

MACHINE-LEARNING BASED MULTI-OMICS DATA INTEGRATION FOR PERSONALIZED ONCOLOGY

Florian Buettner



DKTK

German Cancer
Consortium



dkfz.

German Cancer Consortium
Partner site Frankfurt|Mainz

LATENT VARIABLE MODELS AS DISCOVERY TOOL

LATENT VARIABLE MODELS AS DISCOVERY TOOL

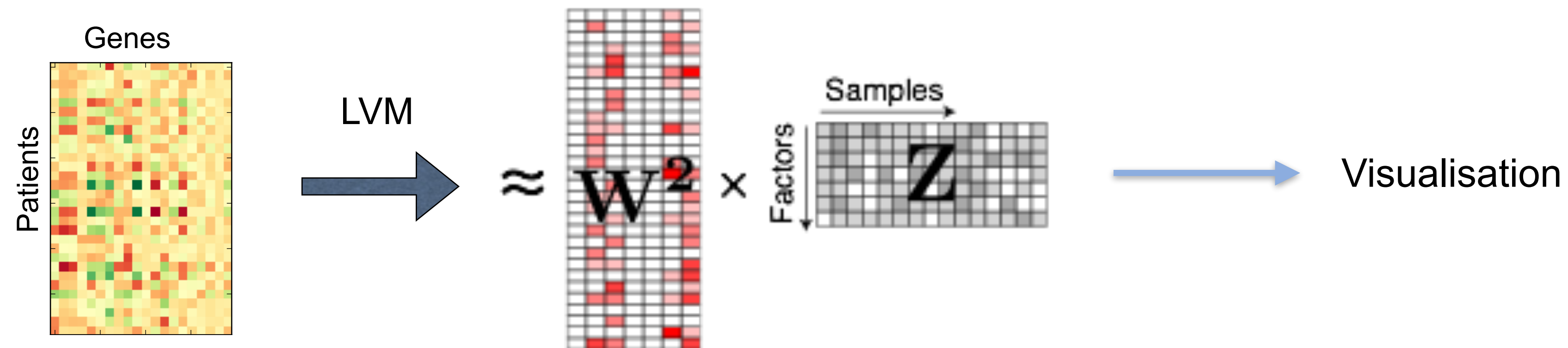
- Omics profiling uncovers novel variation between patients

LATENT VARIABLE MODELS AS DISCOVERY TOOL

- Omics profiling uncovers novel variation between patients
- Latent variable models
 - Statistical tool to infer an unobserved, hidden state of a biological system based on observable data that is often high-dimensional
 - Reduce a high-dimensional dataset of correlated observations into a low-dimensional dataset of uncorrelated and interpretable latent variables

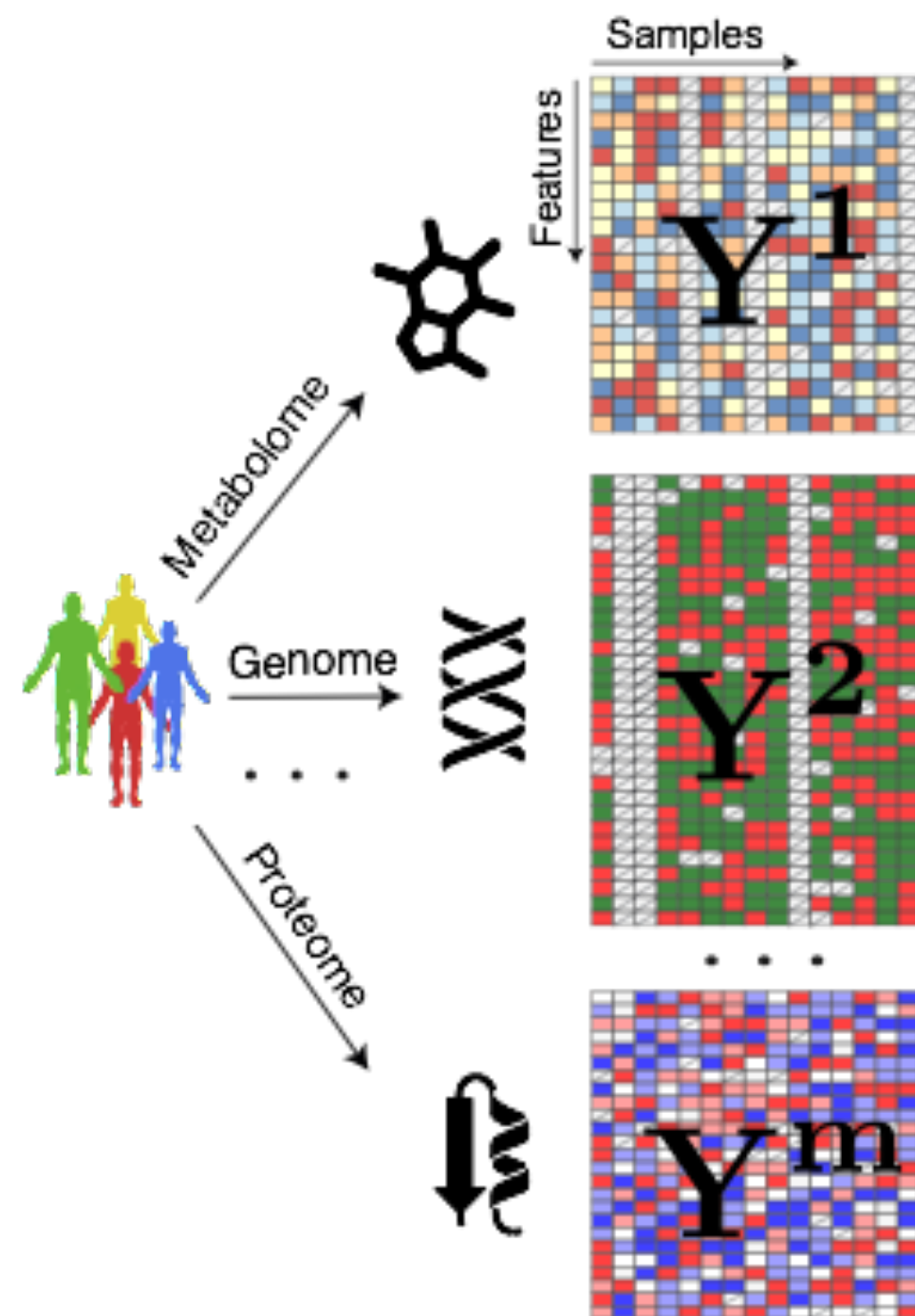
LATENT VARIABLE MODELS AS DISCOVERY TOOL

- Omics profiling uncovers novel variation between patients
- Latent variable models
 - Statistical tool to infer an unobserved, hidden state of a biological system based on observable data that is often high-dimensional
 - Reduce a high-dimensional dataset of correlated observations into a low-dimensional dataset of uncorrelated and interpretable latent variables



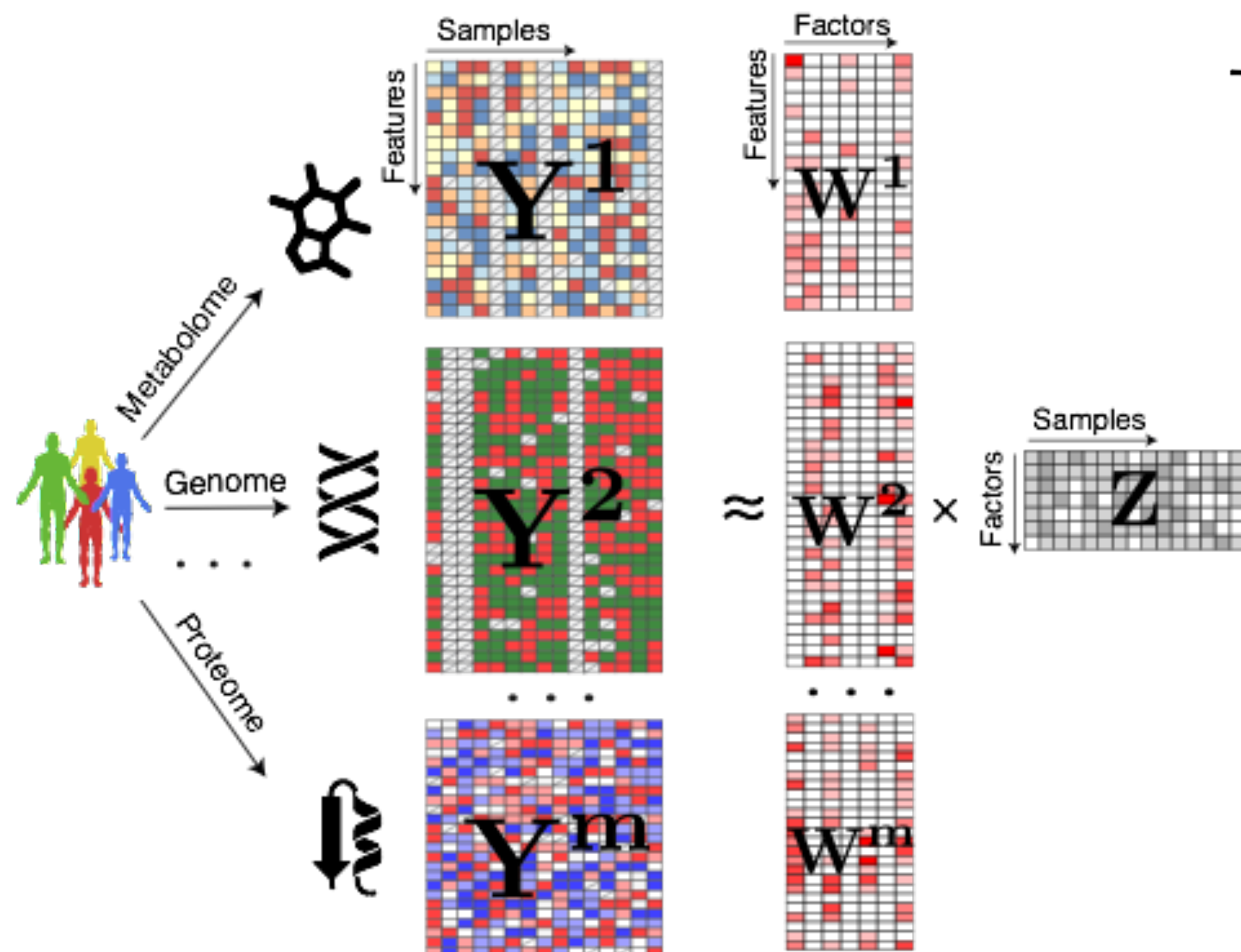
Buettner et al Genome Biology, 2017

MULTI-OMICS FACTOR ANALYSIS



Arguelaget, ..., Huber#/Buettner#/Stegle#, 2018

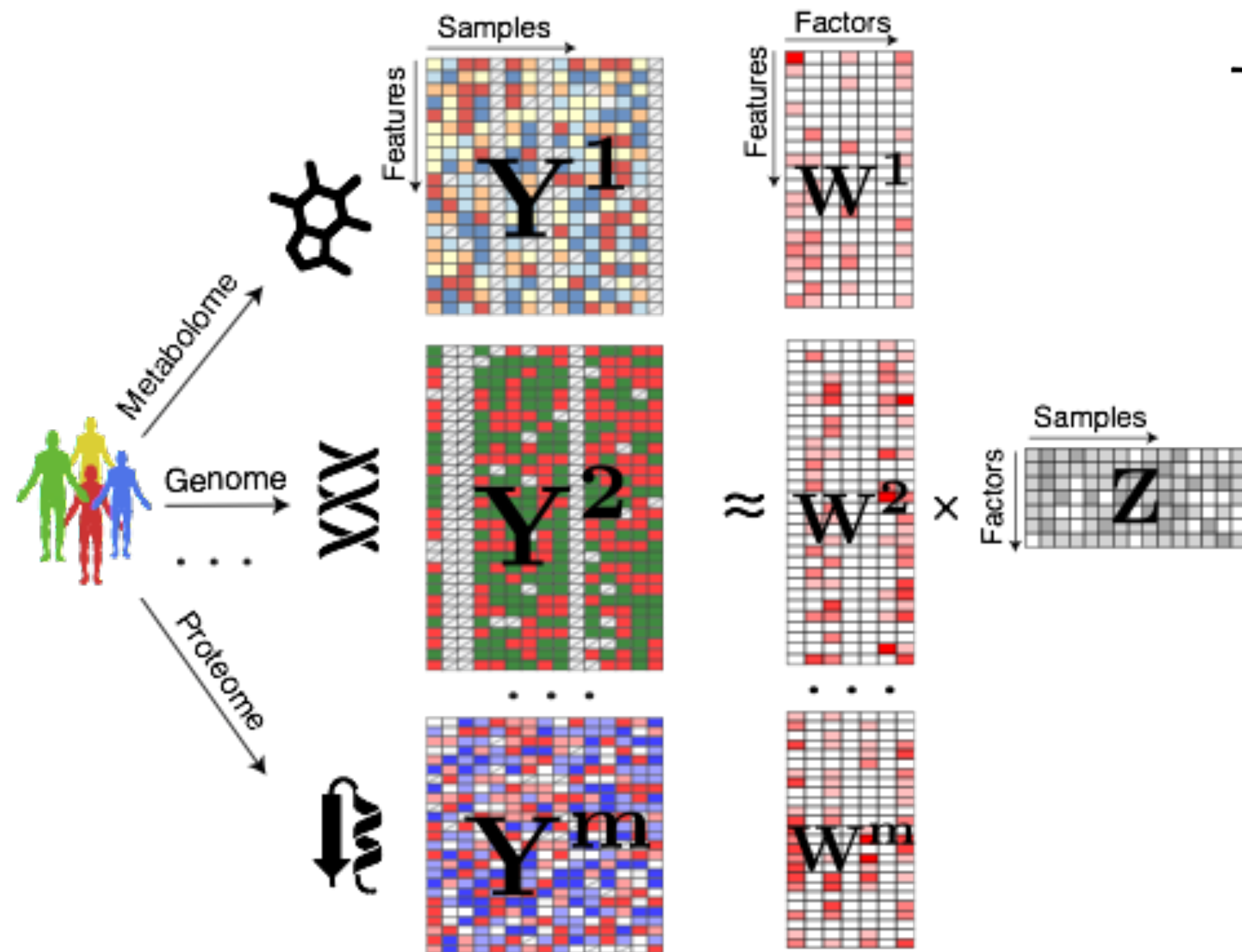
MULTI-OMICS FACTOR ANALYSIS



$$Y^m = ZW^mT$$

Arguelaget, ..., Huber#/Buettner#/Stegle#, 2018

MULTI-OMICS FACTOR ANALYSIS



$$Y^m = ZW^mT$$

- Identify drivers of variation that are shared between omics layers or unique
- Model binary/count data
- Account for missing values
- Approximate Bayesian inference for scalability

Arguelaget, ..., Huber#/Buettner#/Stegle#, 2018

PROTEOGENOMIC CHARACTERISATION OF AML

Jayavelu*, Wolf*, Buettner*, ... ,Oellerich, Cancer Cell 2022

PROTEOGENOMIC CHARACTERISATION OF AML

- Acute myeloid leukemia (AML)
 - Treatment decisions based on genomic classification (~15 mutations)
 - Large fraction in outcome variability remains unexplained

Jayavelu*, Wolf*, Buettner*, ... ,Oellerich, Cancer Cell 2022

PROTEOGENOMIC CHARACTERISATION OF AML

- Acute myeloid leukemia (AML)
 - Treatment decisions based on genomic classification (~15 mutations)
 - Large fraction in outcome variability remains unexplained
- Hypothesis
 - Patient subpopulations with distinct survival profile defined by multi-ome

Jayavelu*, Wolf*, Buettner*, ... ,Oellerich, Cancer Cell 2022

PROTEOGENOMIC CHARACTERISATION OF AML

- Acute myeloid leukemia (AML)
 - Treatment decisions based on genomic classification (~15 mutations)
 - Large fraction in outcome variability remains unexplained
- Hypothesis
 - Patient subpopulations with distinct survival profile defined by multi-ome
- Approach
 - Two independent patient cohorts
 - Discovery cohort (177 patients) and validation cohort (~70 patients), profiled with different technologies
 - Measure Common Mutations, Transcriptome (RNA), Proteome, other risk factors
 - Record survival

Jayavelu*, Wolf*, Buettner*, ... ,Oellerich, Cancer Cell 2022

PROTEOMIC CHARACTERISATION OF PATIENTS

- Compute patient-patient distances and use clustering approach to define patient subpopulations
- Characterize subpopulations via pathway analysis

