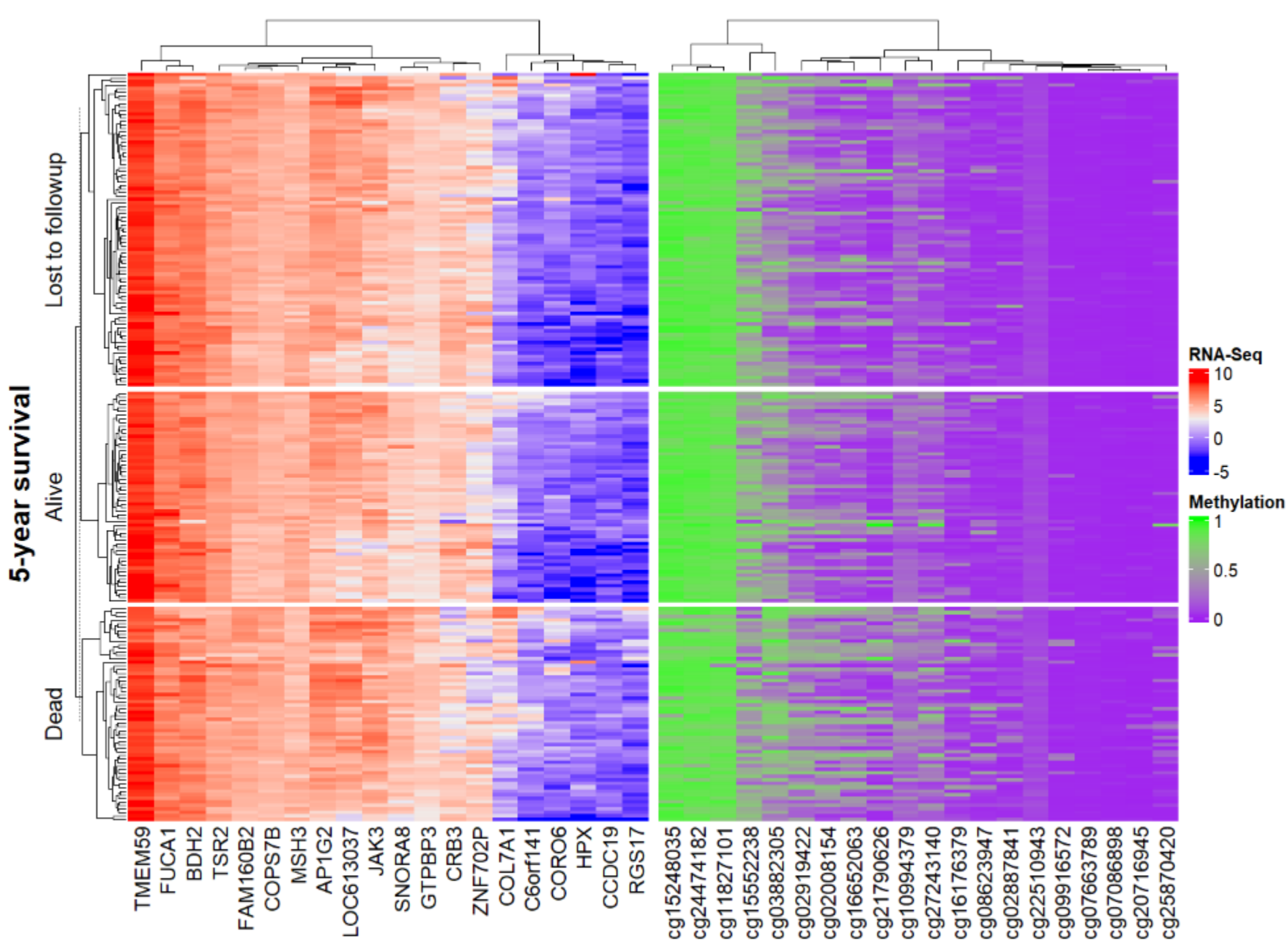
Permutation VIMP: compare prediction error of original model to prediction error after permuting (noising) X_j



RF variable effects on 5-year survival



RF top 20 genes and top 20 CGs vs 5-year survival



Summary

PCA: dimension reduction and visualize samples in 2-D space

sPLS/CCA: identify sets of genes that are correlated with CGs; could use as filtering step before network analysis

ML for survival:

- Regularized Cox model: easy to interpret (HRs and variable selection), more assumptions
- Random forests: more flexible but harder to interpret (use VIMP, partial dependence plots)

Other ideas for analyzing multi-omics

- Clustering to derive disease subtypes - 2 recent review papers:

  Pierre-Jean, Morgane, et al. "Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration." Briefings in bioinformatics 21.6 (2020): 2011-2030.

  Chauvel, Cécile, et al. "Evaluation of integrative clustering methods for the analysis of multi-omics data." Briefings in bioinformatics 21.2 (2020): 541-552.

- Network analysis – derive correlated multi-omics “modules”

  Yan, Jingwen, et al. "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data." Briefings in bioinformatics 19.6 (2018): 1370-1381.

  Shi, W. Jenny, et al. "Unsupervised discovery of phenotype-specific multi-omics networks." Bioinformatics 35.21 (2019): 4336-4343.