Machine Learning RNA Signatures for DLBCL Prognosis

Stacking Ensemble Approach for COO-Specific Survival Prediction

Bioinformatics Analysis

2026-01-19

Section 1

Introduction

Background: DLBCL Heterogeneity

Diffuse Large B-Cell Lymphoma (DLBCL)

- Most common aggressive non-Hodgkin lymphoma
- 30-40% of patients experience treatment failure
- Cell-of-Origin (COO) classification provides prognostic value

COO Subtypes:

Subtype	Biology	Prognosis
GCB	Germinal center B-cell origin	Generally favorable
ABC	Activated B-cell origin	Often aggressive
MHG	Molecular high-grade	Poor prognosis
UNC	Unclassified	Variable

Rationale for ML Approach

Why Machine Learning for Prognostic Signatures?

1. **High-dimensionality**: 29,372 probes vs. 1,300 samples
2. **Feature selection**: Identify most informative genes
3. **Non-linear patterns**: Capture complex biological interactions
4. **Ensemble methods**: Combine complementary model strengths

Key Question:

Can we build RNA expression signatures that predict survival independently within each COO subtype?

Section 2

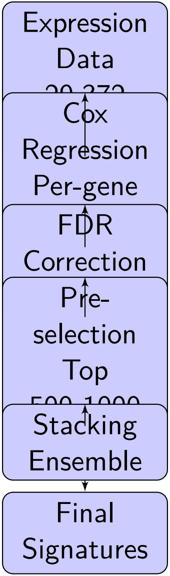
Methods

Study Cohort: HMRN/Lacy Dataset

Total Samples	1,303
Platform	Illumina HumanHT-12 V4
GCB	517 (39.7%)
ABC	345 (26.5%)
MHG	164 (12.6%)
UNC	277 (21.3%)
Events (Deaths)	676 (51.9%)
Median Follow-up	4.2 years

- Population-based cohort from UK
- R-CHOP or R-CHOP-like treatment
- Comprehensive clinical annotation
- COO assigned by gene expression profiling

Analysis Pipeline Overview



Expression Data 20,372

Cox Regression Per-gene

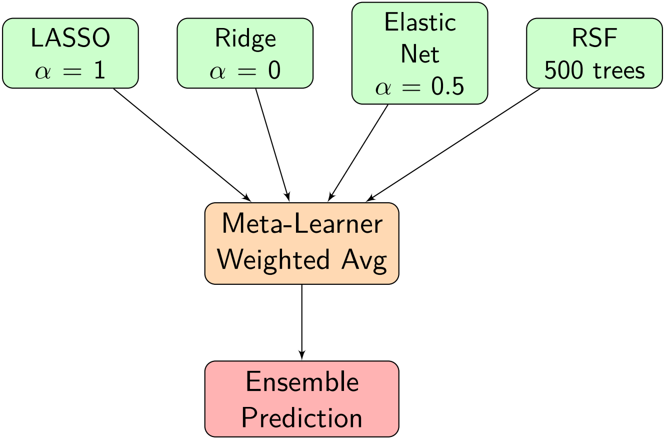
FDR Correction

Pre-selection Top 500-1000

Stacking Ensemble

Final Signatures

Stacking Ensemble Architecture



- **5-fold CV**: Stratified by event status
- **Weights**: Based on discriminative power $|C - 0.5|$
- **Prediction inversion**: When C-index < 0.5

Base Learner Rationale

Model	Strengths	Parameters
LASSO	Sparse solutions, interpretable	$\alpha = 1$, λ by CV
Ridge	Handles collinearity, stable	$\alpha = 0$, λ by CV
Elastic Net	Balance sparsity/stability	$\alpha = 0.5$, λ by CV
RSF	Non-linear, interactions	500 trees, nodesize=10

Why Stacking?