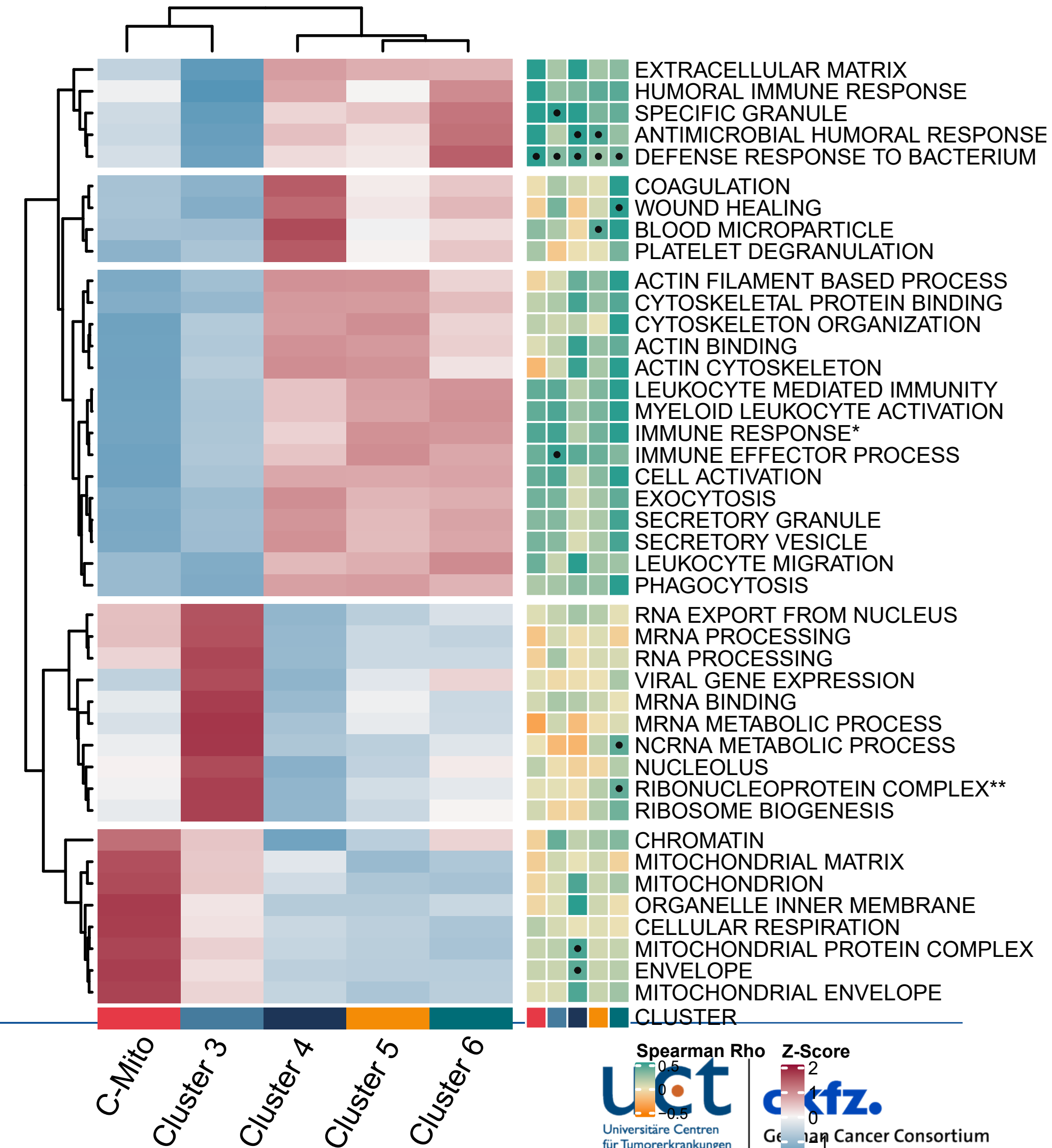


CHARACTERIZATION OF SUBPOPULATIONS

- Use pre-annotated pathways to identify biological processes driving differences between clusters
- Comparative cluster analysis
- Gene set variation analysis (GSVA)

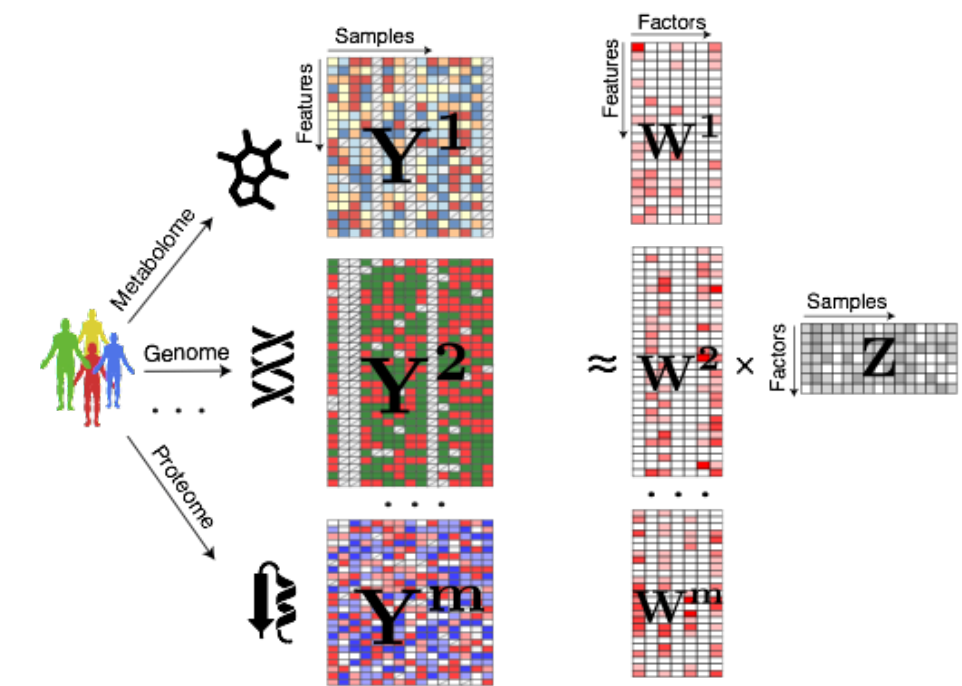
CHARACTERIZATION OF SUBPOPULATIONS

- Use pre-annotated pathways to identify biological processes driving differences between clusters
- Comparative cluster analysis
- Gene set variation analysis (GSVA)



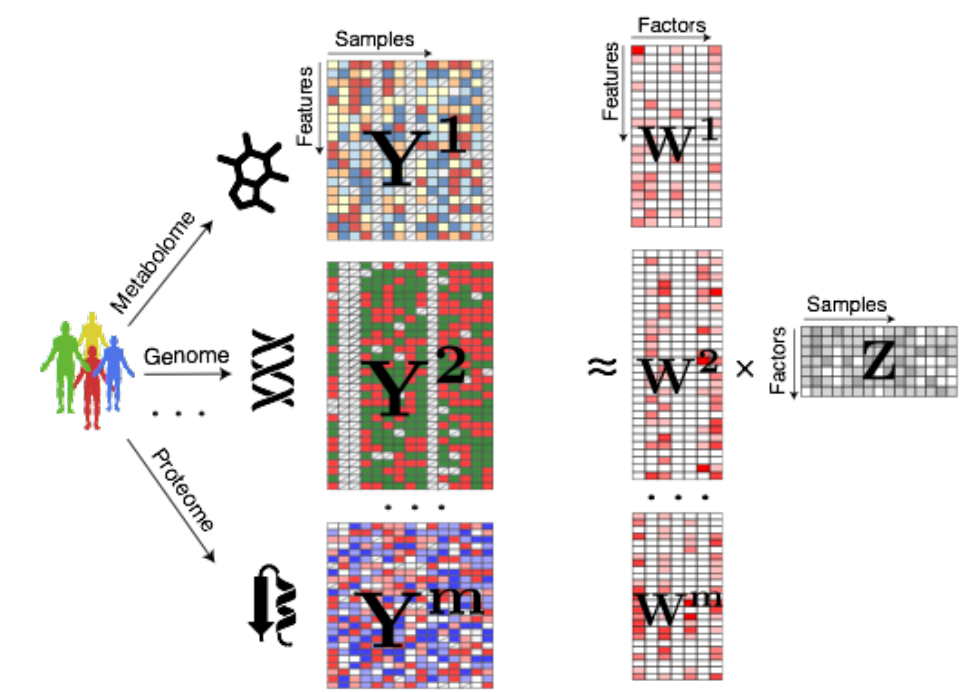
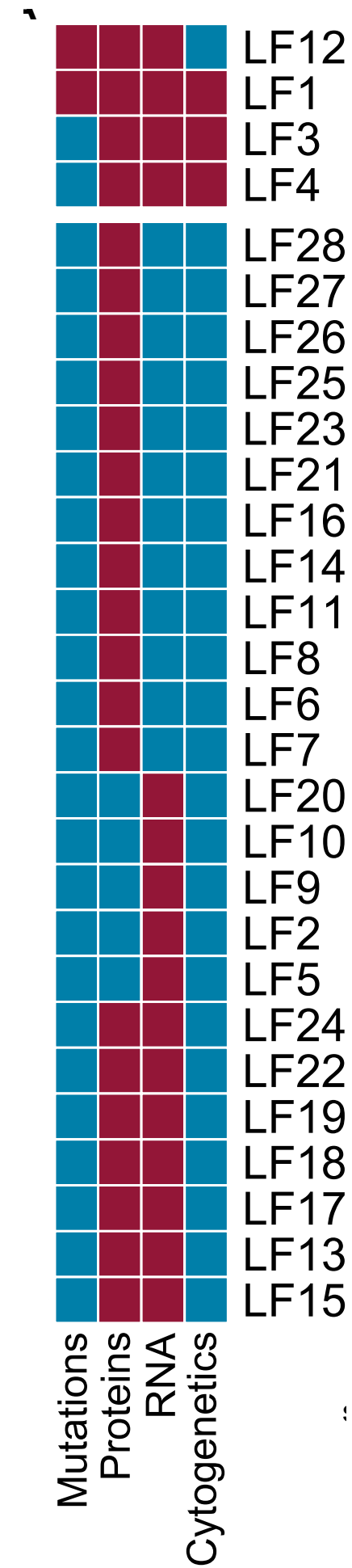
MOFA IDENTIFIES KNOWN DRIVERS OF VARIABILITY

- 28 factors explaining variation in the data
- Mix of shared and unique drivers
- Most dominant factor LF1 active in all 4 data views
- Inspections of loadings reveals LF1 explains known axis of HOX/NPM1



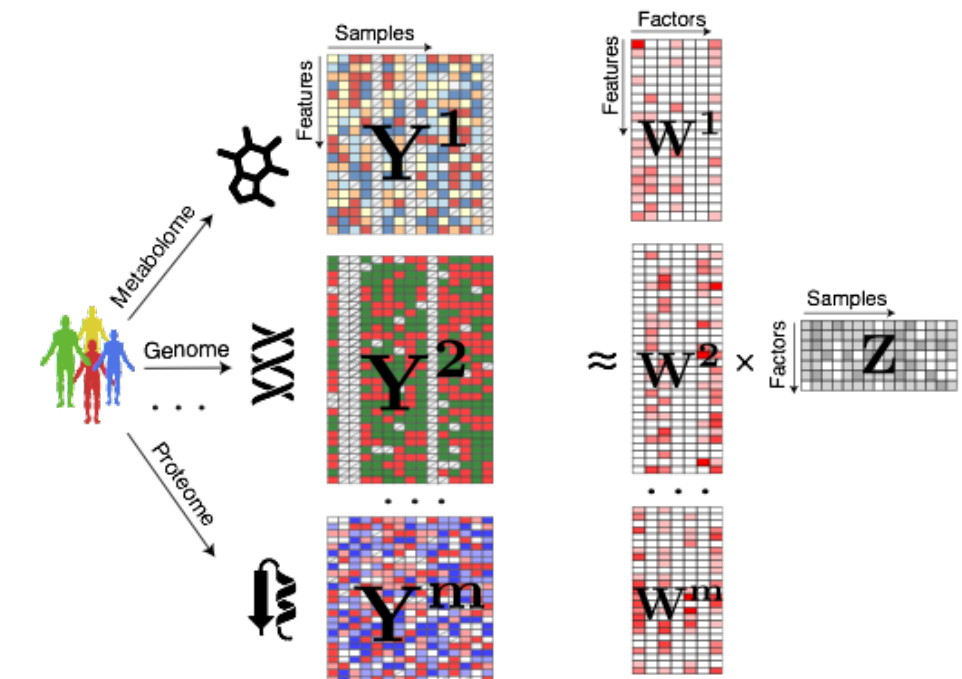
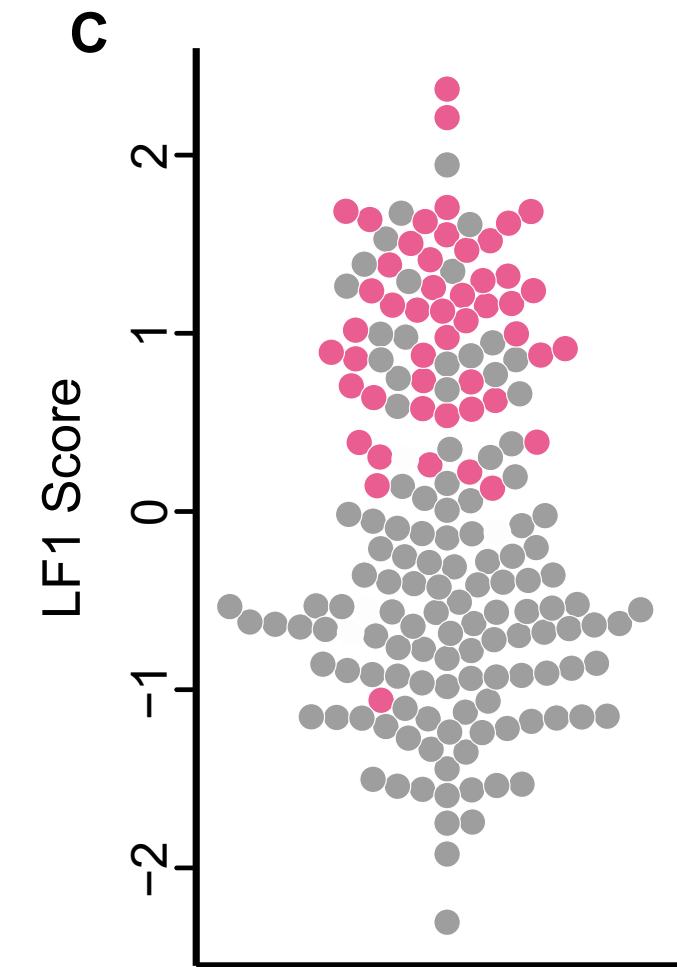
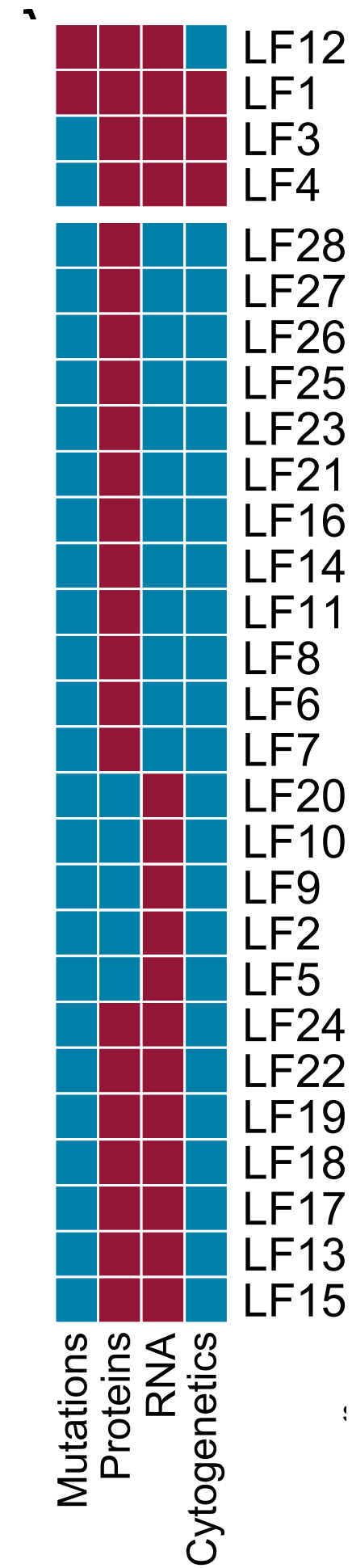
MOFA IDENTIFIES KNOWN DRIVERS OF VARIABILITY

- 28 factors explaining variation in the data
- Mix of shared and unique drivers
- Most dominant factor LF1 active in all 4 data views
- Inspections of loadings reveals LF1 explains known axis of HOX/NPM1



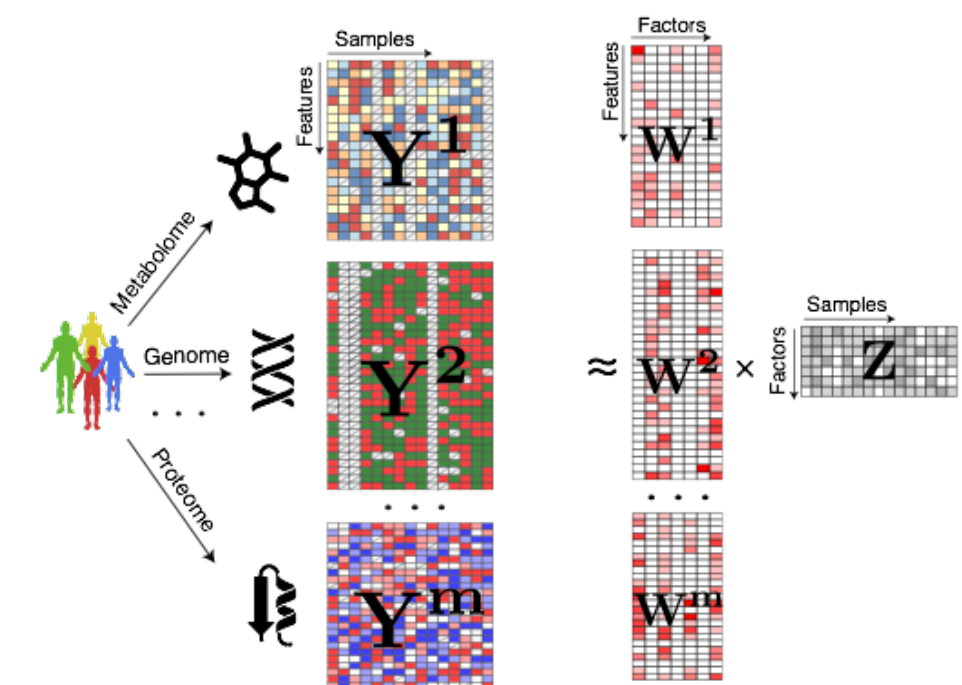
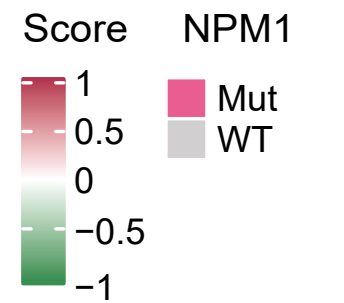
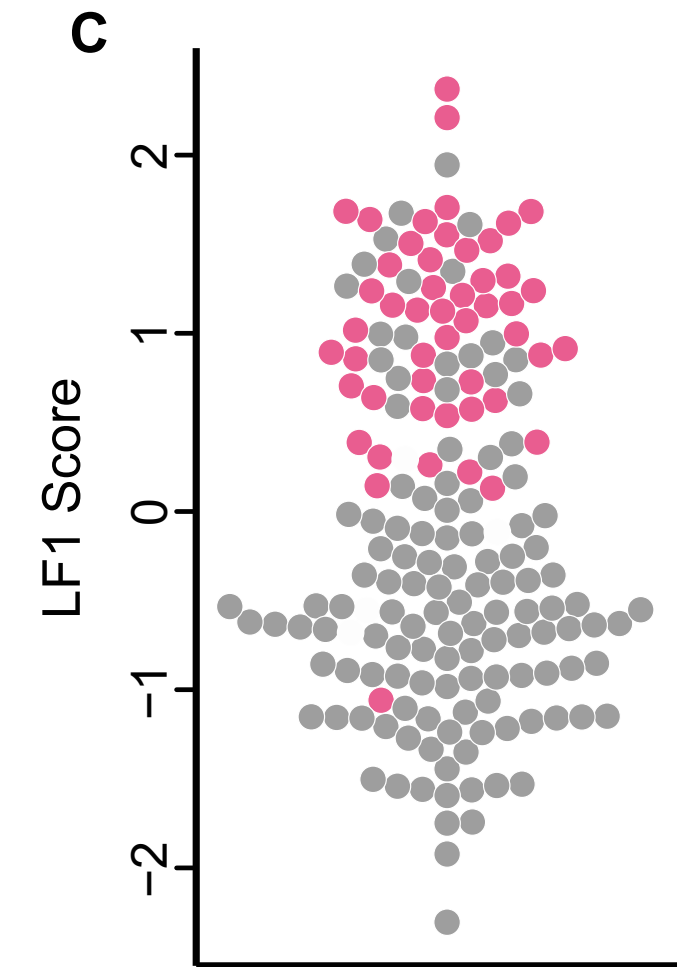
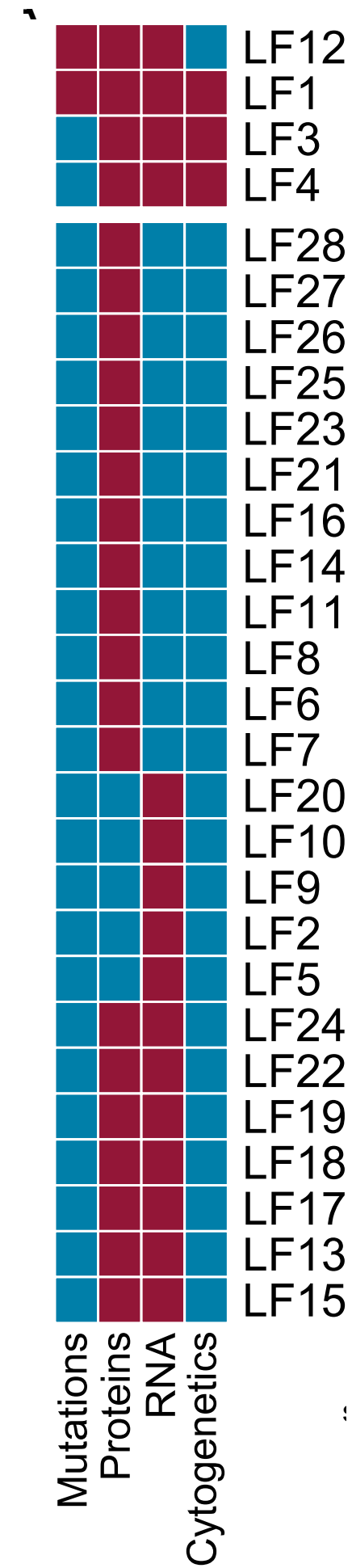
MOFA IDENTIFIES KNOWN DRIVERS OF VARIABILITY

- 28 factors explaining variation in the data
- Mix of shared and unique drivers
- Most dominant factor LF1 active in all 4 data views
- Inspections of loadings reveals LF1 explains known axis of HOX/NPM1



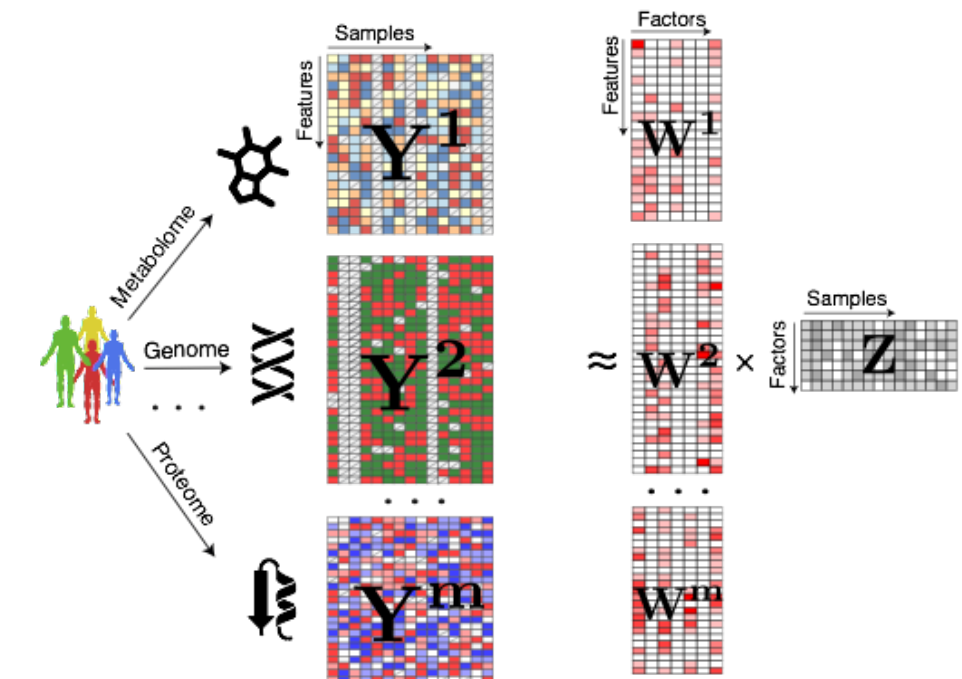
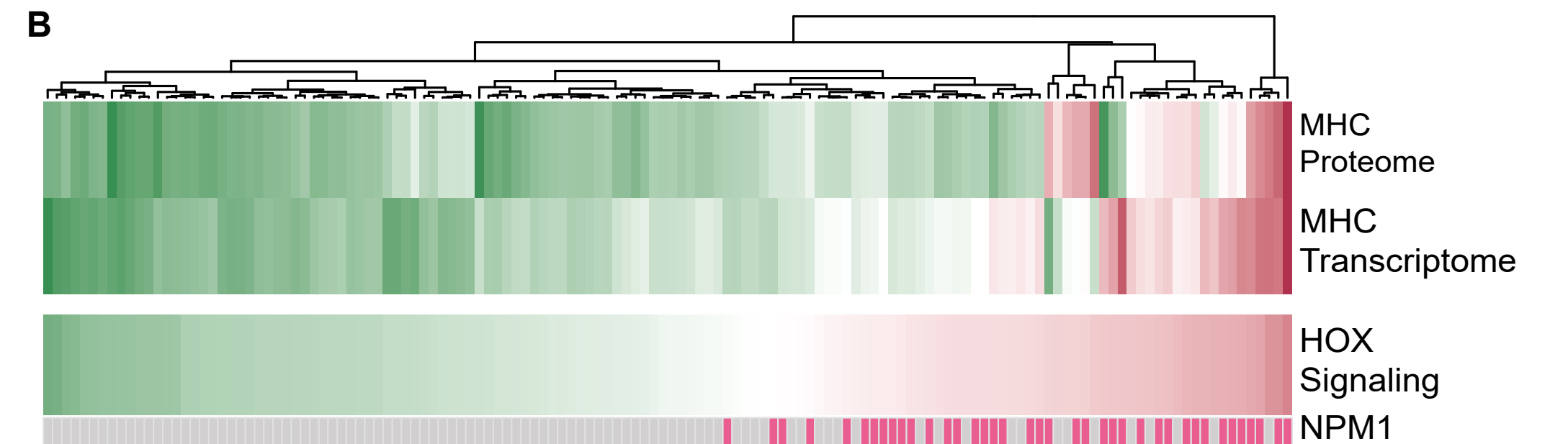
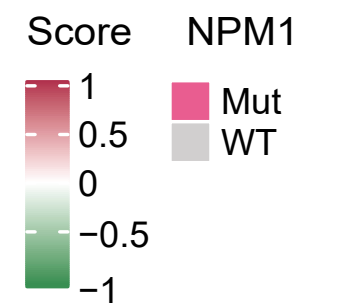
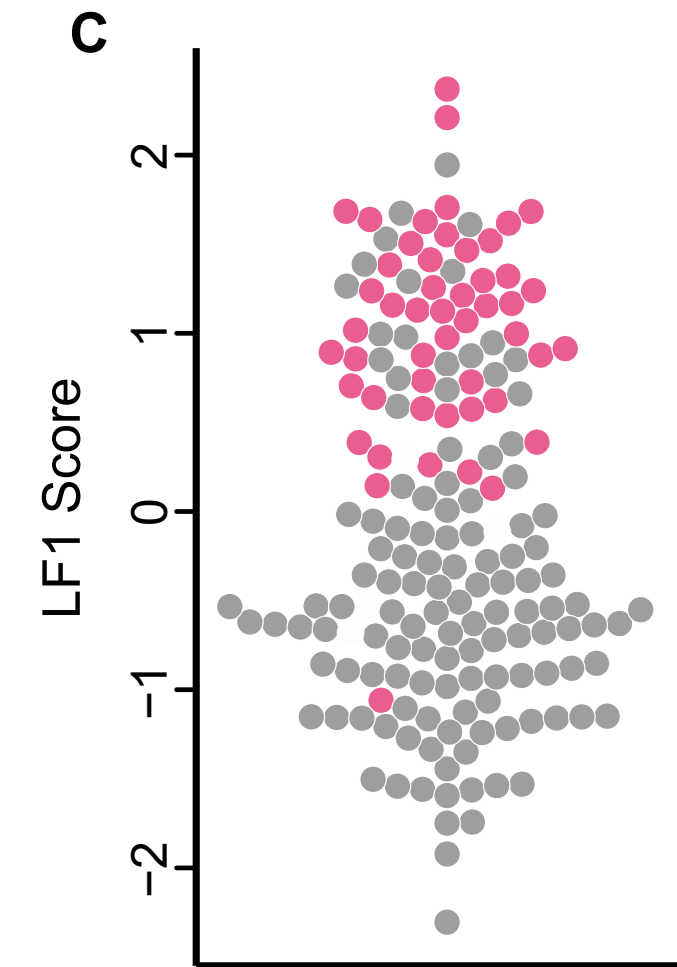
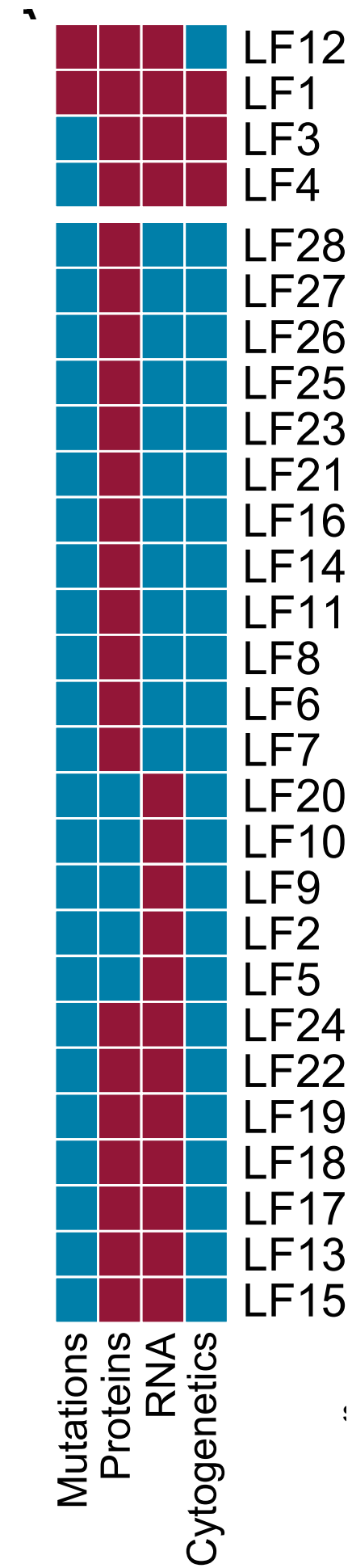
MOFA IDENTIFIES KNOWN DRIVERS OF VARIABILITY

- 28 factors explaining variation in the data
- Mix of shared and unique drivers
- Most dominant factor LF1 active in all 4 data views
- Inspections of loadings reveals LF1 explains known axis of HOX/NPM1



MOFA IDENTIFIES KNOWN DRIVERS OF VARIABILITY

- 28 factors explaining variation in the data
- Mix of shared and unique drivers
- Most dominant factor LF1 active in all 4 data views
- Inspections of loadings reveals LF1 explains known axis of HOX/NPM1

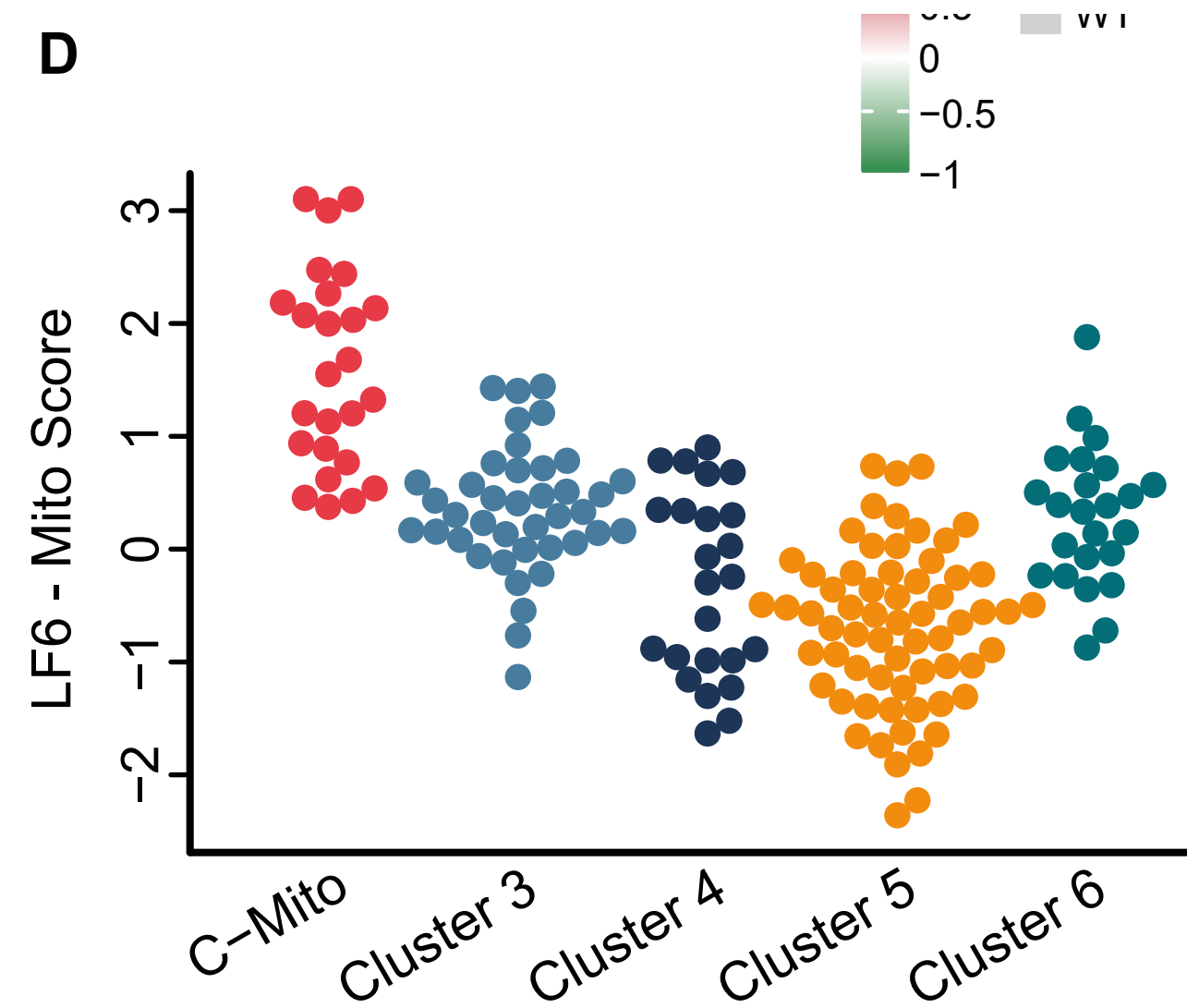


MOFA IDENTIFIES NEW DRIVERS OF VARIATION

- LF6 most dominant factor active in only one view (proteome)
- LF6 represents same Mito processes form cluster 1