- Different regularization captures different signals
- RSF captures non-linear gene-gene interactions
- Meta-learner optimally combines complementary strengths

Section 3

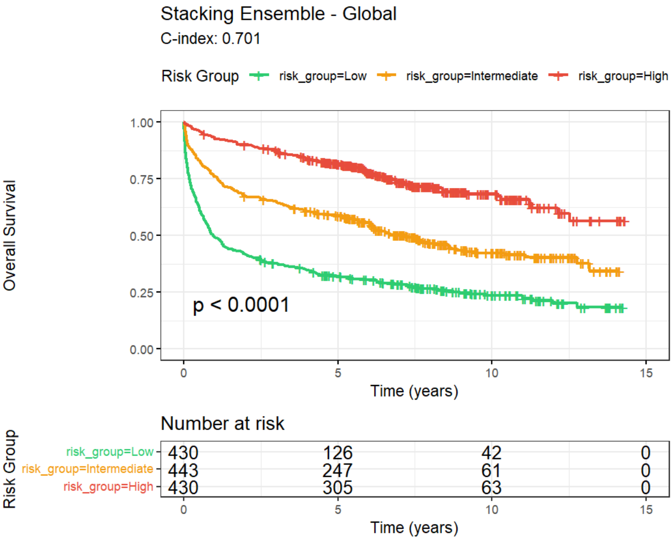
Results

Ensemble Model Performance

Subset	N	Events	C-index	Log-rank p
Global	1303	676	0.701	<0.001
GCB	517	222	0.711	<0.001
ABC	345	229	0.693	<0.001
MHG	164	100	0.728	8.77e-15
UNC	277	125	0.705	3.49e-13

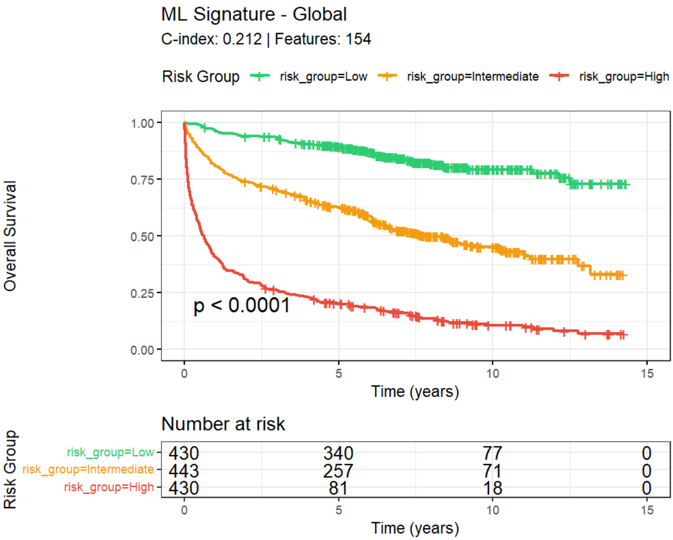
- **C-index $>$ 0.7** across all subsets
- MHG achieves highest discrimination (0.728)
- All signatures highly significant ($p < 0.001$)

Global Signature: Stacking Ensemble KM



Global cohort (n=1,303): Clear separation of Low/Intermediate/High risk groups

Global Signature: Elastic Net Model KM



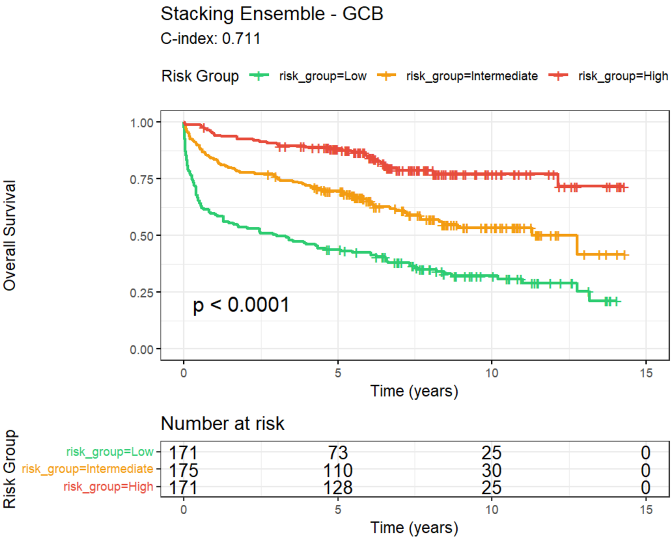
ML Signature - Global

C-index: 0.212 | Features: 154

Risk Group	0	5	10	15
risk_group=Low	430	340	77	0
risk_group=Intermediate	443	257	71	0
risk_group=High	430	81	18	0