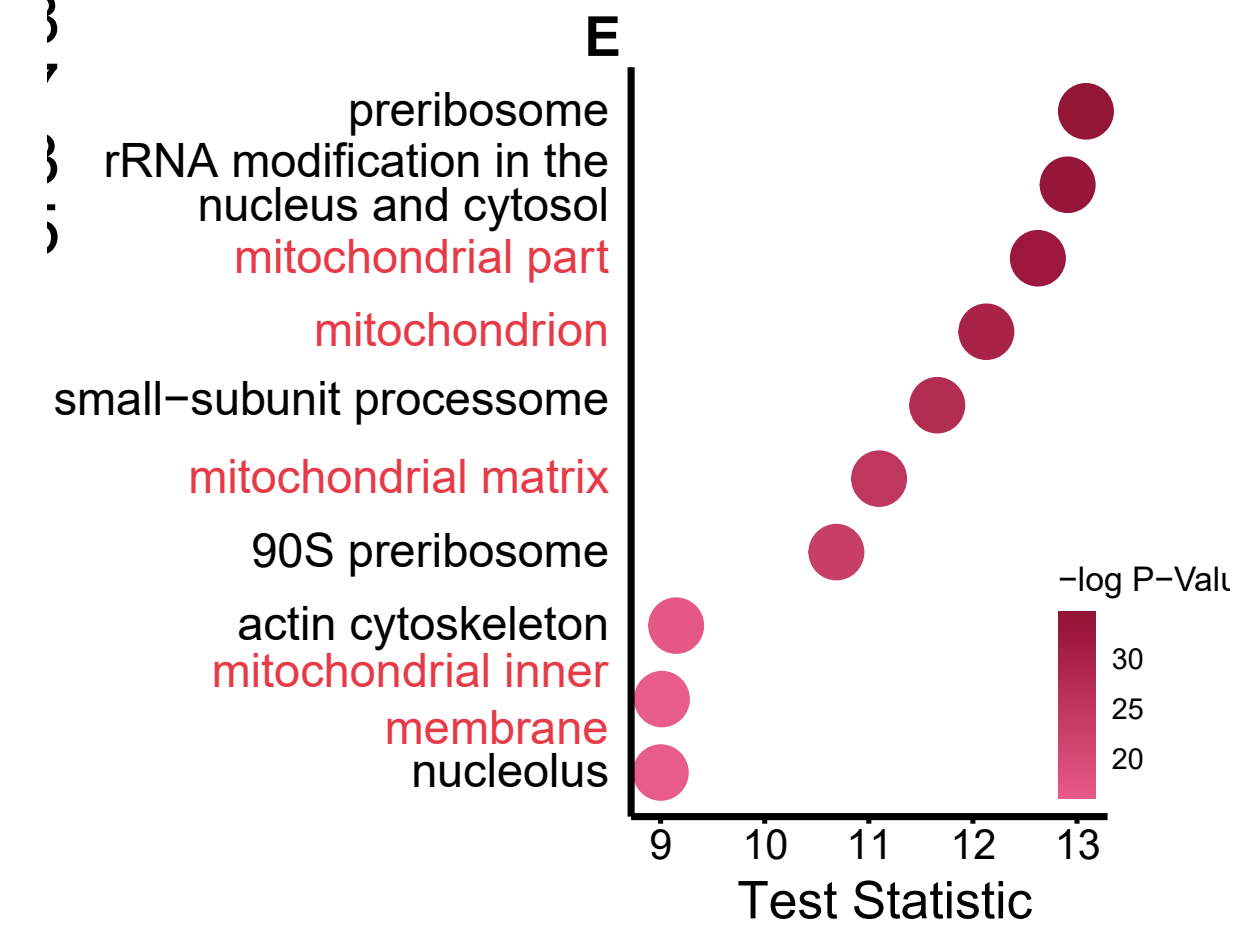
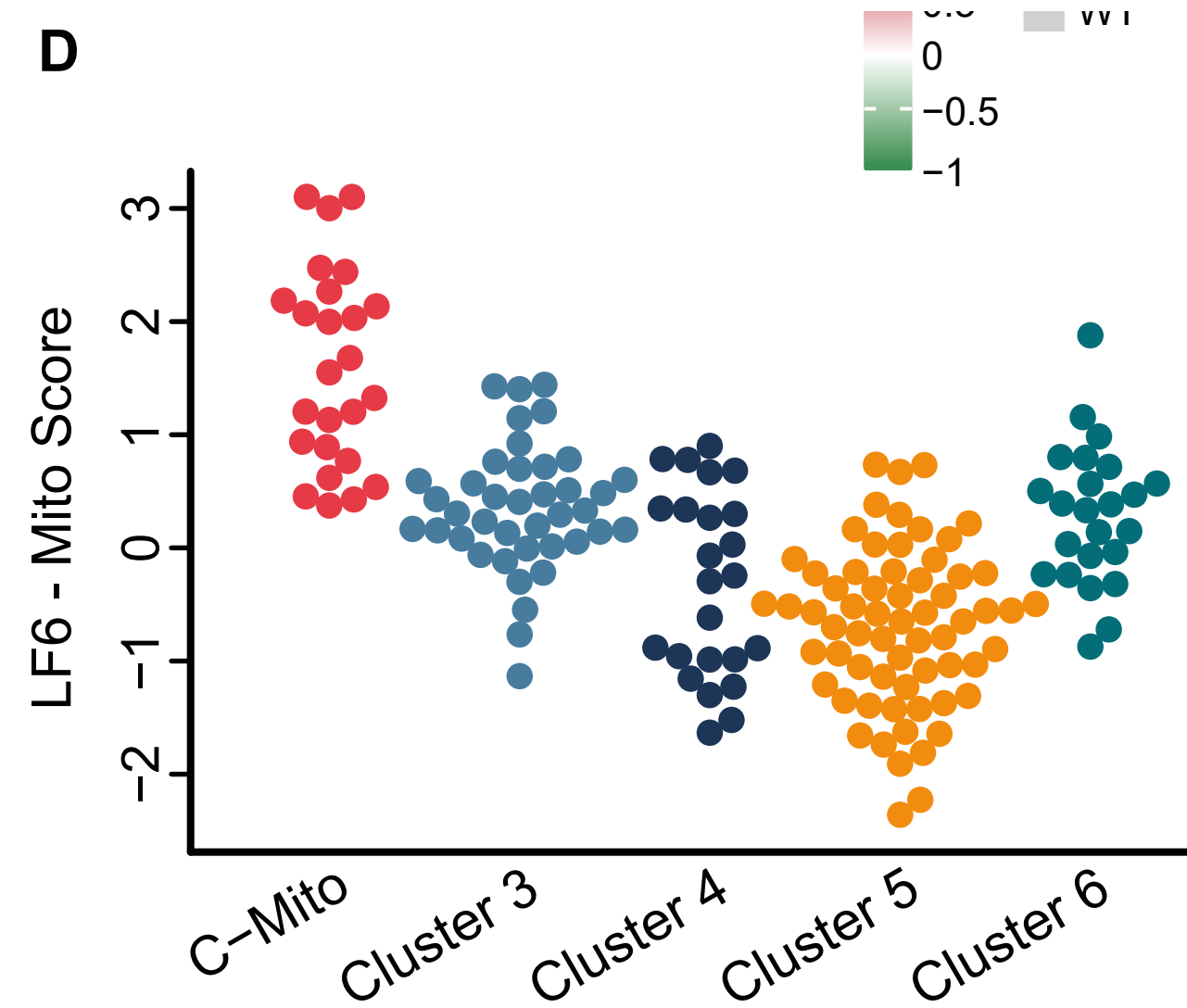
MOFA IDENTIFIES NEW DRIVERS OF VARIATION

- LF6 most dominant factor active in only one view (proteome)
- LF6 represents same Mito processes form cluster 1



MOFA IDENTIFIES NEW DRIVERS OF VARIATION

- LF6 most dominant factor active in only one view (proteome)
- LF6 represents same Mito processes form cluster 1

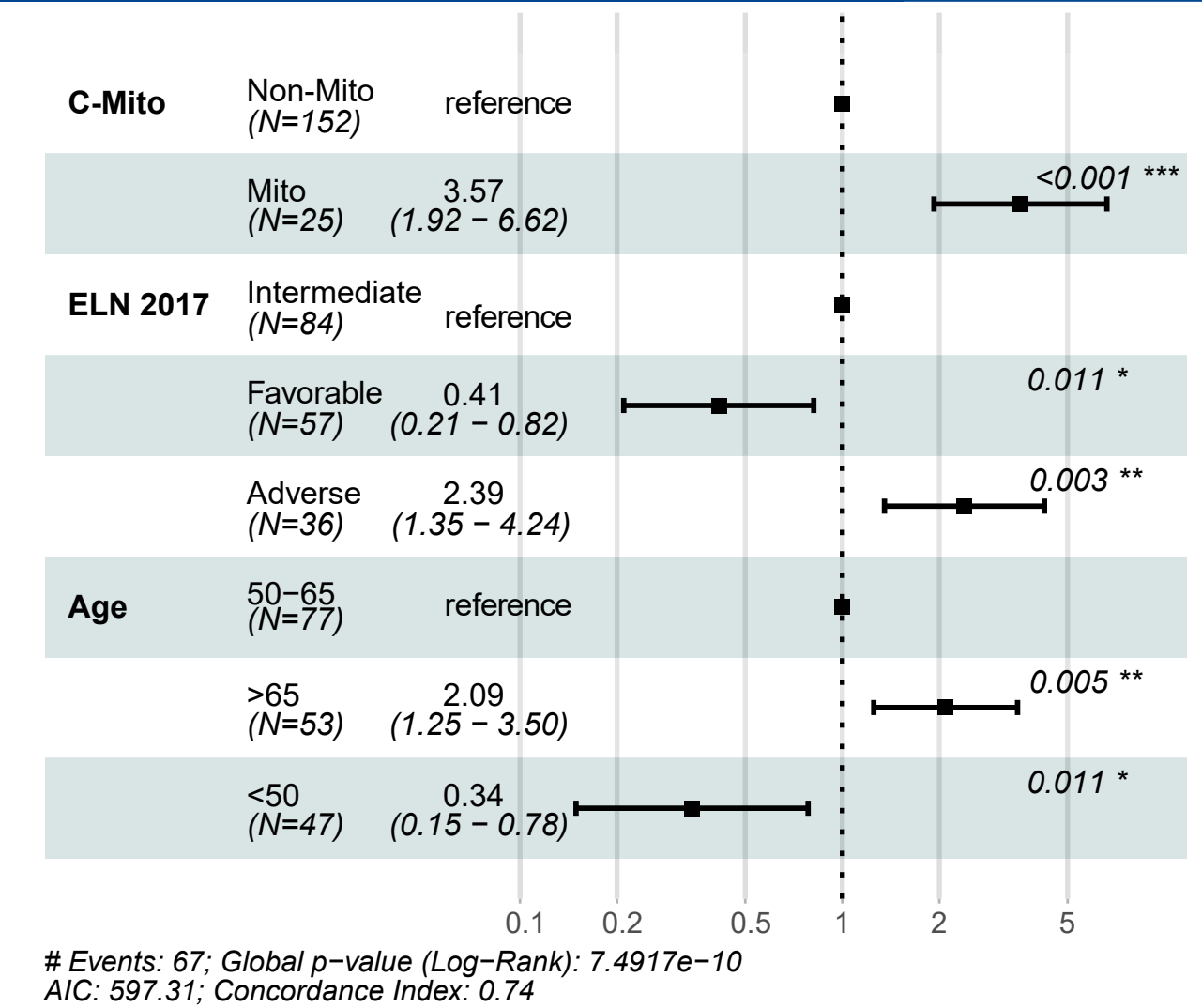


SUBCLUSTERS HAVE A DISTINCT SURVIVAL PROFILE

- Clusters 1 and 5 have distinct survival phenotype
- Survival for patients in cluster 1 significantly worse, for cluster 5 significantly better

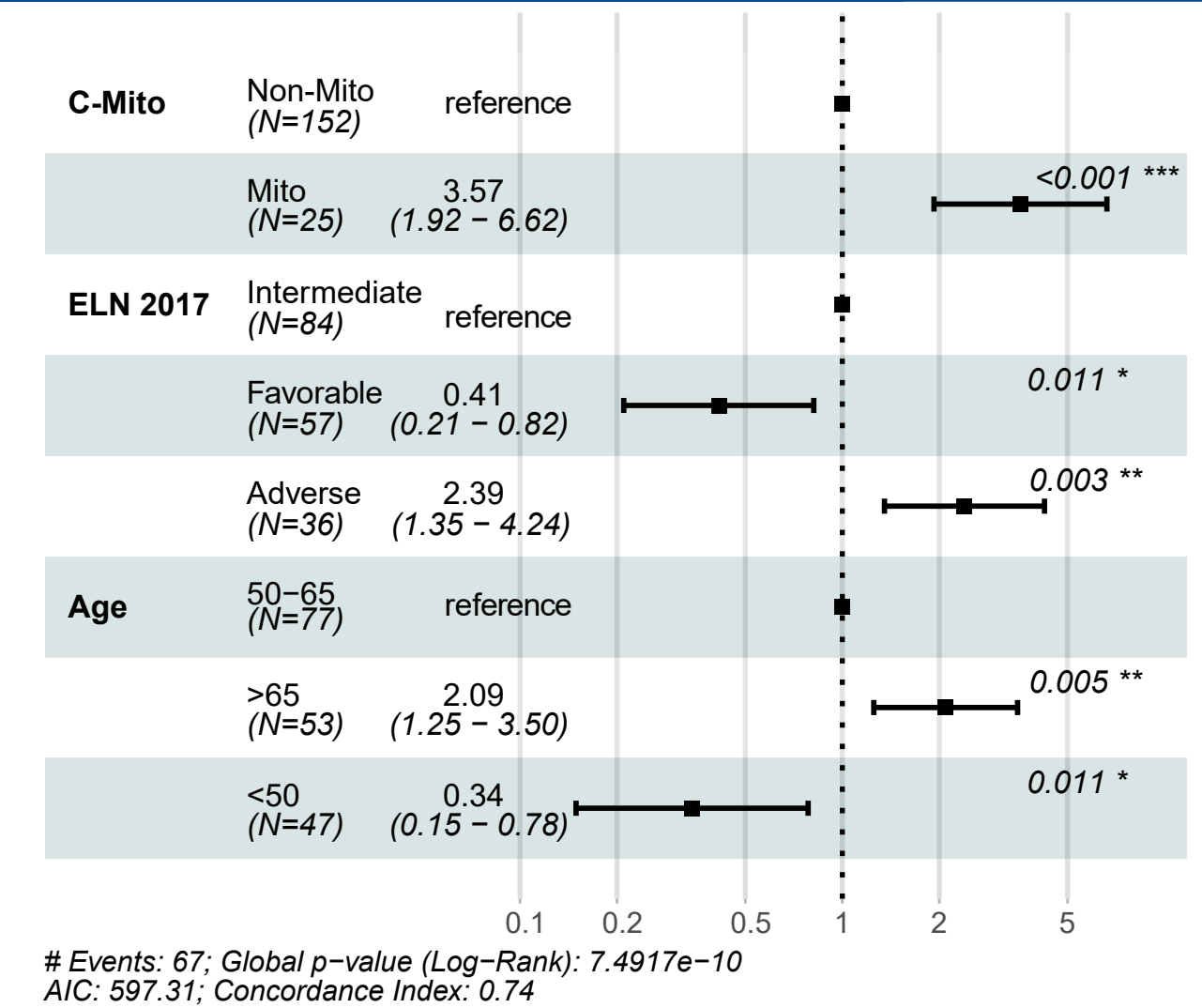
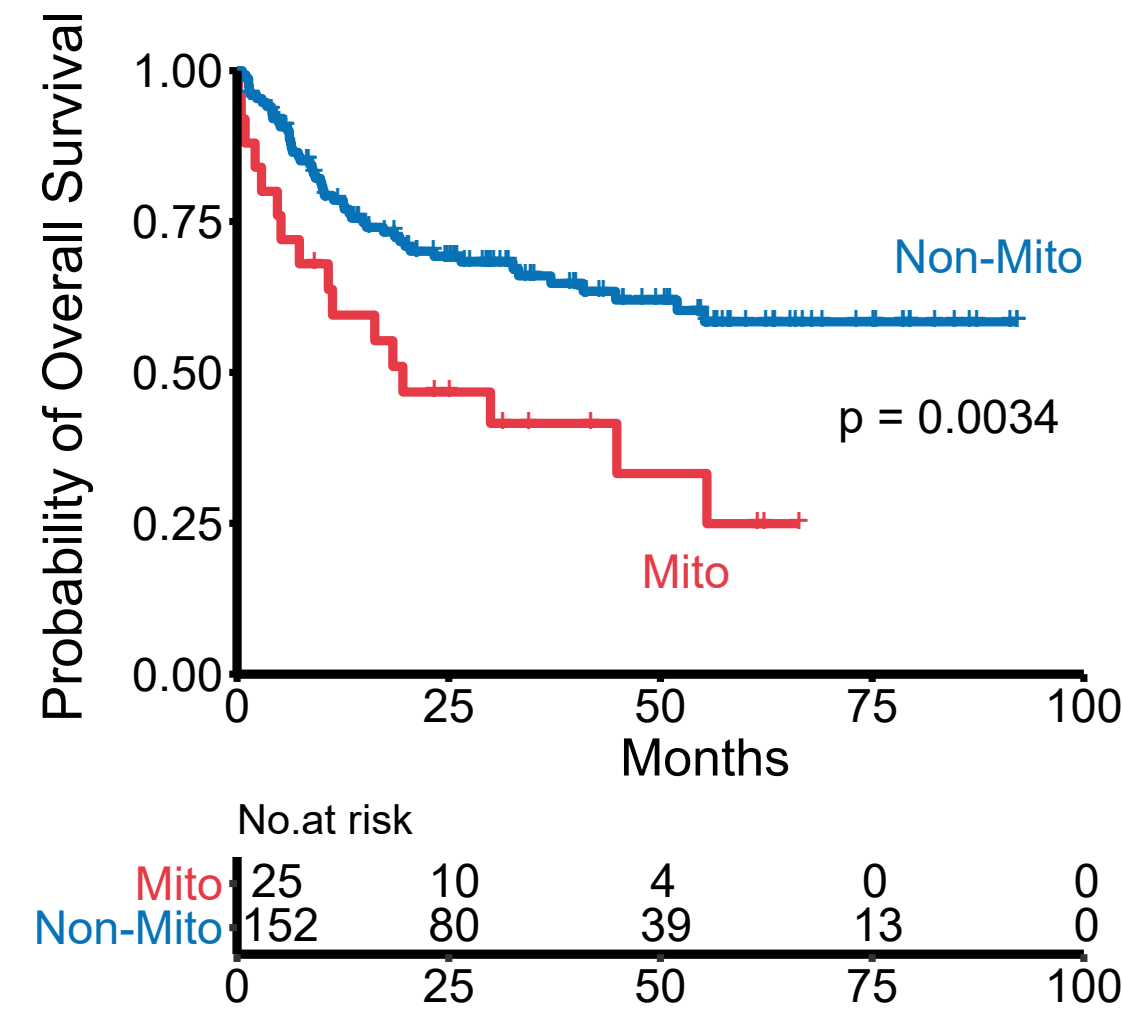
SUBCLUSTERS HAVE A DISTINCT SURVIVAL PROFILE

- Clusters 1 and 5 have distinct survival phenotype
- Survival for patients in cluster 1 significantly worse, for cluster 5 significantly better



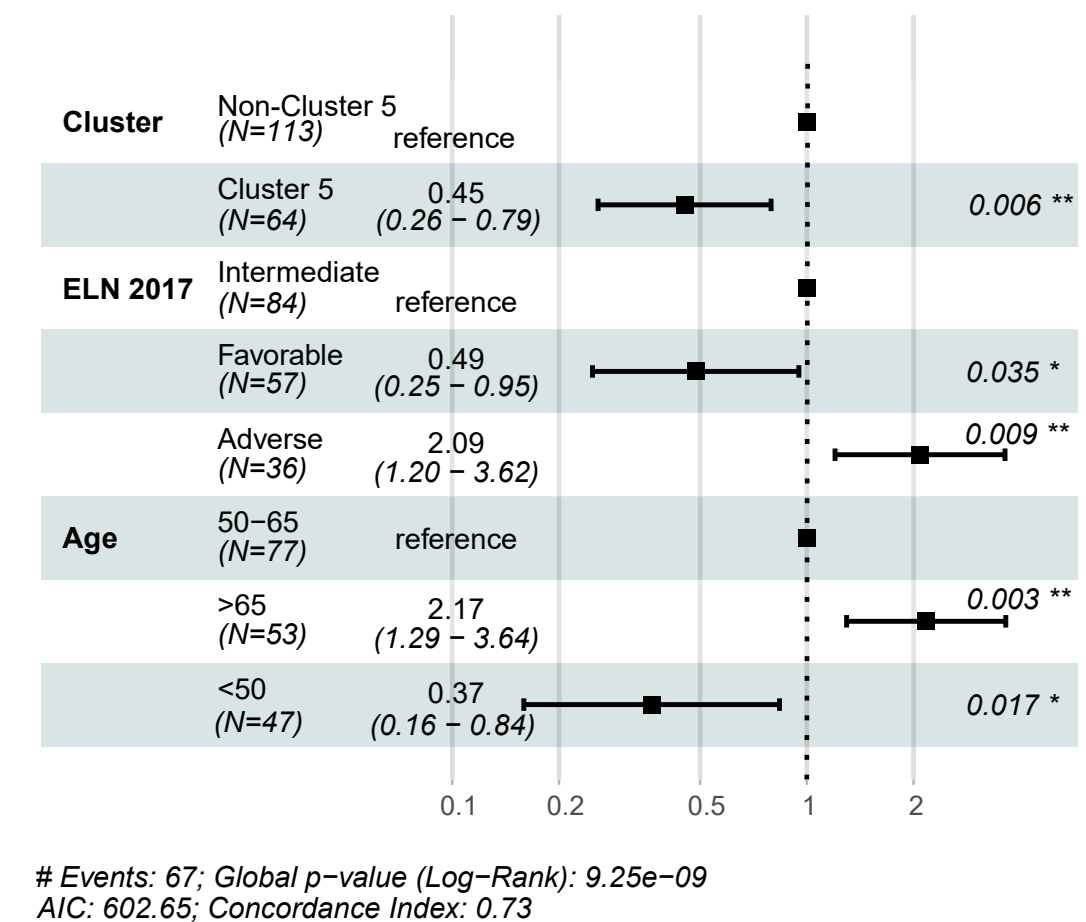
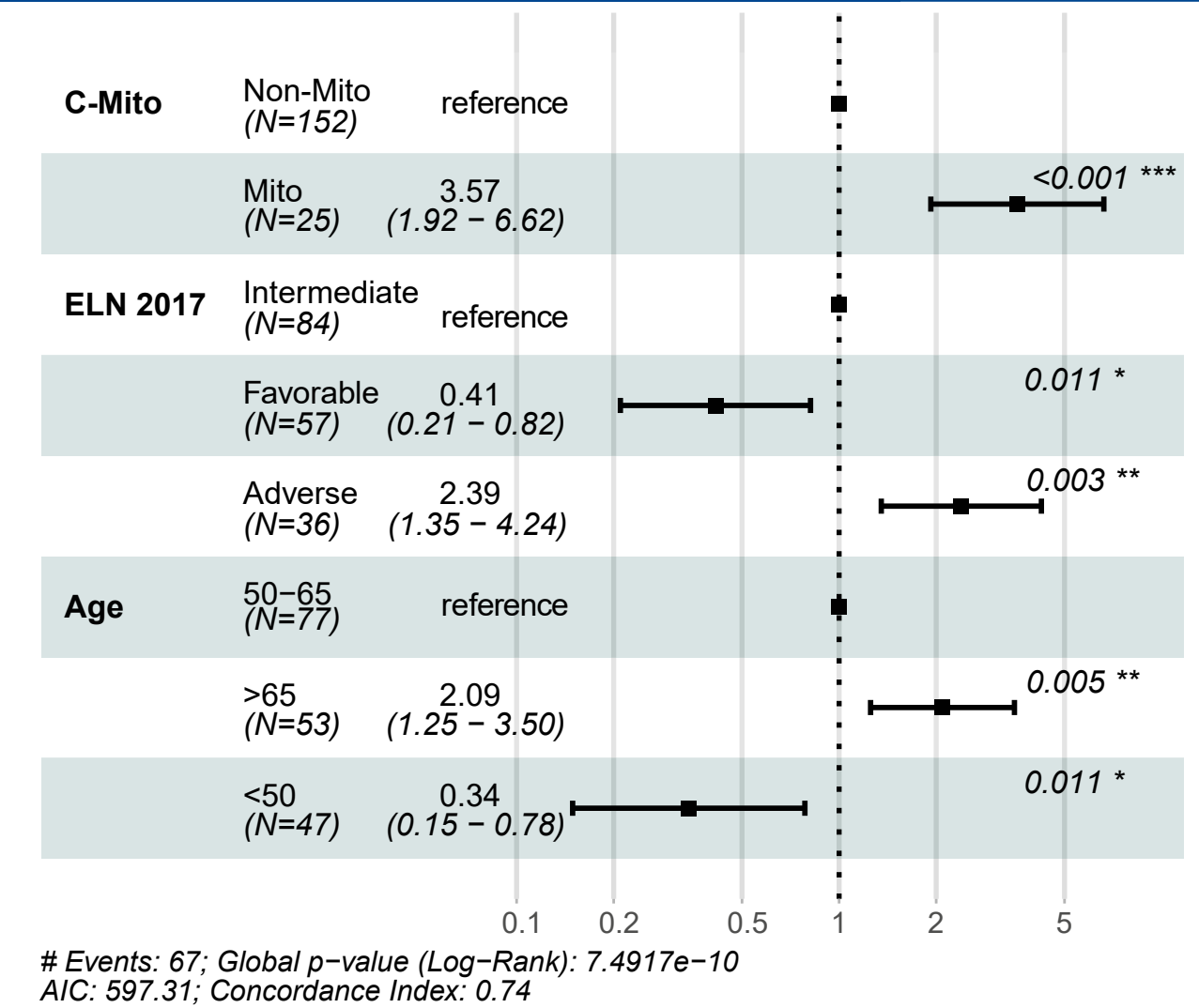
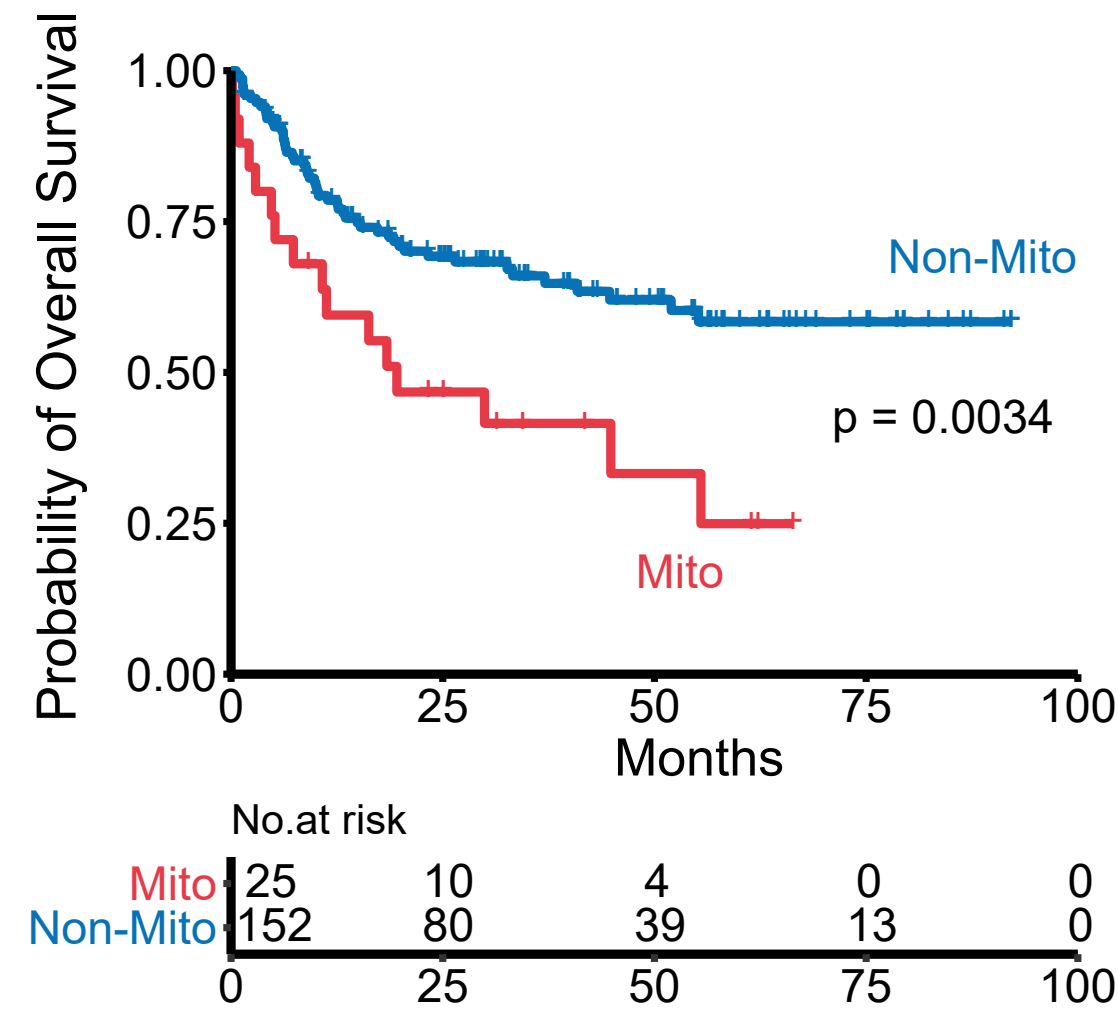
SUBCLUSTERS HAVE A DISTINCT SURVIVAL PROFILE

- Clusters 1 and 5 have distance survival phenotype
- Survival for patients in cluster 1 significantly worse, for cluster 5 significantly better