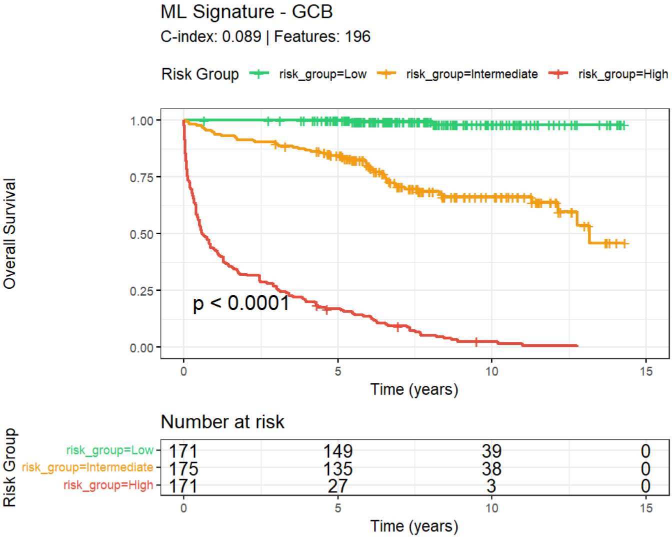
Elastic Net Cox model with 154 selected features

GCB Subtype: Stacking Ensemble KM



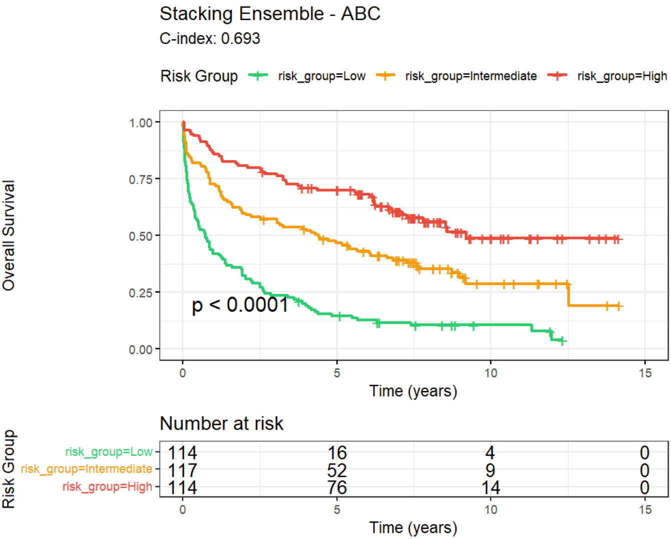
GCB subtype (n=517): C-index = 0.711, strong risk stratification

GCB Subtype: Elastic Net Model KM



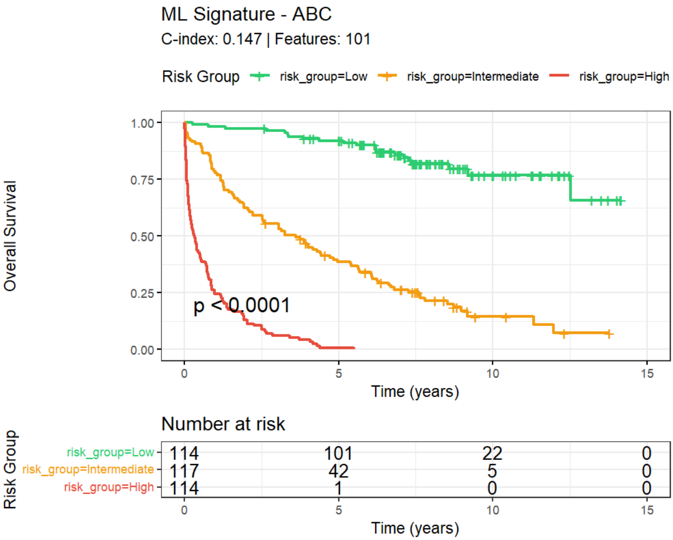
GCB Elastic Net signature with 196 features