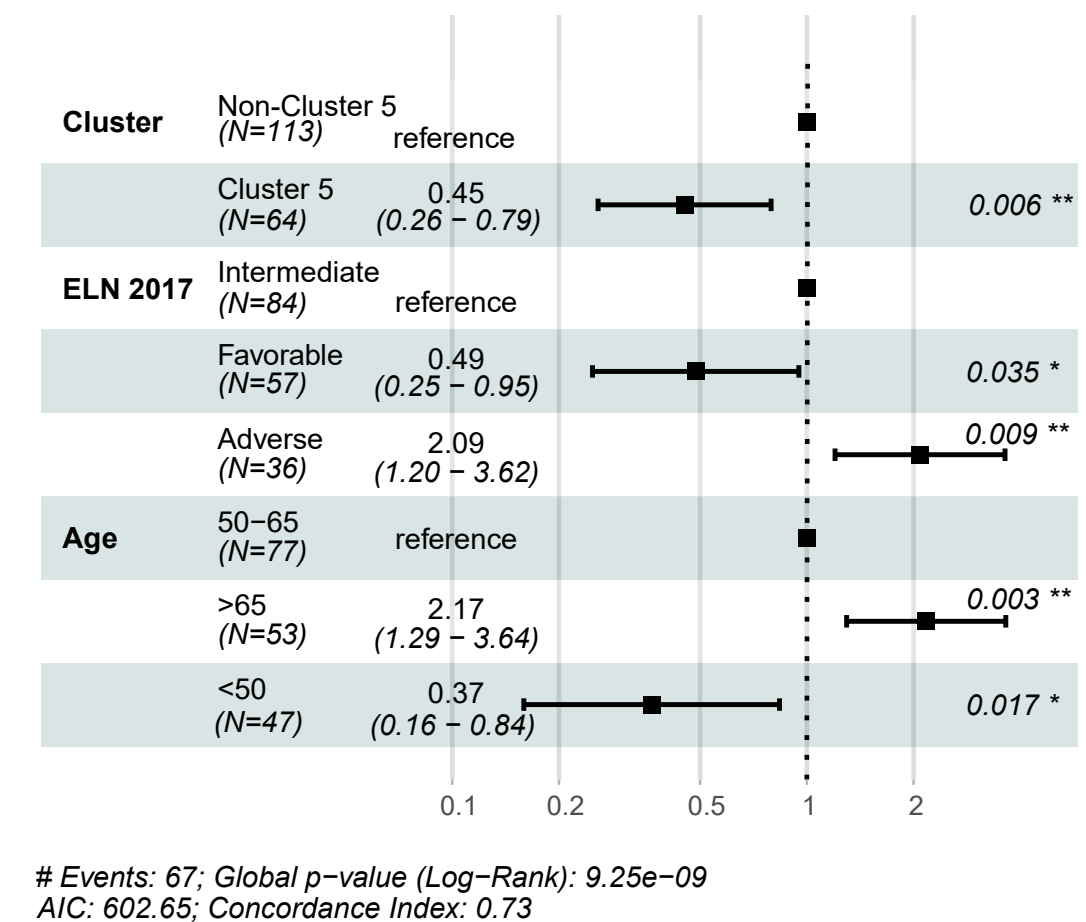
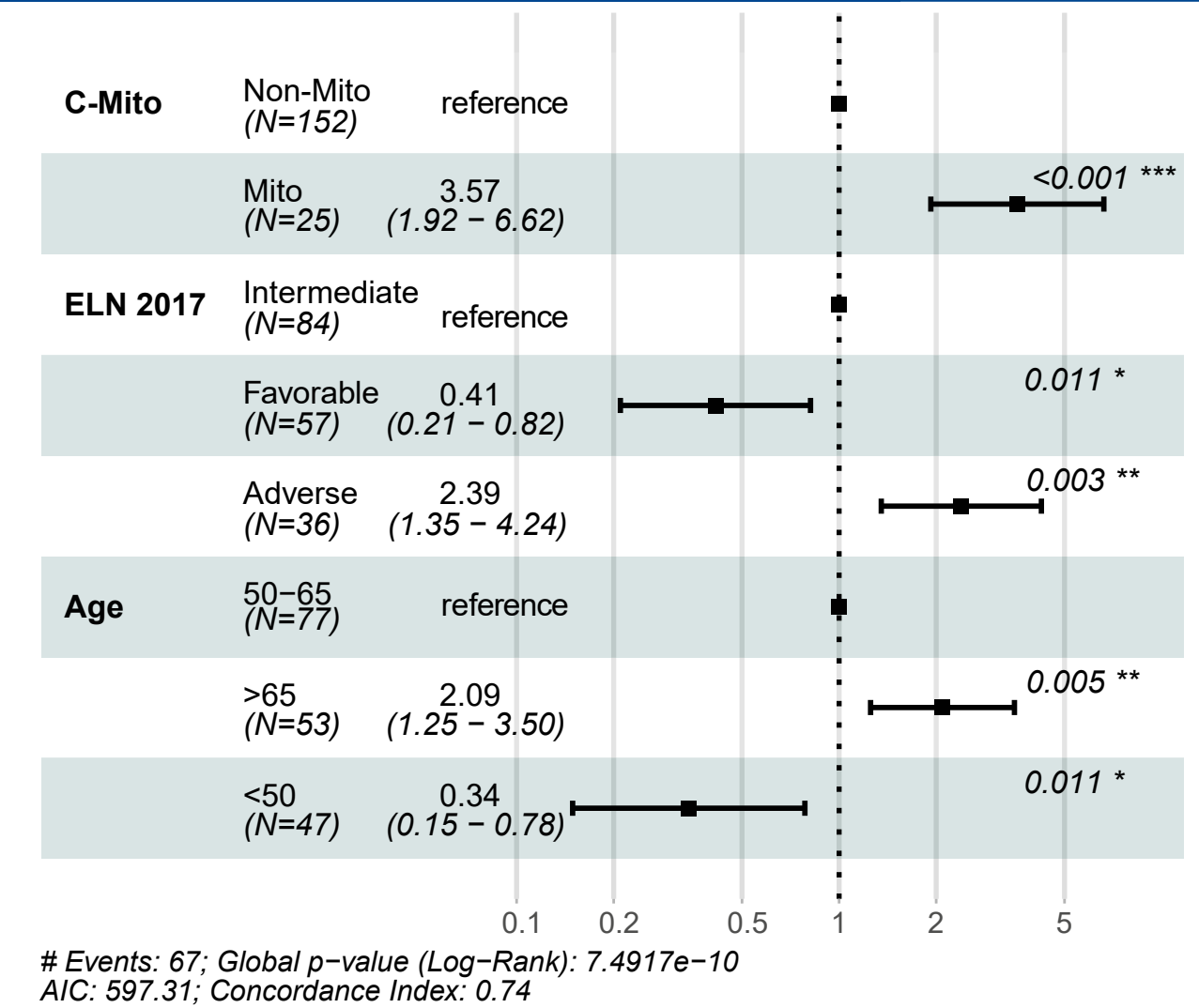
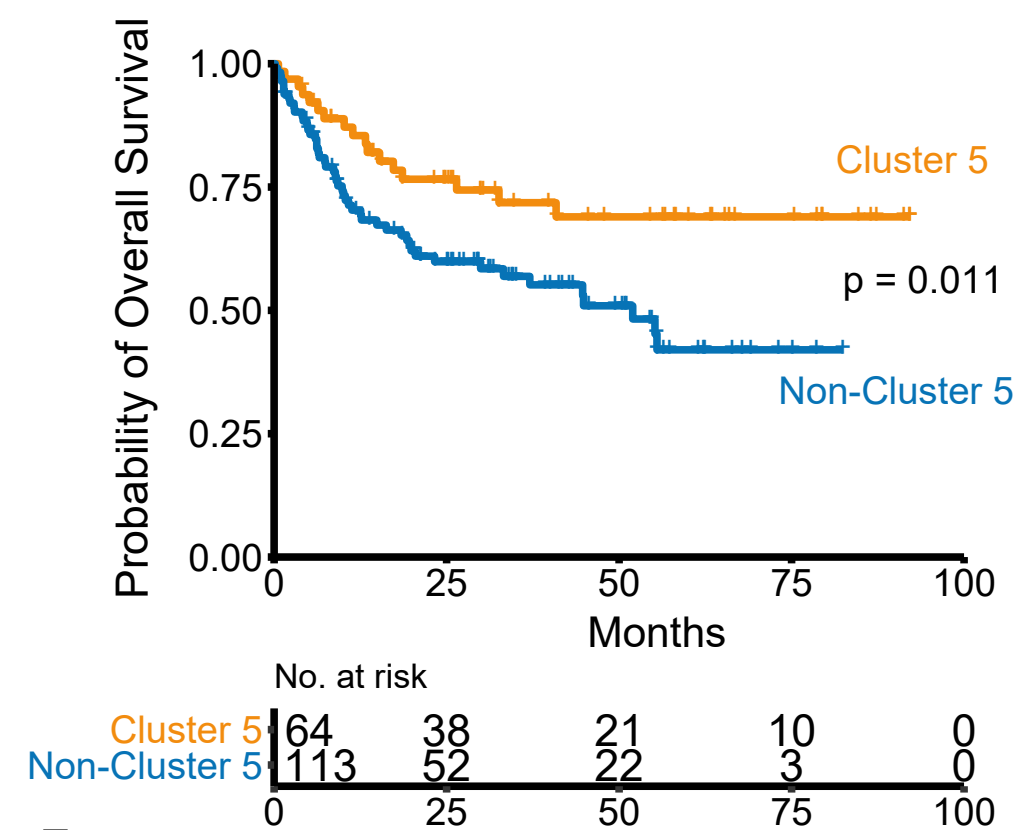
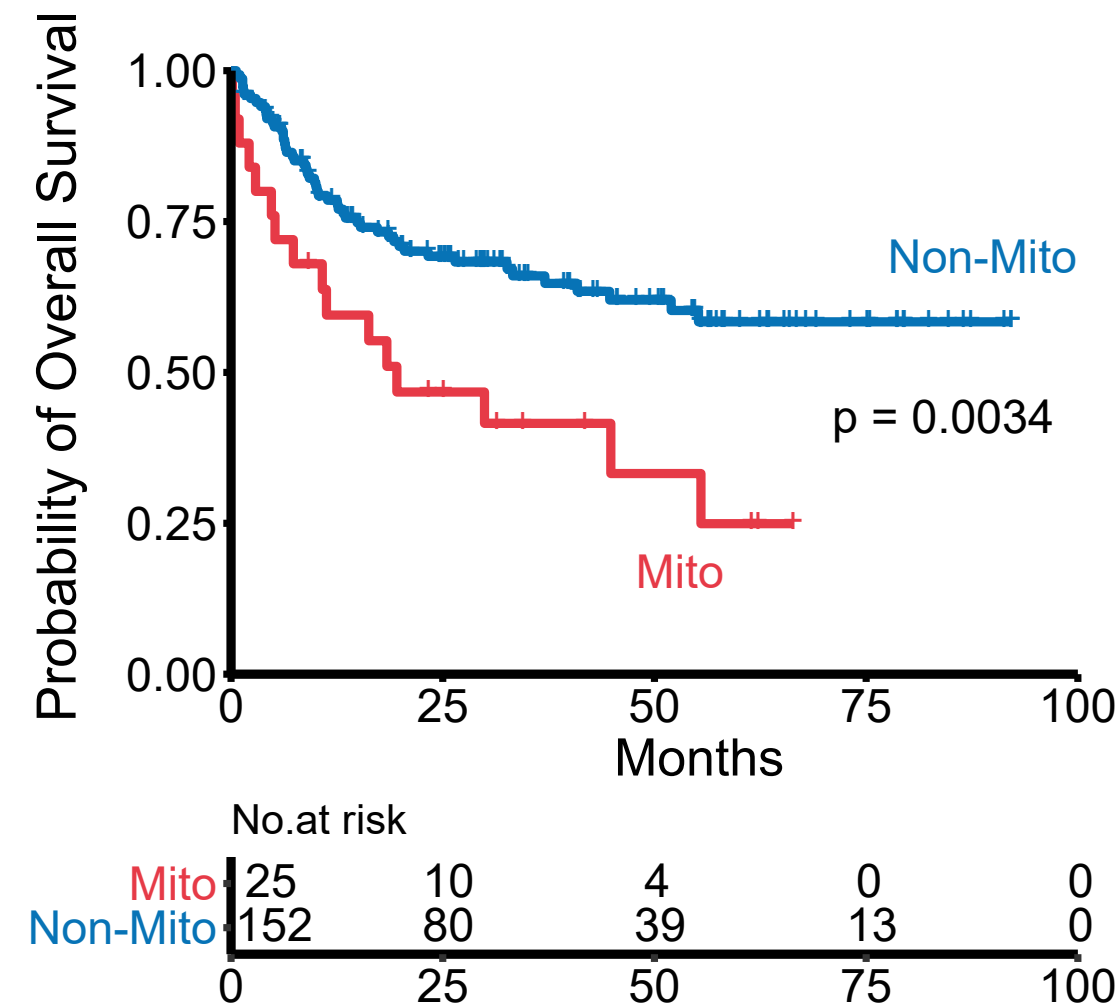
SUBCLUSTERS HAVE A DISTINCT SURVIVAL PROFILE

- Clusters 1 and 5 have distance survival phenotype
- Survival for patients in cluster 1 significantly worse, for cluster 5 significantly better



SUBCLUSTERS HAVE A DISTINCT SURVIVAL PROFILE

- Clusters 1 and 5 have distance survival phenotype
- Survival for patients in cluster 1 significantly worse, for cluster 5 significantly better

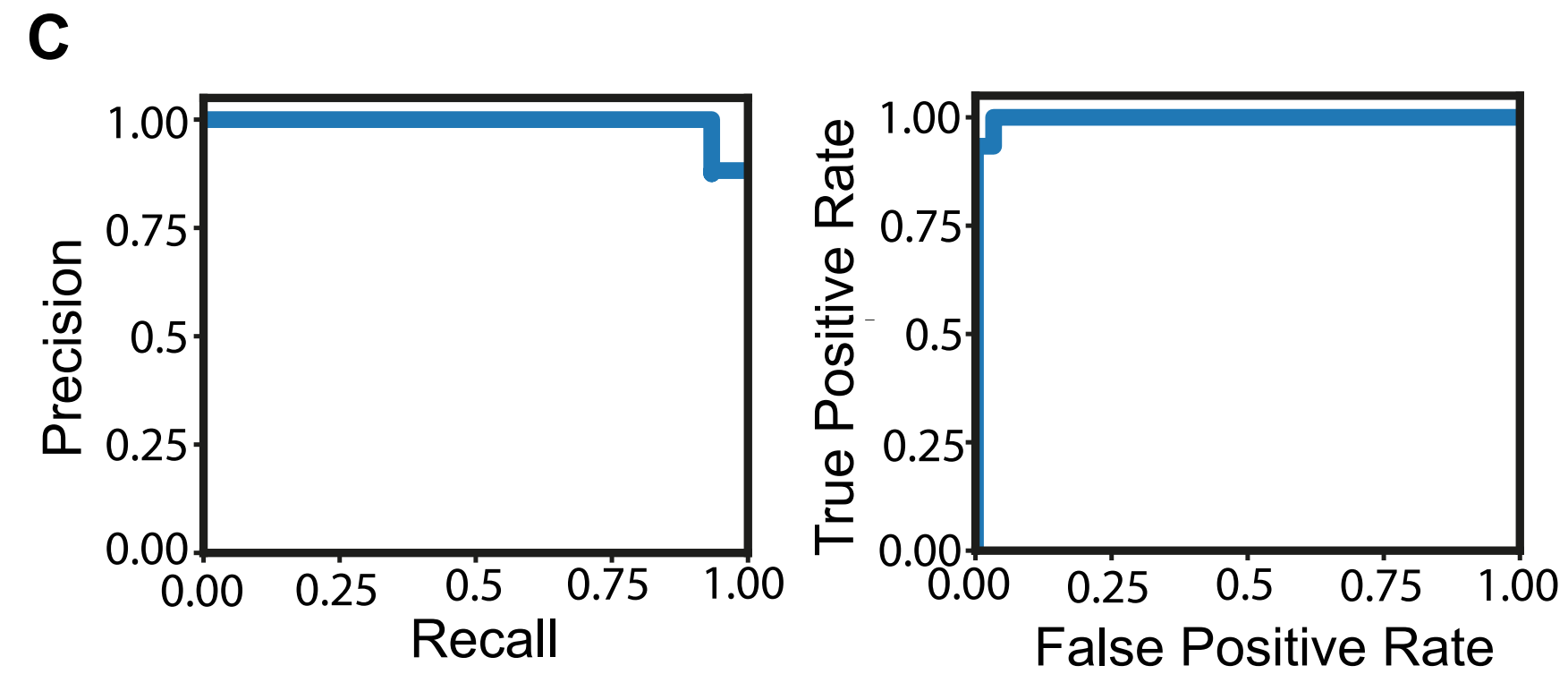


MITO CLUSTER CAN BE PREDICTED PROSPECTIVELY

- Train supervised classifier on discovery cohort
- Identify small set of predictive proteins
- Test on validation cohort

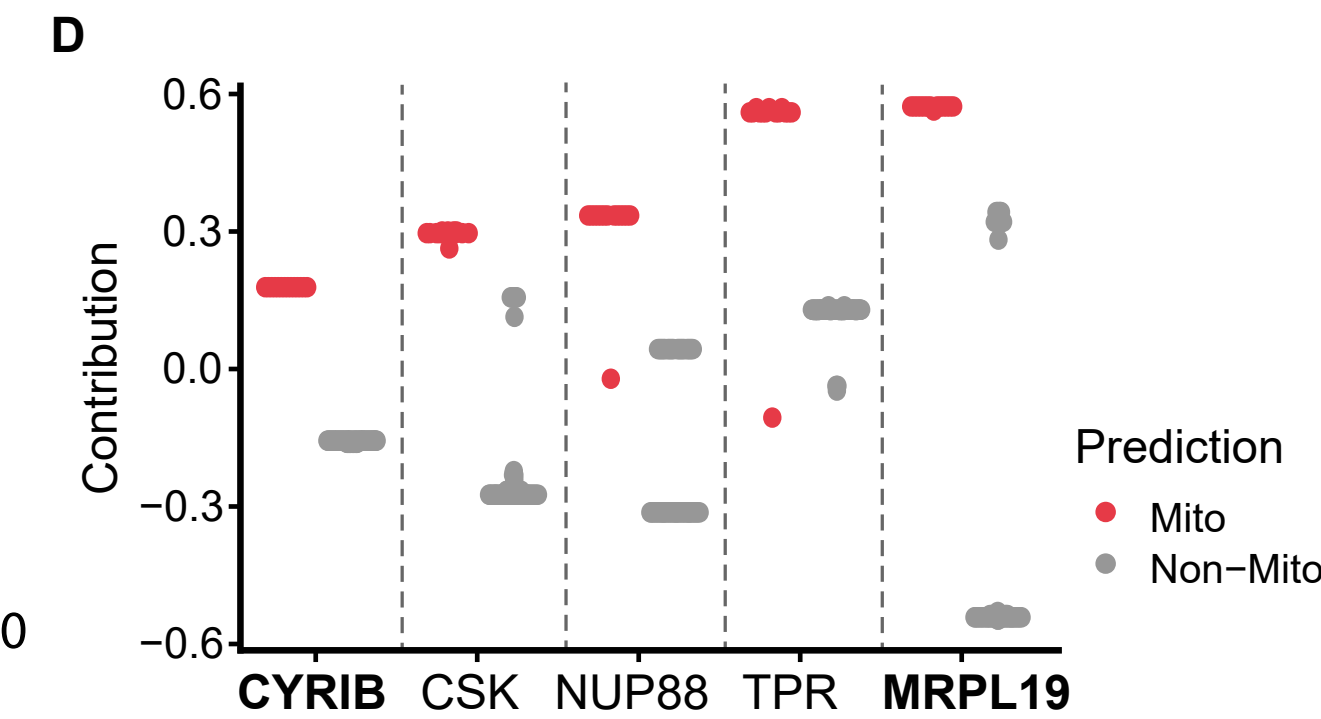
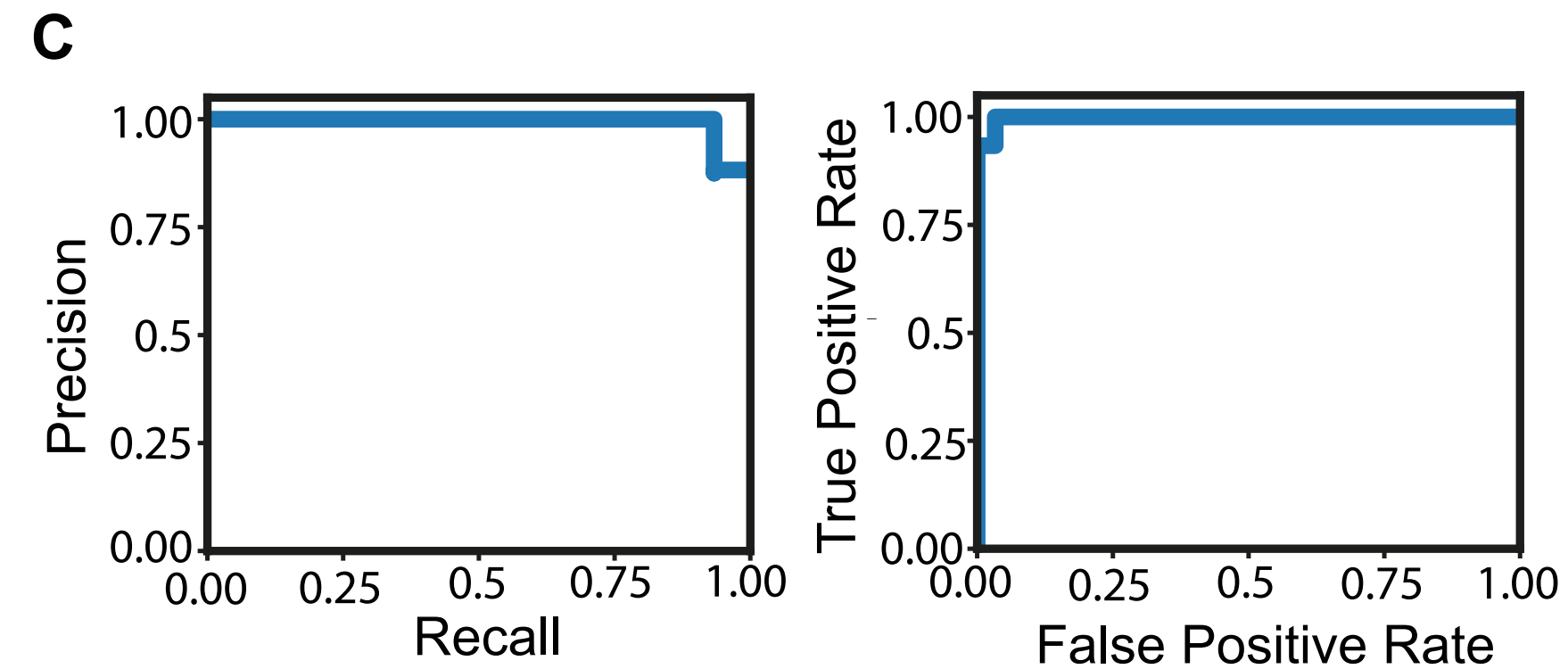
MITO CLUSTER CAN BE PREDICTED PROSPECTIVELY

- Train supervised classifier on discovery cohort
- Identify small set of predictive proteins
- Test on validation cohort



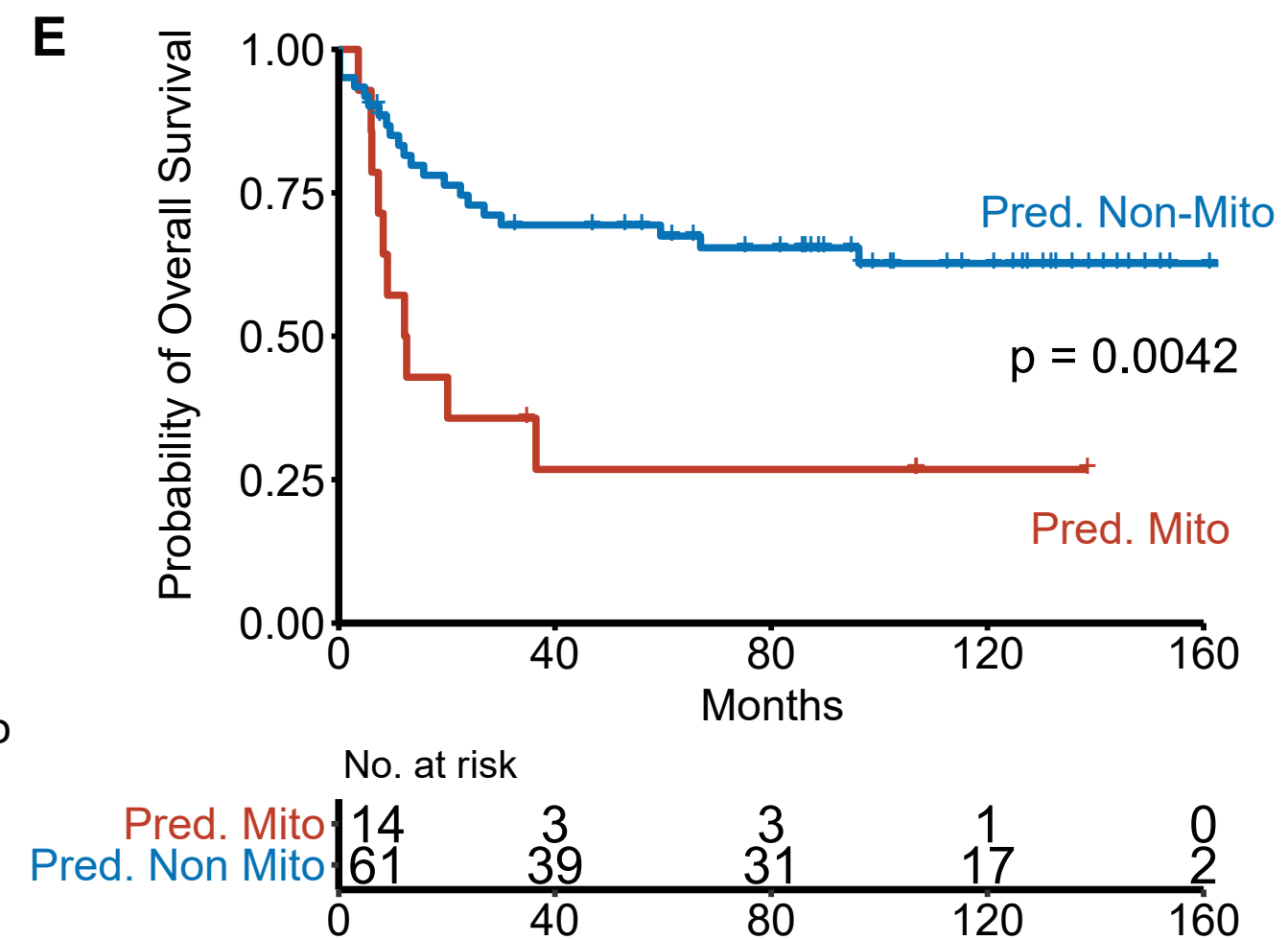
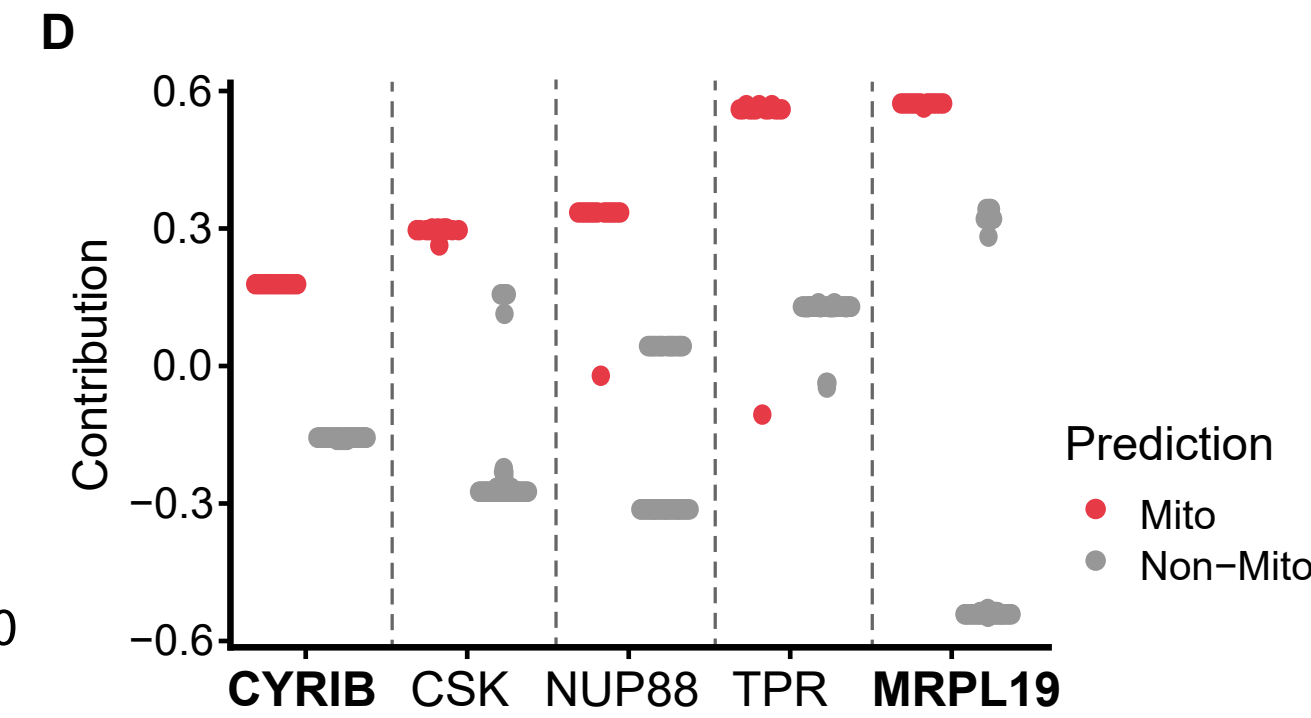
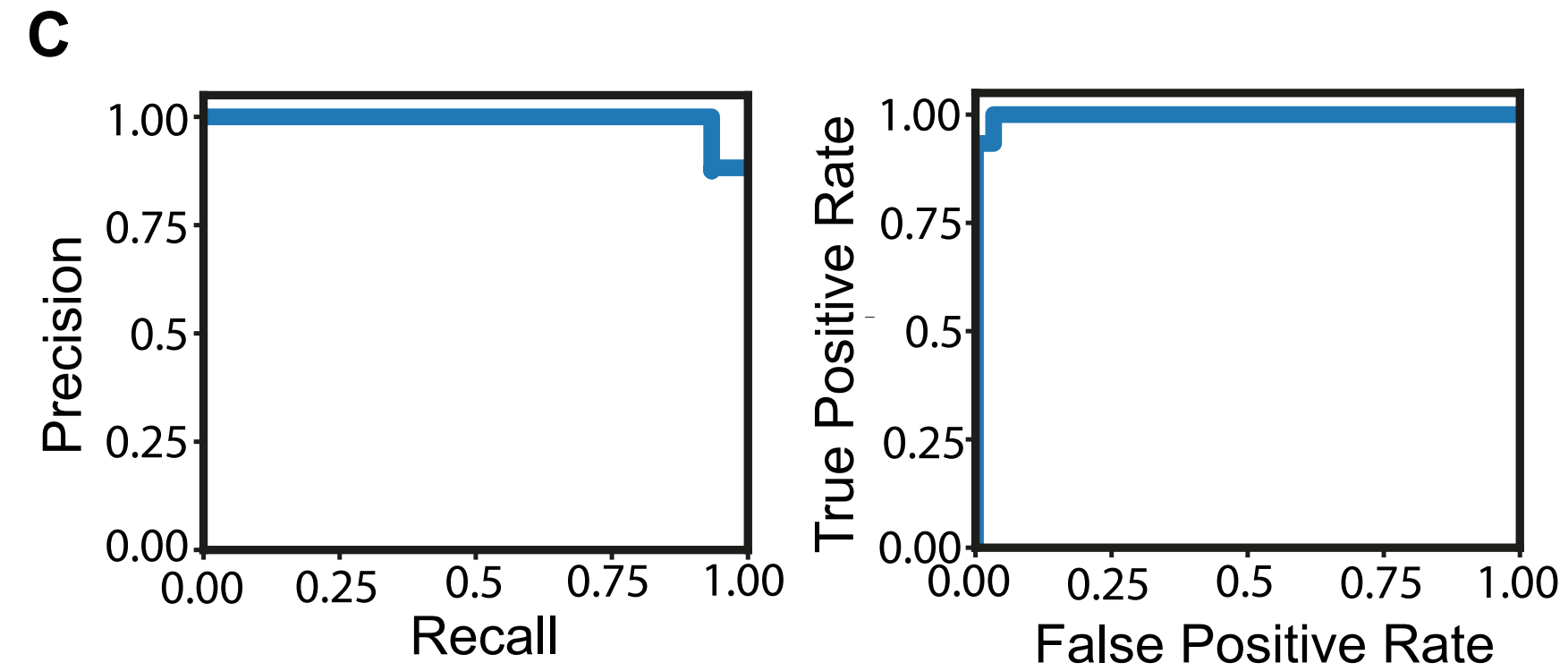
MITO CLUSTER CAN BE PREDICTED PROSPECTIVELY

- Train supervised classifier on discovery cohort
- Identify small set of predictive proteins
- Test on validation cohort



MITO CLUSTER CAN BE PREDICTED PROSPECTIVELY

- Train supervised classifier on discovery cohort
- Identify small set of predictive proteins
- Test on validation cohort



SUMMARY AND OUTLOOK

- Multi-omics approach discovered proteomic AML subtypes with clinical relevance
- Mito-AML is hypersensitive to drugs targeting mitochondrial complex I
- Use mito classifier to stratify patients for venetoclax-based therapies

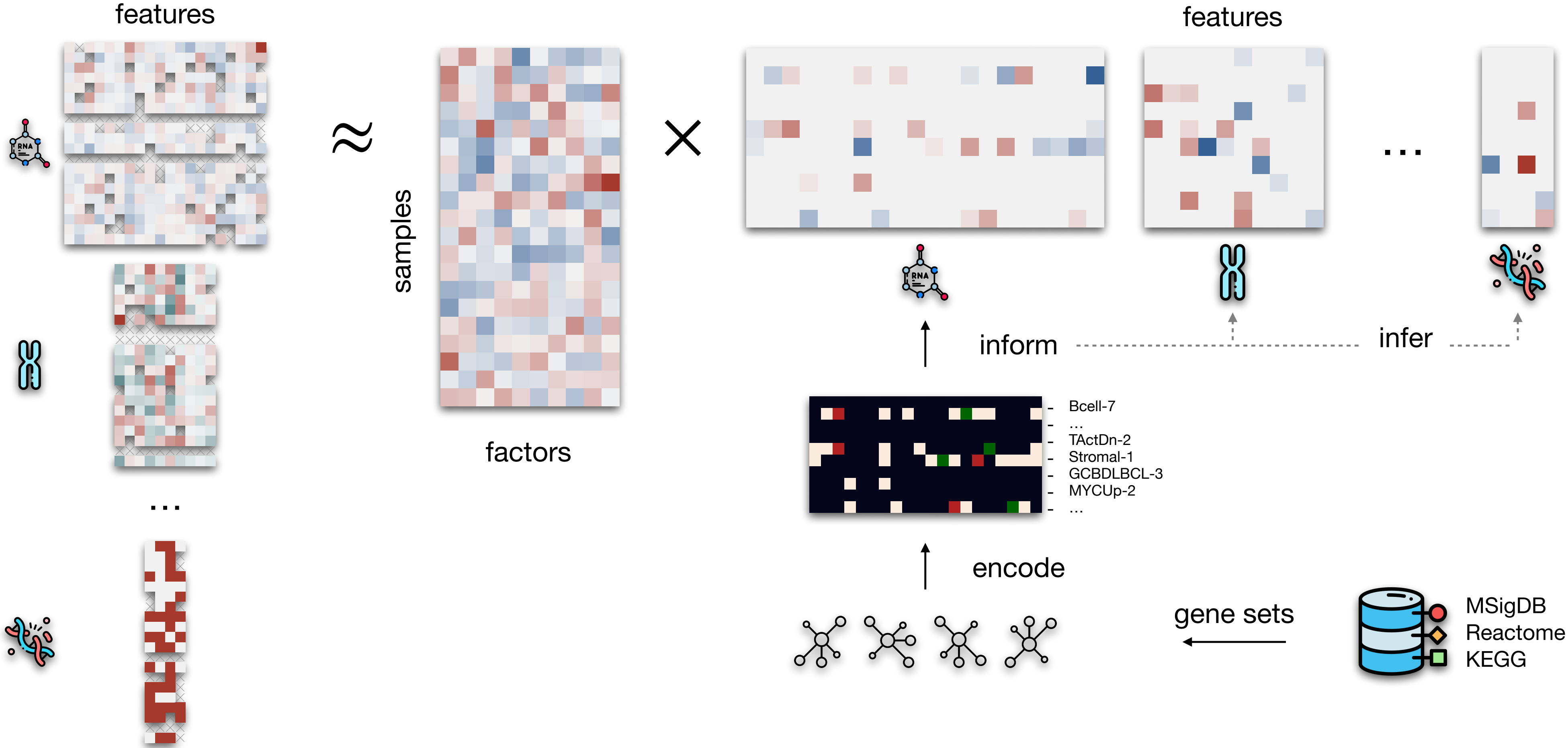
LEVERAGING PRIOR INFORMATION IN MULTI-OMICS MODELLING

- Characterising factors in FA models is challenging
- Use prior knowledge from pathways already during inference
- Associate each factor to a pathway

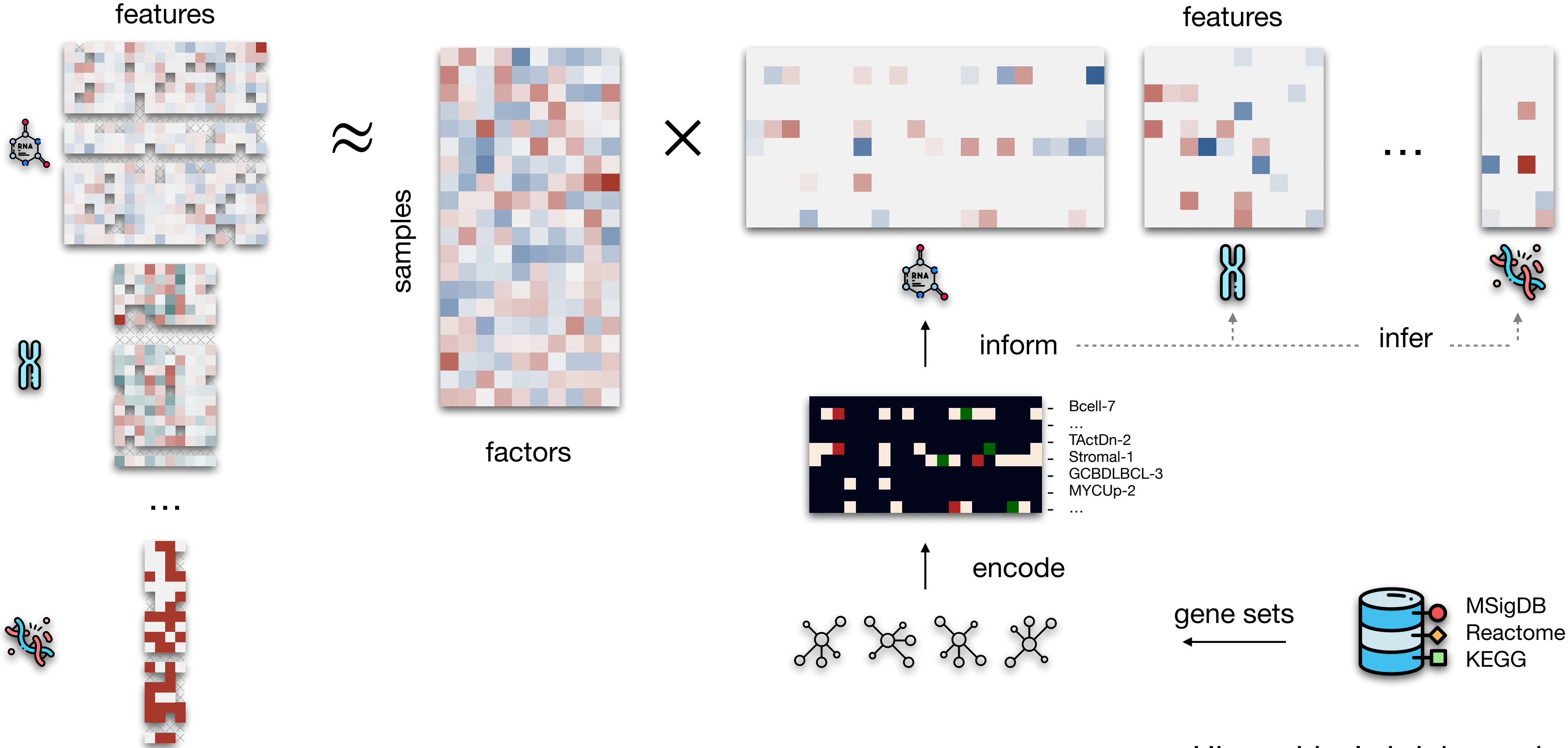
LEVERAGING PRIOR INFORMATION IN MULTI-OMICS MODELLING