ABC Subtype: Stacking Ensemble KM



ABC subtype (n=345): C-index = 0.693, effective despite aggressive biology

ABC Subtype: Elastic Net Model KM

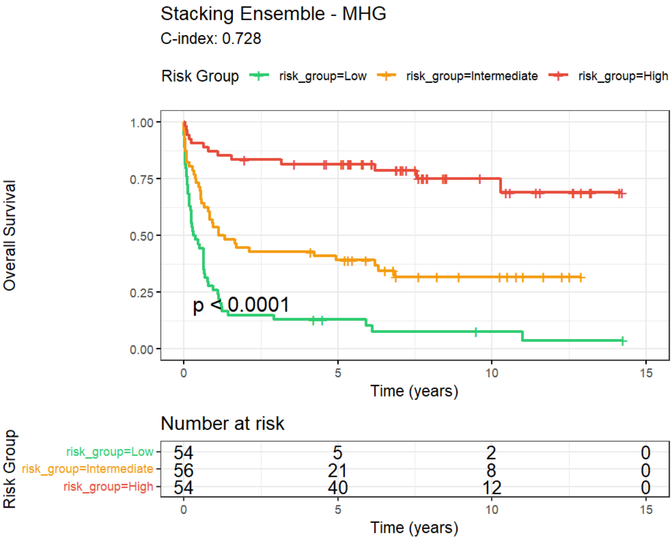


Number at risk

Risk Group	0	5	10	15
risk_group=Low	114	101	22	0
risk_group=Intermediate	117	42	5	0
risk_group=High	114	1	0	0

ABC Elastic Net signature with 101 features

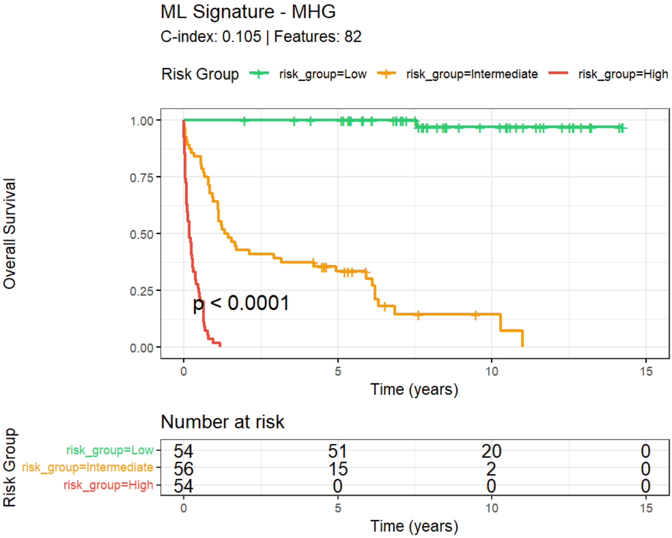
MHG Subtype: Stacking Ensemble KM



Risk Group	0	5	10	15
risk_group=Low	54	5	2	0
risk_group=Intermediate	56	21	8	0
risk_group=High	54	40	12	0

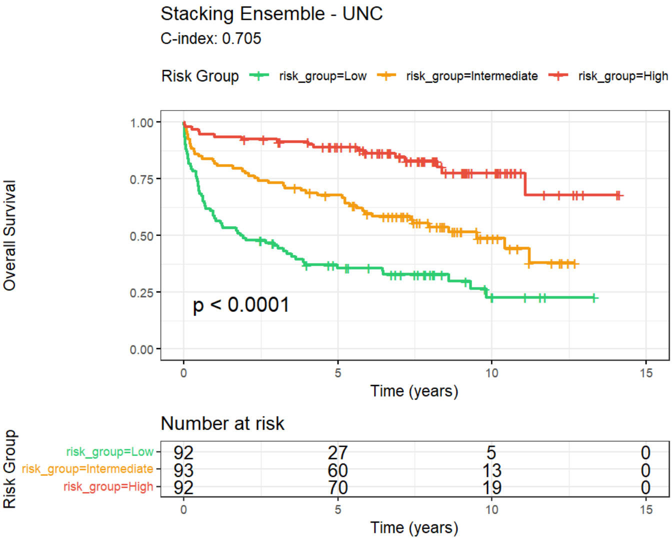
MHG subtype (n=164): **Highest C-index = 0.728** - best discrimination