- Characterising factors in FA models is challenging
- Use prior knowledge from pathways already during inference
- Associate each factor to a pathway
- Infer pathways driving inter-patient variation
- Refine and customise pathway annotations
- Identify interpretable patient sub-populations

MUVI: A MULTI-VIEW LATENT VARIABLE MODEL WITH DOMAIN-INFORMED STRUCTURED SPARSITY

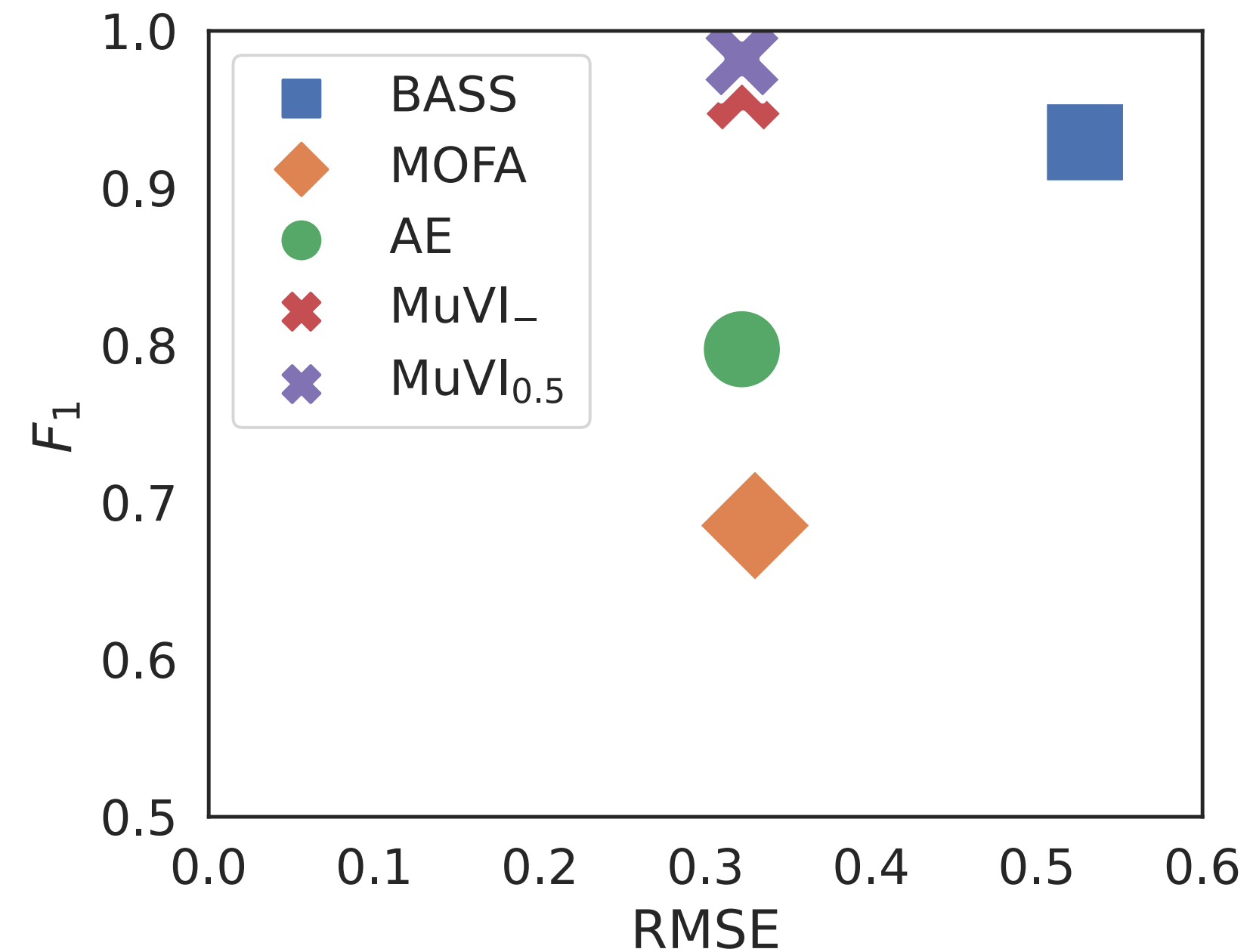


MUVI: A MULTI-VIEW LATENT VARIABLE MODEL WITH DOMAIN-INFORMED STRUCTURED SPARSITY



MUVI COMBINES DOMAIN-INDUCED SPARSITY WITH LOW RECONSTRUCTION ERROR

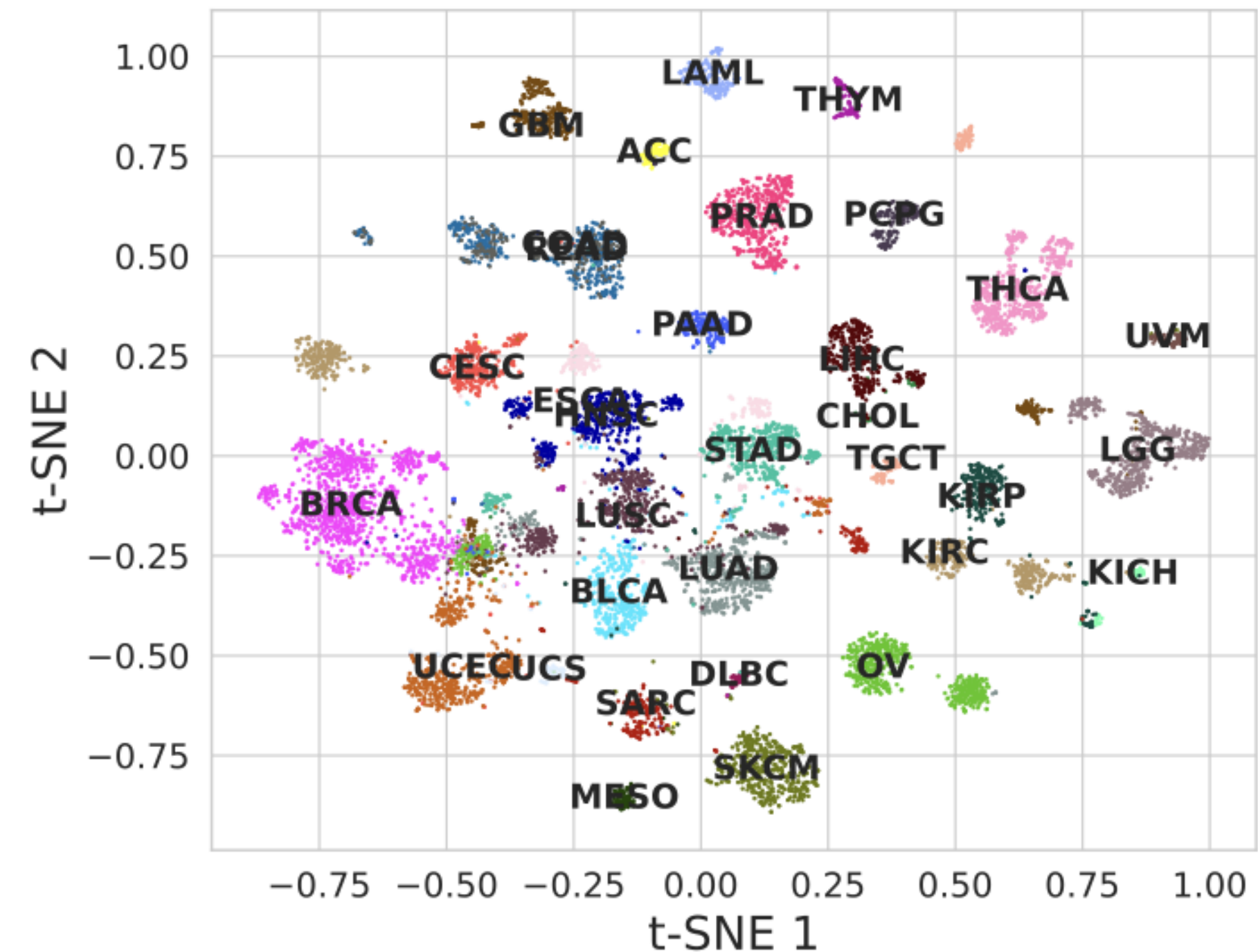
- Simulate multi-view expression data with noisy pathways
- Compare MUVI to
 - MOFA
 - BASS (Bayesian group factor analysis with structured sparsity)
 - Multi-view VAE



PAN-CANCER MULTI-OMICS ANALYSIS OF TCGA

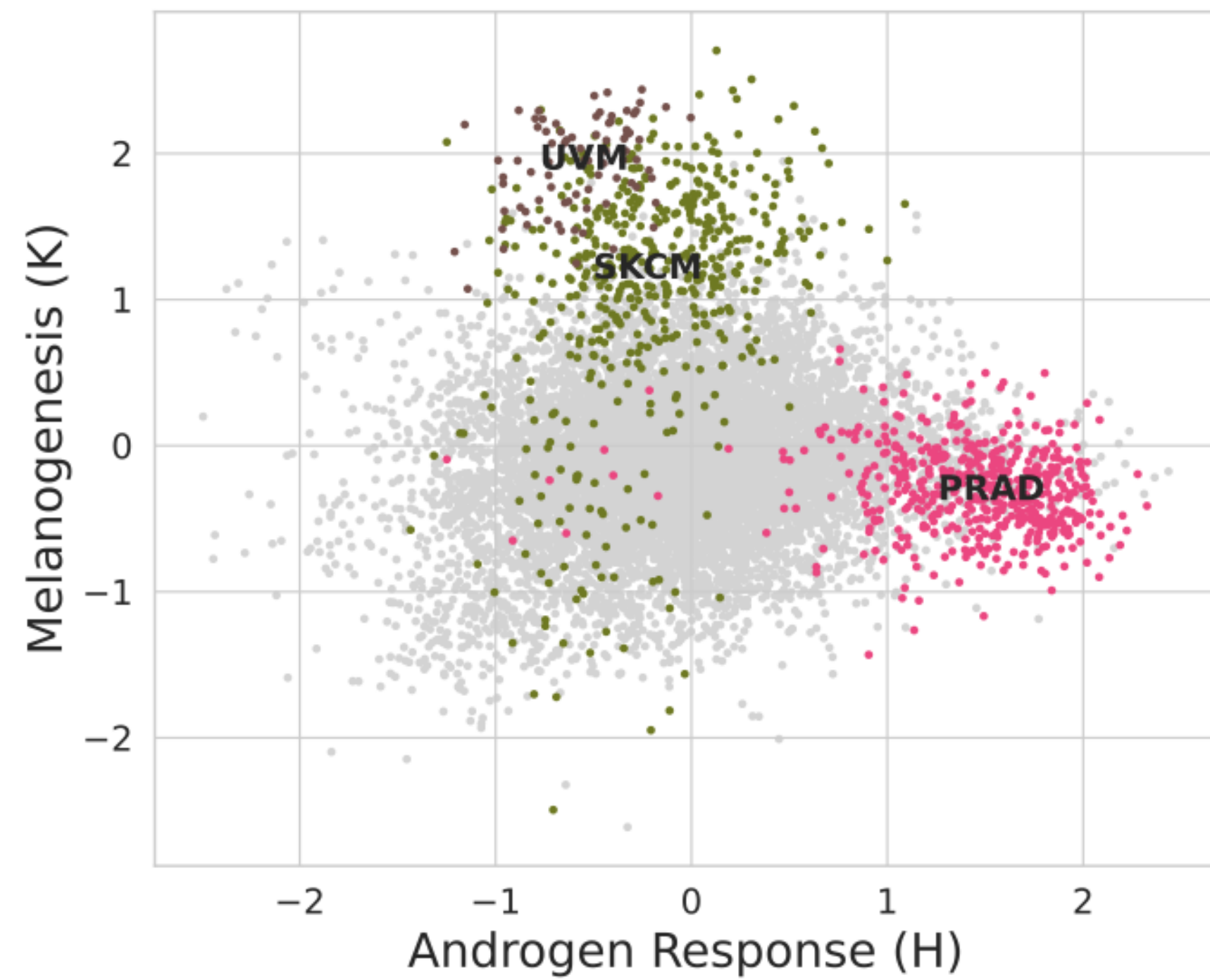
- Use MuVI to analyze TCGA data
 - 4 data views (DNA methylation, mRNA expression, microRNA and reverse phase protein array (RPPA))
 - 11k patients, 33 cancer types

Global view

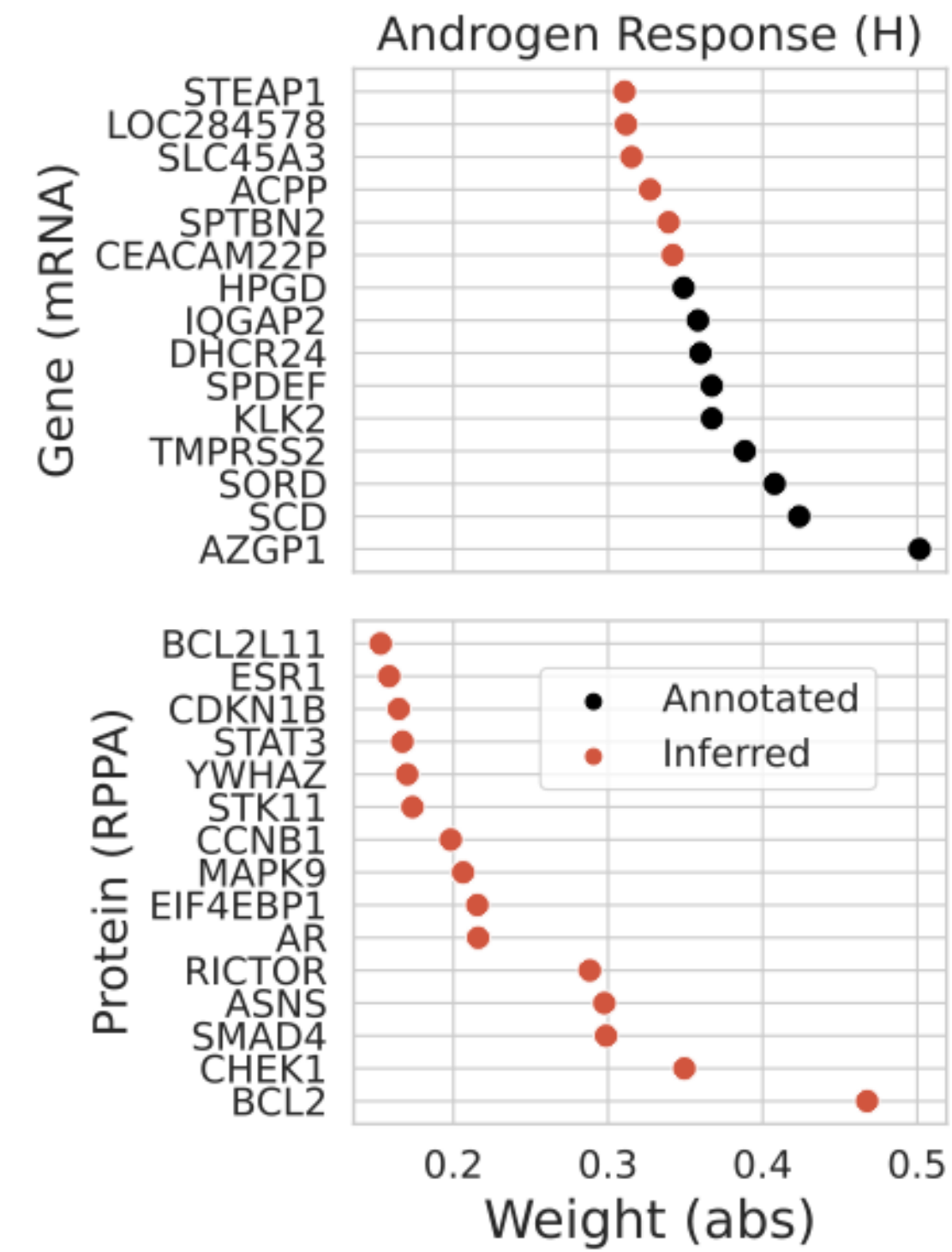


MULTI-SCALE ANALYSIS

Factor



Local



SUMMARY

SUMMARY

- Use latent variable models to infer unobservable hidden states from observable data, taking prior knowledge into account

SUMMARY

- Use latent variable models to infer unobservable hidden states from observable data, taking prior knowledge into account
- Discover shared and unique drivers of variability for multi-omics data of AML patients
- Software

SUMMARY

- Use latent variable models to infer unobservable hidden states from observable data, taking prior knowledge into account
- Discover shared and unique drivers of variability for multi-omics data of AML patients
- Software
 - github.com/bioFAM

SUMMARY

- Use latent variable models to infer unobservable hidden states from observable data, taking prior knowledge into account
- Discover shared and unique drivers of variability for multi-omics data of AML patients
- Software
 - github.com/bioFAM
 - <https://github.com/MLO-lab/MuVI>

SUMMARY

- Use latent variable models to infer unobservable hidden states from observable data, taking prior knowledge into account
- Discover shared and unique drivers of variability for multi-omics data of AML patients
- Software
 - github.com/bioFAM
 - <https://github.com/MLO-lab/MuVI>
 - <https://mlo-lab.github.io/>

ACKNOWLEDGEMENTS

- Ricard Arguelaget, Oliver Stegle (EBI, Cambridge)
- Thomas Oellerich, Sebastian Wolf (Uniklinik Frankfurt)
- Matthias Mann, Ashok Jayavelu (MPI München)
- Arber Qoku, Andreas Kopf, Sebastian Gruber (MLO/DKTK, Frankfurt)