MHG Subtype: Elastic Net Model KM



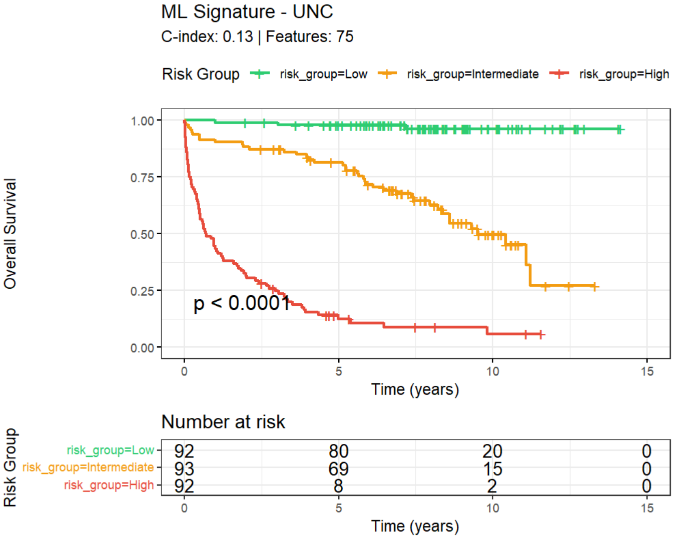
MHG Elastic Net signature with 82 features

UNC Subtype: Stacking Ensemble KM



UNC subtype (n=277): C-index = 0.705, risk stratification in unclassified cases

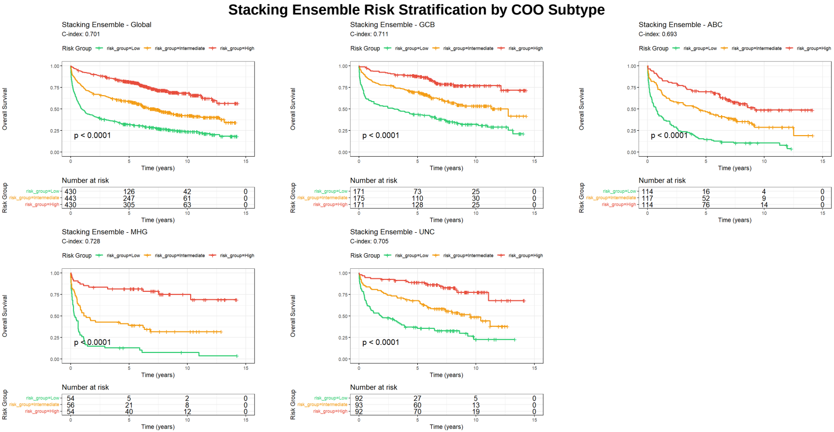
UNC Subtype: Elastic Net Model KM



Risk Group	0	5	10	15
risk_group=Low	92	80	20	0
risk_group=Intermediate	93	69	15	0
risk_group=High	92	8	2	0

UNC Elastic Net signature with 75 features

KM Curves: All Subtypes Comparison



Stacking Ensemble - Global: Number at risk

Risk Group	0	5	10	15
risk_group=Low	430	126	42	0
risk_group=Intermediate	443	247	61	0
risk_group=High	430	305	63	0

Stacking Ensemble - GCB: Number at risk

Risk Group	0	5	10	15
risk_group=Low	171	73	25	0
risk_group=Intermediate	175	110	30	0
risk_group=High	171	128	25	0

Stacking Ensemble - ABC: Number at risk

Risk Group	0	5	10	15
risk_group=Low	114	16	4	0
risk_group=Intermediate	117	52	9	0
risk_group=High	114	76	14	0

Stacking Ensemble - MHG: Number at risk

Risk Group	0	5	10	15
risk_group=Low	54	5	2	0
risk_group=Intermediate	56	21	8	0
risk_group=High	54	40	12	0

Stacking Ensemble - UNC: Number at risk

Risk Group	0	5	10	15
risk_group=Low	92	27	5	0
risk_group=Intermediate	93	60	13	0
risk_group=High	92	70	19	0

Section 4

Biological Insights

Global Signature: Key Genes

Adverse (Higher Expression = Worse Survival)

Gene	HR	Function
PTDSS2	1.11	Phospholipid metabolism
ANGPTL4	1.09	Angiogenesis, metastasis
CD300LG	1.08	Immune checkpoint
MT1G	1.07	Metallothionein, stress response
FCN3	1.07	Complement activation

Global Signature: Protective Genes

Gene	HR	Function
PRND	0.90	Prion protein family
KLRC1	0.92	NK cell receptor (NKG2A)
CUX2	0.93	Transcription factor
JCHAIN	0.94	Immunoglobulin J chain
CD1E	0.97	Lipid antigen presentation

Key Theme: Protective genes enriched for:

- T-cell and NK cell immune markers
- Antigen presentation machinery
- Immune surveillance components

Global: Pathway Enrichment

Adverse Genes (Poor Prognosis):

Pathway	FDR	Key Genes
E2F Targets	8.7e-12	PLK4, RRM2, RAD51C
G2M Checkpoint	2.5e-07	UBE2C, KIF23, CDC25A
MYC Targets V1	1.2e-06	TYMS, HSPD1, MCM4
MYC Targets V2	5.0e-04	PLK4, MCM4
mTORC1 Signaling	0.06	RRM2, CDC25A

Interpretation: Cell cycle/proliferation genes associated with aggressive disease

Global: Protective Pathways

Protective Genes (Better Prognosis):

Pathway	FDR	Key Genes
Allograft Rejection	2.4e-11	CD3D/E/G, CD4, CD8A/B
IL-2/STAT5 Signaling	4.9e-06	EOMES, TNFRSF9, CTLA4
EMT	8.0e-04	IL32, ADAM12
Complement	0.014	GZMK, CD40LG, LCK
Coagulation	0.028	PROC, MMP9

Interpretation: T-cell infiltration and immune activation predict favorable outcomes

GCB-Specific Insights

Signature Size: 196 genes (C-index = 0.711)

Key Biological Themes:

Theme	Genes	Interpretation
Apoptosis	CD2, BID, LEF1	Death pathway regulation
Angiogenesis	CCND2, LUM, S100A4	Tumor vascularization
E2F/Cell Cycle	TCF19, KIF18B, E2F8	Proliferation markers
Wnt/β-catenin	LEF1, AXIN1	GCB developmental pathway

GCB-Specific: Strong enrichment for apoptosis pathway genes suggests GCB tumors retain sensitivity to programmed cell death

ABC-Specific Insights

Signature Size: 101 genes (C-index = 0.693)

Key Biological Themes:

Theme	Genes	Interpretation
E2F Targets	DSCC1, BIRC5, CDC25A	DNA replication
MYC Targets	NOP2, WDR74, HSPD1	Ribosome biogenesis
Unfolded Protein	ATF4, EIF4E	ER stress response
IL-2/STAT5	IL2RB, IL3RA	Cytokine signaling

ABC-Specific: MYC target enrichment reflects constitutive NF-κB and MYC pathway activation characteristic of ABC-DLBCL

MHG-Specific Insights

Signature Size: 82 genes (C-index = 0.728)

Highest Performing Signature - Key Genes:

Gene	HR	Direction	Function
PNLDC1	1.23	Adverse	piRNA biogenesis
C8B	1.22	Adverse	Complement component
SLC12A3	1.22	Adverse	Ion transport
ARHGAP22	0.82	Protective	RhoGAP, cell migration
H2BU1	0.84	Protective	Histone variant

MHG Insight: Best discrimination suggests MHG represents a distinct biological entity with unique prognostic features

UNC-Specific Insights

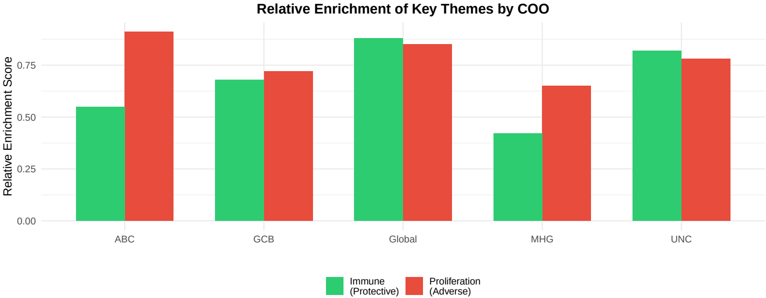
Signature Size: 75 genes (C-index = 0.705)

Key Biological Themes:

Theme	Genes	Interpretation
Allograft Rejection	CD8A, CD3D/E/G, LCP2	T-cell immunity
Estrogen Response	TIAM1, BCL11B, MYC	Hormone signaling
E2F Targets	RRM2, CDKN2A, CHEK1	Cell cycle control
Wnt Signaling	MYC, TCF7, SKP2	Developmental pathway

UNC Insight: Shares features of both GCB and ABC, with strong immune infiltration signal predicting better outcomes

Cross-Subtype Comparison



Shared vs. Unique Gene Patterns

Genes Appearing in Multiple Signatures:

Gene	Subtypes	Direction	Function
KLRC1	Global, UNC	Protective	NK receptor
CUX2	Global, ABC	Protective	Transcription factor
CD1E	Global, GCB, ABC	Protective	Antigen presentation
RRM2	ABC, UNC	Adverse	DNA synthesis
MYC targets	ABC, UNC	Adverse	Proliferation

Key Finding:

Immune genes (especially T-cell/NK markers) are **protective** across all subtypes, while proliferation genes are **adverse**

Section 5

Summary

Key Findings

1. Effective Risk Stratification

Stacking ensemble achieves C-index 0.70-0.73 across all COO subtypes

2. Universal Prognostic Themes

- Adverse: Cell cycle, E2F/MYC targets, proliferation
- Protective: T-cell infiltration, NK cells, antigen presentation

3. COO-Specific Biology

- GCB: Apoptosis pathways, Wnt signaling
- ABC: MYC/NF-κB targets, ER stress
- MHG: Best discrimination, unique epigenetic features
- UNC: Hybrid features, strong immune signal

Clinical Implications

Potential Applications:

1. **Risk stratification** beyond IPI and COO
2. **Treatment selection** based on molecular features
3. **Clinical trial** patient enrichment

Therapeutic Insights:

Finding	Implication
Immune infiltration protective	Immunotherapy potential
Proliferation genes adverse	CDK/cell cycle inhibitors
MYC enrichment in ABC	BET inhibitors
Apoptosis defects	BCL2 inhibitors

Limitations & Future Directions

Limitations:

- Single cohort analysis (requires external validation)
- Array platform (RNA-seq may capture additional biology)
- Retrospective study design

Future Work:

1. Validate in independent DLBCL cohorts
2. Integrate with genetic/mutational data
3. Develop clinical-grade assay
4. Prospective clinical validation

Acknowledgments

Data Source:

- HMRN/Lacy cohort (GSE181063)
- Illumina HumanHT-12 V4 BeadChip

Methods:

- R packages: glmnet, randomForestSRC, survival
- GSEA via Enrichr (MSigDB Hallmark)

Analysis Pipeline:

- Stacking ensemble with 4 base learners
- 5-fold stratified cross-validation
- Weighted average meta-learner

Questions?

Code and data available upon request